Book of Short Papers SIS 2018

Editors: Antonino Abbruzzo - Eugenio Brentari
Marcello Chiodi - Davide Piacentino

SIS 2018
PALERMO
49 TH SCIENTIFIC MEETING
OF THE ITALIAN
STATISTICAL SOCIETY
20–22 JUNE

Copyright © 2018

PUBLISHED BY PEARSON

WWW.PEARSON.COM

First printing, November 2018, ISBN-9788891910233

SIS
2018
49 TH SCIENTIFIC MEETING
OF THE ITALIAN
STATISTICAL SOCIETY
PALERMO 20-22 JUNE

Contents

5.3 Advances in Statistical Models 855

5.4 Advances in Time Series 892

5.5 Data Management 922

5.6 Developments in Graphical Models 959

SIS
2018
49 TH SCIENTIFIC MEETING
OF THE ITALIAN
STATISTICAL SOCIETY
PALERMO 20-22 JUNE

1. Preface

This book includes the papers presented at the "49th Meeting of the Italian Statistic Society". The conference has registered 445 participants, 350 reports divided into 4 plenary sessions, 20 specialised sessions, 25 sessions solicited, 27 sessions spontaneous, 2 poster sessions. The high number of participants, the high quality of the interventions, the productive spirit of the conference, the ability to respect the time table, are the main indices of the full success of this conference. The meeting hosted also, as plenary sessions, the ISTAT annual report 2018, and a round table on statistics and job markets. Methodological plenary sessions concerned with ordinal data, the dynamics of climate change and models in biomedicine.

Moreover, two related events were held: Start-up Research (SUR) and Stats Under the Stars (SUS4). The SUS4 event attracted many sponsors of statistical, financial, editorial fields as well as numerous students, not only from Italy but also from abroad (Groningen, Tyumen, Barcelona, and Valencia): 98 students for a total of 25 teams. The SUR was a 2-day meeting where small research groups of young scholars, advised by senior researchers with a well-established experience in different areas of Statistics, was asked to develop innovative methods and models to analyse a common dataset from the Neurosciences.

2. Plenary Sessions

A new paradigm for rating data models
Un nuovo paradigma per modelli su dati di preferenza

Domenico Piccolo

Abstract Rating data arise in several disciplines and the class of Generalized Linear Models (GLM) provides a consolidated methodology for their analysis: such structures (and a plethora of variants) model the cumulative probabilities of ordinal scores as functions of subjects' covariates. A different perspective can be adopted when considering that discrete choices as ordinal assessments are the result of a complex interaction between subjective perception and external circumstances. Thus, an explicit specification of the inherent uncertainty of the data generating process is needed. This paradigm has triggered a variety of researches and applications, conveying in the unifying framework of GEneralized Mixtures with uncertainty (GEM) which encompasses also classical cumulative models. Some critical discussions conclude the paper.

Abstract *Dati di preferenza sono presenti in differenti ambiti e la classe dei modelli lineari generalizzati fornisce una metodologia consolidata per la loro analisi. Una prospettiva differente deriva dal considerare le scelte discrete che producono dati ordinali come un processo derivante da (almeno) due componenti: una percezione soggettiva (feeling) ed una ineliminabile indecisione (uncertainty). Cosí, una specificazione esplicita dell'indecisione nel processo generatore dei dati di preferenza si é resa necessaria. Tale paradigma ha generato numerose varianti e sviluppi che includono anche l'approccio piú tradizionale. Il lavoro introduce i modelli* GEM *e si conclude con alcune considerazioni critiche.*

Key words: Scientific paradigm, Rating data, Empirical evidence, Statistical models, CUB models, GEM models

Domenico Piccolo
Department of Political Sciences, University of Naples Federico II, Naples, Italy. e-mail: domenico.piccolo@unina.it

1 Introduction

Statistical models are formal tools to describe, understand and predict real phenomena on the basis of some recognizable mechanism to be effectively estimated and tested on empirical data. This approach is ubiquitous in modern research and establishes the basis to advance disputable topics in any field: a model can be falsified and rejected to favour a more sustainable alternative, as it is for any progress in scientific research and human knowledge.

In this perspective, the specification step of a model derives from empirical evidence, rational deductions, analogy and similarities. Given a set of data and contextual information, statistical methods are nowadays sufficiently developed to provide suitable specifications: time series, longitudinal models, qualitative variables, count data, survival measures, continuous or discrete observations, experimental design, reliability and quality control, etc. are settings where consolidated theories and effective methods suggest consistent specifications.

These topics will be discussed in the field of ordinal data modelling, as ratings of objects or activities, opinions towards facts, expressions of preferences, agreement to sentences, judgements, evaluations of items, perceptions, subjective sensations, concerns, worry, fear, anxiety, etc. All these expressions are collected as ratings/scores by means of verbal or pseudo-verbal categories. Such data are available in different fields: Marketing researches, Socio-political surveys, Psychological enquiries, Sensory sciences and Medicine, among others. For these analyses, several approaches are available as log-linear and marginal models [48], contingency tables inference [47], and so on. In fact, the leitmotif of our discussion is that observed rating data are realizations of a (genuine) discrete or of a discretized process derived by an intrinsically continuous (latent) variable [6].

2 The classical paradigm

Emulating the approach introduced for the logistic regression, the paradigm of cumulative models considers the probability of responses less or equal to a rating r as a function of selected regressors. A convenient link is necessary to establish a one-to-one correspondence between a 0-1 measure (the probability) and a real quantity (the value of the regressors); as a consequence, probit, logit, complementary log-log links, etc. are proposed to specify the corresponding models.

The vast quantity of results derived from such procedures represents the dominant paradigm in theoretical and empirical literature about the statistical analysis of rating data. The process of data generation is introduced by invoking -for each subject- a latent variable Y_i^*, taking values along the real line and related to the values $\boldsymbol{t}_i$ of p explanatory subjects' covariates via the classical regression model: $Y_i^* = \boldsymbol{t}_i\boldsymbol{\beta} + \varepsilon_i, i = 1,2,\dots,n$. For a given m, the relationship with the discrete ordinal variable R_i, defined over the discrete support $\{1,2,\dots,m\}$, is provided by:

$$\alpha_{r-1} < Y_i^* \leq \alpha_r \qquad \Longleftrightarrow \qquad R_i = r\,, \quad r = 1,2,\dots,m\,,$$

where $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_m = +\infty$ are the thresholds (cutpoints) of the continuous scale of the latent variables Y_i^*.

Then, if $\varepsilon_i \sim F_\varepsilon(.)$, the probability mass function of R_i, for $r = 1,2,\dots,m$, is:

$$Pr(R_i = r) = Pr(\alpha_{r-1} < Y_i^* \leq \alpha_r) = F_\varepsilon(\alpha_r - \boldsymbol{t}_i\boldsymbol{\beta}) - F_\varepsilon(\alpha_{r-1} - \boldsymbol{t}_i\boldsymbol{\beta}), \qquad (1)$$

where $Pr(R_i \leq r|\boldsymbol{\theta},\boldsymbol{t}_i) = F_\varepsilon(\alpha_r - \boldsymbol{t}_i\boldsymbol{\beta})$ for $i = 1,2,\dots,n$ and $r = 1,2,\dots,m$.

This specification requires the knowledge of $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ parameters, that is the $(m-1)$ intercept values $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_{m-1})'$ in addition to the explicit covariate parameters $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)'$.

When selecting logistic random variables ε_i, the probability of a single rating turns out to be:

$$Pr(R_i = r|\boldsymbol{\theta},\boldsymbol{t}_i) = \frac{1}{1+\exp(-[\alpha_r - \boldsymbol{t}_i\boldsymbol{\beta}])} - \frac{1}{1+\exp(-[\alpha_{r-1} - \boldsymbol{t}_i\boldsymbol{\beta}])}\,, \qquad (2)$$

for $i = 1,2,\dots,n$ and $r = 1,2,\dots,m$. As a consequence of proportionality properties, these models are known as *proportional odds models* (POM) [1, 66].

Cumulative models have been embedded into the GLM perspective [51] and generalized in several directions: multilevel, varying choice of thresholds, multivariate setting, *conditional logits* including both subjects' and objects' covariates [52], variable effect $\boldsymbol{t}_i\boldsymbol{\beta}_j$ as in *stereotype* models [5], or thresholds which depend on covariates as in *partial proportional odds models* [57], and models with dispersion effects as the *location-scale models* [50] or *location-shift models* [67].

Quite often, the interpretation of these models takes advantage of odds and log-odds measures which are quantities easily manageable by Medicine and Biomedical researchers; similarly, plotting devices explain the direction and the effect of significant covariates on the ordinal responses and new graphical solutions have been recently advanced [68]. In order to anchor the estimated results to easier and more interpretable indexes [2, 3] some difficulties have been emphasized.

Indeed, some drawbacks of the classical specification should be noticed:

- Data generating process refers to a latent variable whose unobservable distribution defines the discrete distribution for the observable ratings.
- It is difficult to accept that -in any circumstance- subjects perform choices by considering ratings not greater than a fixed one, whereas it is more common to consider choices as determined by the “stimulus” associated to a single category and its surrounding values. In fact, the relationship (2) is difficult to manage for deriving immediately the effect of a set of covariates on the categorical response.
- If ratings generated by several items have to be clustered, ranked or classified, unconditionally from covariates, the classical setting leads to a saturated model, which implies an arithmetic equivalence between observed and assumed distributions.

A different paradigm has been proposed for rating data [58, 28] which is substantially based on the explicit modelling of the data generating process of the ordinal observations [44].

3 A generating process for rating data

Finite mixtures have been advanced by several Authors for analysing ordinal data (see [62] for a review) and most of them are motivated for improving fitting. They assume a convex combinations of probability distributions belonging to the same class of models or discretize continuous random variables to get appropriate probability mass functions for the observed ratings: in this vein, the reference to the Beta distribution with several variants is frequent: [55, 32, 65, 31, 70], among others. A different proposal arises from stochastic binary search algorithm [8].

In this scenario, psychological rationale related to the choice of ordinal scores leads to the introduction of CUB models [58]. A mixture distribution is specified to model the *propensity* to adhere to a meditated choice (formally described by a shifted Binomial random variable) and to a totally random one (described by a discrete Uniform distribution). Original motivations for the selection of these random variables were mostly based on a simplicity criterion [28]. However, the Binomial random model may be interpreted as a counting process of a sequential selection among the m categories and accounts for the genuine feeling of the response (appendix of [43]). Then, the Uniform distribution has been introduced as the most extreme among all discrete alternatives and accounts for the inherent uncertainty of the choice [39, 34, 63]. Independently, a psychological support to a direct modelling of a discrete component in the ordinal choices has been recently proposed by Allik [4].

Then, for a known $m > 3$, given a matrix $\boldsymbol{T}$ of values for subjects' covariates, a CUB model for the i-th subject implies both stochastic and deterministic relationships defined, respectively, as:

$$\begin{cases} Pr\,(R_i = r \mid \boldsymbol{T}) = \pi_i\, b_r(\xi_i) + (1-\pi_i)\, p_r^U\,, & r = 1,2,\dots,m; \\ \pi_i = \pi_i(\boldsymbol{\beta}) = \dfrac{1}{1+e^{-\boldsymbol{z}_i\boldsymbol{\beta}}}\,; \qquad \xi_i = \xi_i(\boldsymbol{\gamma}) = \dfrac{1}{1+e^{-\boldsymbol{w}_i\boldsymbol{\gamma}}}\,; & i = 1,2,\dots,n\,. \end{cases} \tag{3}$$

We set $b_r\,(\xi_i) = \binom{m-1}{r-1}\xi_i^{m-r}\,(1-\xi_i)^{r-1}$ and $p_r^U = 1/m$, $r = 1,2,\dots,m$, for the shifted Binomial and Uniform probability mass functions introduced to model feeling and uncertainty, respectively. Here, $(\boldsymbol{z}_i', \boldsymbol{w}_i')'$ are the values of the subjects' covariates extracted from the subjects' covariates data matrix $\boldsymbol{T}$. The acronym CUB stands for *C*ombination of a (discrete) *U*niform and (shifted) *B*inomial distribution.

Given the parameterization (3), it should be noticed that:

- the set of values $\boldsymbol{z}_i$ and $\boldsymbol{w}_i$ may be coincident, overlapped or completely different;

- whereas $1-\xi_i$ characterize the Binomial distribution of the mixture and are immediately related to the strength of the feeling component (they involve the modal value of the responses), the uncertainty parameters $1-\pi_i$ are just the weights of the Uniform distribution assumed for the indecision in the responses (thus, they are not involved in the specification of p_r^U);
- although psychological arguments are sufficient for motivating uncertainty as an important component of an ordinal choice, it is possible to show that uncertainty may be also generated when genuine Binomial choices manifest small variations in the feeling parameters.
- CUB models are able to detect a possible relationship between feeling and uncertainty when a common covariate is significant for those components.

Any one-to-one mapping $\mathbb{R}^p \leftrightarrow [0,1]$ between parameters and covariates is legitimate but the logistic function is to be preferred for simplicity and robustness properties [42]. Thus, the relationships:

$$logit\,(1-\pi_i)\,=\,-\boldsymbol{z}_i\boldsymbol{\beta}\,;\quad logit\,(1-\xi_i)\,=\,-\boldsymbol{w}_i\boldsymbol{\gamma};\qquad i=1,2,\ldots,n. \tag{4}$$

immediately relates uncertainty weights and feeling measures to subjects' covariates.

Although model (3) has been introduced with covariates, one may specify a CUB distribution without such a constraint:

- if $\pi = aver(\pi_i)$ and $\xi = aver(\xi_i)$ are some averages of the individual parameters, the parameters $(\pi,\,\xi)$ can be used to compare the responses to different items;
- for a given i-th subject, the features of the implied CUB model conditional to $(\boldsymbol{z}_i,\,\boldsymbol{w}_i)$ may be investigated by letting $\pi_i = \pi$ and $\xi_i = \xi$.

In this way, a CUB probability distribution is defined as:

$$Pr\,(R=r\mid\boldsymbol{\theta}) = \pi\,b_r\,(\xi) + (1-\pi)\,p_r^U,\qquad r=1,2,\ldots,m. \tag{5}$$

where $\boldsymbol{\theta} = (\pi,\xi)' \in \Omega\,(\boldsymbol{\theta}) = \{(\pi,\xi):\ 0<\pi\leq 1, 0\leq\xi\leq 1\}$ and the parameter space $\Omega\,(\boldsymbol{\theta})$ is the (open left) unit square.

A CUB model is *identifiable* [36] for any $m>3$ whereas $m=3$ implies a saturated model. Not all the well-defined discrete distributions are *admissible* for a CUB model as, for instance, bimodal probability functions. However, if multimodality is a consequence of latent classes then a CUB model with an explanatory variable related to those classes is adequately fitted according to (3).

A qualifying feature of CUB models is their visualization: given the one-to-one correspondence between the probability mass function (5) and a single point in the parameter space $\Omega\,(\boldsymbol{\theta})$, it is immediate to compare different items in terms of feeling and uncertainty; furthermore, by including subjects' covariates as in (3), the effect of each covariate on either components is visualized and response profiles can be identified. Finally, when CUB models estimated for different clusters –specified by time, space, circumstances and/or covariates– the changes in the responses (in terms of feeling and uncertainty) are immediately shown in the same parameter space. In this respect, also more refined devices –as, for instance, Scatter Plot of Estimates (*SPE*)– have been advanced [45].

Inferential issues have been solved by asymptotically efficient maximum likelihood methods performed by EM procedures [59] which are implemented in an R package [45]. Residual analysis may be conducted [29]. Alternative inferential procedures as Bayesian approaches [25] and permutation tests [9] have been also established.

Successful applications of CUB models have been obtained in different fields and for dissimilar objectives [21, 26, 10, 15, 11, 24, 13, 17]. In addition, formal and empirical comparisons with cumulative models have been extensively discussed on both real and simulated data sets [62]. Especially, the performance of the novel class of models has been checked when heterogeneity is a heavy component: in these circumstances, the better performance of CUB model as measured by *BIC* criterion, for instance, is the consequence of parsimony. Moreover, CUB models correctly assign heterogeneity to a possible uncertainty as related to explanatory covariates, whereas in POM heterogeneity is scattered over the categories by means of estimated cutpoints.

In this class of models, feeling causes no interpretation problem since it is immediately related to the attraction towards the item (modal values of the distribution and feeling parameters are strictly related); on the contrary, some doubts concern the nature of the uncertainty component [33], especially when no subjects' covariates are introduced. Surely, the measure conveyed by $1-\pi$ includes at least three meanings:

- *subjective indecision*: the measure $1-\pi_i$ is related to the personal indecision of the i-th respondent;
- *heterogeneity*: for CUB model without covariates, $1-\pi$ summarizes heterogeneity of the responses with respect to a given item [12, 16];
- *predictability*: in case of predicting ordinal outcomes, π is a direct measure of predictability of the model with respect to two extremes: a discrete Uniform ($\pi \to 0$) and shifted Binomial ($\pi = 1$) distributions.

Indeed, the special mixture proposed for ordinal data allows for a single rating to convey information on both *feeling* and *uncertainty* parameters:

i) the observed r is directly related to ξ since the probability of high/low values of r may increase or decrease with $1-\xi$;

ii) the observed r has also a bond with π since, for each score r, it increases or decreases the relative frequencies of: $\left|Pr(R=r)-\dfrac{1}{m}\right| \propto \pi$. Then, modifying the distance from the Uniform situation, each response modifies also the information concerning the uncertainty.

4 The family of CUB models

Previous models have been generalized in several directions to adequately fit rating data and exploit parsimony in different contexts. Among the many variants, we mention the inclusion of *shelter effect* [37], also with covariates (G*e*CUB) [43], the cases when almost one dominant preference is expressed (CUSH) [16] or some proposal with varying uncertainty (as VCUB [34] and CAUB [63] for *response style effects*) and also by including random effects in the components (HCUB [38] and RCUB [64]) and connections with fuzzy analysis of questionnaires [30].

Then, CUB models have been usefully applied in presence of "don't know" answers (DK-CUB [54, 41]), for detecting latent classes (LC-CUB [35]), for managing missing values in ordered data [24], for considering a MIMIC jointly with a CUB model [19] and for implementing model-based trees [18]. Noticeable is the discussion of similarities with log-linear models [56].

Further significant issues are the multi-items treatments, when also objects' characteristics are considered [61], and some genuine and promising multivariate proposals [7, 22, 23, 20].

An important advancement considers *overdispersion* as a further component to be examined. Thus, the feeling parameter is interpreted as a random variable specified by a Beta distribution: so, a (shifted) Beta-Binomial distribution is considered for the feeling component and CUBE models are introduced [39, 60, 46]. This class is truly general since it includes several previous specifications (as IHG [27], for instance); in addition, it adds higher flexibility to CUB models with an extra parameter. Since CUB are nested into CUBE models, relative testing is immediate.

The importance to take explicitly uncertainty into account in designing statistical models for rating data has received an increasing consideration (as in CUP models: [69]); indeed, the inclusion of classical approach in the emerging paradigms is a desirable opportunity.

In fact, due to the complex functioning of decision making process, two components are definitely dominant: i) a primary attraction/repulsion towards the item which is related to the personal history of the subject and his/her beliefs with respect to the item; ii) an inherent indecision derived by the circumstances surrounding the choice. This argument is grounded on empirical evidence since any choice is a human decision, where taste, emotion, sentiment are substantial issues which manifest themselves as a random component denoted as *feeling*. Moreover, choice changes over time, with respect to the environment, the modality (verbal, visual, numerical, etc.) and the tool (face-to-face response, telephone, questionnaire, online, etc.) by which data are collected: these circumstances cause an inherent indecision denoted as *uncertainty*.

Then, each respondent provides observations which are realizations of a choice R_i generated by mixing the feeling X_i and the uncertainty V_i. Formally, let m be the number of categories and $\boldsymbol{t}_i^{(X)} \in \boldsymbol{T}$, $\boldsymbol{t}_i^{(V)} \in \boldsymbol{T}$, where $\boldsymbol{T}$ includes the values of the selected covariates to explain feeling and uncertainty, respectively.

For any well-defined discrete distributions for feeling X_i and uncertainty V_i, respectively, a *GEneralized Mixture model with uncertainty* (GEM) is defined as follows:

$$Pr\,(R_i = r \mid \boldsymbol{\theta}) = \pi_i \, Pr\left(X_i = r \mid \boldsymbol{t}_i^{(X)}, \boldsymbol{\Psi}\right) + (1-\pi_i)\, Pr\,(V_i = r)\,, \tag{6}$$

for $i = 1,\dots,n$ and $r = 1,\dots,m$. Here, $\pi_i = \pi(\boldsymbol{t}_i^{(V)}, \boldsymbol{\beta}) \in (0,\,1]$ are introduced to weight the two components and $\boldsymbol{\Psi}$ includes all the parameters necessary for a full specification of X_i. The probability distribution of the *uncertainty* component V_i is assumed known over the support $\{1,\dots,m\}$ on the basis of *a priori* assumptions.

GEM models (6) are a very flexible class of models to interpret and fit rating data since they may specify: main flavour of the responses and their uncertainty/heterogeneity, overdispersion in respondents' feeling, presence of a possible inflated category and different probability distributions for given covariates: a distribution function for a latent variable to be discretized (classical approach); a probability mass function (novel approach). Both of them may include uncertainty in their specification for better fitting and interpretation.

Thus, GEM is the new paradigm since it is an *inclusive class of models for ordinal data* which encompasses both classical and mixture models in a unique representation where distributions are not compelled to belong to the exponential family.

5 Occam's razor: believe it or not

Occam's razor refers to one of the first principle in scientific research: "*Non sunt multiplicanda entia sine necessitate*" (Entities are not to be multiplied without necessity).

Cumulative models without covariates are saturated and this circumstance causes logical problems which have not been considered too much in the literature. The point is that estimable thresholds act as frequency parameters to perfectly fit the observed distribution; the introduction of explanatory variables modifies this perfect fit by adding variability to data. The paradox is that the starting point is a deterministic model and this structure becomes a stochastic one by adding further information.

Now, let $\mathfrak{E}(\mathscr{M} \mid A)$ be the measure of the explicative power of a model $\mathscr{M}$ when the information set is A: it may be log-likelihood, pseudo-R^2, fitting measure, divergence, BIC, deviance, etc. Then, let us denote by $\mathfrak{E}(\mathscr{M} \mid A \cup B)$ the same quantity when the information set is enlarged by a not-empty set B disjoint with A. Of course, $A \subset (A \cup B)$ and $\mathfrak{E}(\mathscr{M} \mid A) < \mathfrak{E}(\mathscr{M} \mid A \cup B)$ for any informative data set B. According to Occam's razor principle, a statistician should prefer a model based on $(A \cup B)$ if and only if the enlarged information improves the explicative power of A, otherwise B is a "*multiplicanda entia sine necessitate*".

In the class $\mathscr{M}$ of cumulative models, let A be the information set consisting of observed ratings and B the set of values of explanatory variables. Then, $\mathfrak{E}(\mathscr{M} \mid A \cup B) < \mathfrak{E}(\mathscr{M} \mid A)$ since any explicative power measure will reach its maximum in case of perfect coincidence of predictions and observations. Thus, under this perspective, the application of these models might appear controversial.

These considerations favour the novel paradigm since in the parametric mixture models modelling it is possible to fit and interpret data by splitting the explicative power of the probability structure from the added contribution given by explanatory covariates. This approach starts with a stochastic model and improves its performance by adding genuine information.

6 Conclusions

The core of the paper is to underline the paradigmatic nature of the modelling approach to rating data. According to Kuhn [49], a paradigm includes "the practices that define a scientific discipline at a certain point in time". Indeed, changes in paradigms have been recently proposed in order to better comprehend the subjective mechanism of a discrete selection out of a list of ordinal categories. The crisis of the established approach is more evident for clustering of items, visualization and interpretation purposes, especially in the big data era. Simulated and real data analysis support the usefulness of the chosen mixture modelling approach as a prominent alternative, thus fostering an integrated rationale.

Although some of them are useful, "all models are substantially wrong" and this aphorism is valid also for a novel paradigm. The substantive problem is to establish the starting point for further advances in order to achieve better models which should ever be improved. As statisticians, we must be aware of the role and importance of uncertainty in human decisions and CUB models may be considered as building blocks of more complex statistical specifications; above all, they act as a benchmark for more refined analyses.

An open-minded research process implies that a new paradigm is to be contrasted by the current one, being able to pursue all previous commitments and even solve new challenges. A convincing proposal easily captures new followers; it may be applied in different circumstances and former paradigms may be more and more critically considered. Unfortunately, since models are questionable by definition and given that human mind has inertial attitude towards novelties (it is a heavy effort to assume a new paradigm in consolidated procedures), the breaking point is often deferred in time.

Probably, time is not ripe yet for a paradigm shift. Nevertheless, a comprehensive family of models with appealing interpretation and parsimony features, a number of published papers supporting the new approach in different fields, an increasing diffusion of models which include uncertainty with a prominent role, the availability of a free software which effectively performs inferential procedures and graphical analysis are convergent signals that the prospective paradigm is slowly emerging.

Acknowledgements: Author thanks Stefania Capecchi, Francesca Di Iorio and Rosaria Simone for stimulating discussions and critical comments which improved the final version of the paper. This work has been partially supported by CUBREMOT project, University of Naples Federico II.

References

1. Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, 2^{nd} edition. Hoboken: Wiley.
2. Agresti, A. and Kateri, M. (2017). Ordinal Probability Effect Measures for Group Comparisons in Multinomial Cumulative Link Models. *Biometrics*, **73**, 214–219.
3. Agresti, A. and Tarantola, C. (2018). Simple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica*, 1–14. DOI: 10.1111/stan.12130.
4. Allik, J. (2014). A mixed-binomial model for Likert-type personality measures. *Frontiers in Psychology*, **5**, 371. Published online 2014 May 9. doi: 10.3389/fpsyg.2014.00371. PMCID: PMC4023022.
5. Anderson, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, **46**, 1–30.
6. Anderson, J.A. and Philips, P.R. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Journal of the Royal Statistical Society, Series C*, **30**, 22–31.
7. Andreis, F., Ferrari, P.A. (2013) On a copula model with CUB margins, *QdS. Journal of Methodological Applied Statistics*, **15**, 33–51.
8. Biernacki, C. and Jacques, C. (2016). Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing*, **26**, 929–943.
9. Bonnini, S., Piccolo, D., Salmaso, L. and Solmi, F. (2012). Permutation inference for a class of mixture models, *Journal Communications in Statistics - Theory and Methods*, **41**(16-17), 2879–2895.
10. Capecchi, S. (2015). Modelling the perception of conflict in working conditions. *Electronic Journal of Applied Statistics*, **8**(3), 298–311.
11. Capecchi, S., Endrizzi, I., Gasperi, F. and Piccolo, D. (2016). A multi-product approach for detecting subjects' and objects' covariates in consumer preferences. *British Food Journal*, **118**(3), 515–526.
12. Capecchi, S. and Iannario, M. (2016). Gini heterogeneity index for detecting uncertainty in ordinal data surveys. *METRON*, **74**, 223–232.
13. Capecchi, S., Iannario, M. and Simone, R. (2018). Well-being and relational goods: a model-based approach to detect significant relationships. *Social Indicators Research*, **135**(2), 729–750.
14. Capecchi, S. and Piccolo, D. (2014). Modelling the latent components of personal happiness. In: C. Perna and M. Sibillo (eds.), *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, Springer International, Springer-Verlag, pp.49–52.
15. Capecchi, S. and Piccolo, D. (2016). Investigating the determinants of job satisfaction of Italian graduates: a model-based approach. *Journal of Applied Statistics*, **43**(1), 169–179.
16. Capecchi, S. and Piccolo, D. (2017). Dealing with heterogeneity in ordinal responses. *Quality & Quantity*, **51**, 2375–2393.
17. Capecchi, S. and Simone, R. (2018). A proposal for a model-based composite indicators: experience on perceived discrimination in Europe. *Social Indicators Research*, DOI: 10.1007\ s11205-018-1849-9.
18. Cappelli, C., Simone, R. and Di Iorio, F. (2017). Growing happiness: a model-based tree, in: A. Petrucci and R. Verde (eds.). "Statistics and Data Science: new challenges, new generations", *Proceedings of the 2017 Conference of the Italian Statistical Society, Florence*, 261–266. ISBN (online): 978-88-6453-521-0.
19. Carpita, M., Ciavolino, E. and Nitti, M. (2018). The MIMIC-CUB model for the prediction of the economic public opinion in Europe, *Social Indicators Research*, 1–19. DOI:10.1007\ s11205-018-1885-4.

20. Colombi, R. and Giordano, S. (2016). A class of mixture models for multidimensional ordinal data, *Statistical Modelling*, **16**(4), 322–340.
21. Corduas, M. (2011a) Assessing similarity of rating distributions by Kullback-Liebler divergence. In: Fichet A. et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures, Studies in Classification, Data Analysis, and Knowledge Organization*. Springer-Verlag, Berlin, Heidelberg, pp.221–228.
22. Corduas, M. (2011b) Modelling correlated bivariate ordinal data with CUB marginals, *Quaderni di Statistica*, **13**, 109–119.
23. Corduas, M. (2015). Analyzing bivariate ordinal data with CUB margins, *Statistical Modelling*, **15**(5), 411–432.
24. Cugnata, F. and Salini, S. (2017). Comparison of alternative imputation methods for ordinal data. *Communications in Statistics. Simulation and Computation*, **46**(1), 315–330.
25. Deldossi, L. and Paroli, R. (2015). Bayesian variable selection in a class of mixture models for ordinal data: a comparative study, *Journal of Statistical Computation and Simulation*, **85**(10), 1926–1944.
26. Deldossi, L. and Zappa, D. (2014). A Novel Approach to Evaluate Repeatability and Reproducibility for Ordinal Data, *Communications in Statistics - Theory and Methods*, **43**(4), 851–866.
27. D'Elia, A. (2003). Modelling ranks using the inverse hypergeometric distribution, *Statistical Modelling*, **3**, 65–78.
28. D'Elia, A. and Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**, 917–934.
29. Di Iorio, F. and Iannario, M. (2012). Residual diagnostics for interpreting CUB models. *STATISTICA*, **LXXII**, 163–172.
30. Di Nardo, E. and Simone, R. (2016). CUB models: a preliminary fuzzy approach to heterogeneity. *Proceedings of the 48th Scientific Meeting of the Italian Statistical Society*, Eds. M. Pratesi and C. Perna (ISBN: 9788861970618), pp. 1–10.
31. Fasola, S. and Sciandra, M. (2015). New Flexible Probability Distributions for Ranking Data. In: *Advances in Statistical Models for Data Analysis*. Springer, Springer-Verlag, pp. 117–124.
32. Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics*, **31**, 799–815.
33. Golia, S. (2015). On the interpretation of the uncertainty parameter in CUB models, *Electronic Journal of Applied Statistical Analysis*, **8**(3), 312–328.
34. Gottard, A., Iannario, M. and Piccolo, D. (2014). Varying uncertainty in CUB models. *Advances in Data Analysis and Classifications*, **10**, 225–244.
35. Grilli, L., Iannario, M., Piccolo, D. and Rampichini, C. (2014). Latent class CUB models. *Advances in Data Analysis and Classifications*, **8**, 105–119.
36. Iannario, M. (2010). On the identifiability of a mixture model for ordinal data. *Metron*, **LXVIII**, 87–94.
37. Iannario, M. (2012a). Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications*, **21**, 1–22.
38. Iannario, M. (2012b). Hierarchical CUB models for ordinal variables. *Communications in Statistics. Theory and Methods*, **41**, 3110–3125.
39. Iannario, M. (2013). Modelling uncertainty and overdispersion in ordinal data. *Communications in Statistics. Theory and Methods*, **43**, 771–786.
40. Iannario, M., Manisera, M., Piccolo, D. and Zuccolotto, P. (2012). Sensory analysis in the food industry as a tool for marketing decisions. *Advances in Data Analysis and Classification*, **6**(4), 303–321.
41. Iannario, M., Manisera, M., Piccolo, D. and Zuccolotto, P. (2018). Ordinal data models for No-opinion responses in attitude surveys. *Sociological Methods & Research*, **6**(4), 1–27. DOI: 10.1177/0049124118769081.
42. Iannario, M., Monti, A.C., Piccolo, D. and Ronchetti, E. (2017). Robust inference for ordinal response models. *Electronic Journal of Statistics*, **11**, 3407–3445.
43. Iannario, M. and Piccolo, D. (2016a). A generalized framework for modelling ordinal data, *Statistical Methods and Applications*, **25**, 163–189.

44. Iannario, M. and Piccolo, D. (2016b). A comprehensive framework of regression models for ordinal data, *METRON*, **74**, 233–252.
45. Iannario, M., Piccolo, D. and Simone, R. (2018). CUB: a class of mixture models for ordinal data. R package version 1.1.2, http://CRAN.R-project.org/package=CUB.
46. Iannario M. and Simone R. (2017). Mixture models for rating data: the method of moments via Gröebner basis. *Journal of Algebraic Statistics*, **8**(2), 1–28.
47. Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementations Using R*. New York: Birkäuser/Springer.
48. Lang, J.B. (1996). Maximum Likelihood Methods for a Generalized Class of Log-Linear Models, *The Annals of Statistics*, **24**(2), 726–752.
49. Kuhn, T. (1962). *The Structure of Scientific Revolutions* (fourth edition 2012). University of Chicago Press, Chicago.
50. McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109–142.
51. McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models*, 2^{nd} edition. Chapman & Hall, London.
52. McFadden, K. (1978). Modeling the choice of residential location, in: Karlqvist, A. et al. *Spatial Interaction Theory and Residential Location*. Amsterdam: North-Holland, pp.75–76.
53. Manisera, M. and Zuccolotto, P. (2014a). Modeling rating data with Nonlinear CUB models. *Computational Statistics & Data Analysis*, **78**, 100–118.
54. Manisera, M. and Zuccolotto, P. (2014b). Modelling "Don't know" responses in rating scales. *Pattern Recognition Letters*, **45**, 226–234.
55. Morrison, D.G. (1979). Purchase intentions and purchase behavior. *Journal of Marketing*, **43**, 65–74.
56. Oberski, D.L. and Vermunt, J.K. (2015). The relationship between CUB and loglinear models with latent variables. *Electronic Journal of Applied Statistical Analysis*, **8**(3), 374–383.
57. Peterson, B. and Harrell, F.E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, **39**, 205–217.
58. Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85–104.
59. Piccolo, D. (2006). Observed information matrix in MUB models. *Quaderni di Statistica*, **8**, 33–78.
60. Piccolo, D. (2015). Inferential issues on CUBE models with covariates. *Communications in Statistics. Theory and Methods*, **44**, 5023–5036.
61. Piccolo, D. and D'Elia, A. (2008). A new approach for modelling consumers' preferences. *Food Quality and Preference*, **19**, 247–259.
62. Piccolo, D., Simone R. and Iannario, M. (2018). Cumulative and CUB models for rating data: a comparative analysis, *International Statistical Review*, forthcoming.
63. Simone R. and Tutz G. (2018). Modelling uncertainty and response styles in ordinal data, *Statistica Neerlandica*, doi: 10.1111/stan.12129
64. Simone, R., Tutz, G. and Iannario, M. (2018). Subjective heterogeneity and uncertainty for ordinal repeated measurements, *preliminary report*.
65. Taverne, C. and Lambert, P. (2014). Inflated Discrete Beta Regression Models for Likert and Discrete Rating Scale Outcomes, *arXiv:1405.4637v1*, 19 May 2014.
66. Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.
67. Tutz, G. and Berger, M. (2017). Separating location and dispersion in ordinal regression models, *Econometrics and Statistics*, **2**, 131–148.
68. Tutz, G. and Schauberger, G. (2010). Visualization of categorical response models: from data glyphs to parameter glyphs. *Journal of Computational and Graphical Statistics*, **22**, 156–177.
69. Tutz G., Schneider M., Iannario M. and Piccolo D. (2017). Mixture Models for Ordinal Responses to Account for Uncertainty of Choice. *Advances in Data Analysis and Classification*, **11**(2), 281–305.
70. Ursino, M. and Gasparini, M. (2018). A new parsimonious model for ordinal longitudinal data with application to subjective evaluations of a gastrointestinal disease, *Statistical Methods in Medical Research*, **27**(5), 1376–1393.

Statistical challenges and opportunities in modelling coupled behaviour-disease dynamics of vaccine refusal

Modellazione del rifiuto vaccinale come interazione delle dinamiche delle malattie infettive e del comportamento individuale: sfide ed opportunità per le discipline statistiche

Chris T. Bauch

Abstract Vaccine refusal has proven to be a persistent foe in efforts to protect populations from infectious diseases. One hypothesis about its origin posits a coupling between vaccinating behaviour and disease transmission: when infection prevalence is sufficiently low, individuals become complacent and vaccinating becomes less desirable, causing a decline in vaccine coverage and resurgence of the disease. This dynamic is being explored in a growing number of mathematical models. Here, I present a differential equation model of coupled behaviour-disease dynamics for vaccine-preventable paediatric infections, and I discuss previous research that has applied various statistical methodologies to parameterize and validate the model. I will show how methodologies such as model selection analysis and statistical learning, in conjunction with mechanistic modelling, can be used to test for the presence of phenomena related to coupled behaviour-disease dynamics during episodes of vaccine refusal. These phenomena include social learning and imitation, social norms, criticality, and coupling between vaccinating behaviour and disease prevalence. Some of these methodologies exploit new data sources such as online social media. I conclude that the study and modelling of vaccine refusal can greatly benefit from using mechanistic models informed by both traditional and state-of-the-art statistical methodologies.

Abstract *L'opposizione ai vaccini è un fenomeno persistente che indebolisce la capacità delle comunità di difendersi dalle malattie infettive. La spiegazione di base del rifiuto vaccinale postula l'esistenza di un'interazione tra decisione di vaccinare e trasmissione dell'infezione: quando la prevalenza di infezione è sufficientemente bassa il beneficio percepito dalla vaccinazione è a sua volta basso, il che a lungo andare causerà una discesa della copertura vaccinale, creando le premesse per una "risorgenza" della malattia. Queste dinamiche sono state analizzate in un numero crescente di studi modellistici. In questo lavoro presento un modello per in-*

Chris T. Bauch
Department of Applied Mathematics, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 3G1, e-mail: cbauch@uwaterloo.ca

fezioni pediatriche prevenibili da vaccino, come il morbillo, che accoppia le dinamiche dell'infezione con quelle delle decisioni vaccinali, e discuto le ricerche di tipo statistico che si sono occupate della parametrizzazione e validazione di queste classi di modelli. In particolare cerco di mettere in luce il ruolo delle metodologie statistiche (p.e., teoria dell'informazione, statistical learning, etc) per verificare la presenza di fenomeni di interazione tra decisioni individuali e prevalenza dell'infezione durante epoche di rifiuto dei vaccini. Lo studio di questi fenomeni, che includono l'apprendimento via imitazione, il ruolo delle norme sociali, e la presenza di vari effetti "critici", ha potuto in tempi recenti sfruttare le potenzialità dei dati forniti dai "social". Concludo discutendo l'importanza di combinare le metodologie statistiche tradizionali con le nuove tecniche della "data science" nello studio dell'opposizione ai vaccini.

Key words: behavioural epidemiology, vaccine refusal, coupled behaviour-disease systems, statistical learning, model selection

1 Vaccine-preventable infectious diseases: some background

Infectious diseases have long imposed a considerable burden on human health [36]. Improvements in nutrition, sanitation, hygiene and vaccines have considerably reduced this burden [9]. Smallpox was globally eradicated largely due to use of ring vaccination [20]. Even measles–which is highly transmissible–has been eliminated through universal vaccination programs in many countries, and the elimination of measles from the WHO Region of the Americas raises the possibility that even measles could one day be globally eradicated [16]. As our vaccine technologies and ability to administer them improve, universal vaccine access could become replaced as the primary barrier to elimination and eradication by vaccine refusal. In high- and low-income countries alike, vaccine refusal has led to resurgence of previously eliminated diseases such as measles [28], and has even delayed the eradication of polio by at least a decade [33].

Vaccines provide direct protection to vaccinated individuals by stimulating their immune response to specific antigens, but most vaccines also provide indirect protection for unvaccinated individuals by interrupting pathogen transmission [1, 31]. The transmission of infectious diseases can be mathematically modelled through compartmental models, which assume that individuals are divided into mutually exclusive compartments based on their infection status, and which tracks the transitions between these compartments through linear or nonlinear processes [1, 31]. For instance, the classic Susceptible-Infectious-Recovered (SIR) deterministic model with vaccination, births and deaths assumes that the population is divided into susceptible, infectious and recovered (immune) individuals, and is represented by:

$$\frac{dS}{dt} = \mu(1-p\varepsilon) - \beta SI - \mu S, \tag{1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I, \tag{2}$$

$$\frac{dR}{dt} = \mu p\varepsilon + \gamma I - \mu R. \tag{3}$$

where S is the proportion of the population that is susceptible, I is the proportion infectious, R is the proportion recovered, $\mu > 0$ is the mean birth/death rate, $\beta > 0$ is the mean transmission rate, $1/\gamma > 0$ is the mean duration of the infectious period, $0 \leq p \leq 1$ is the vaccine coverage, and $0 \leq \varepsilon \leq 1$ is the vaccine efficacy [31]. We assume that a proportion p of individuals are vaccinated at birth and, moreover, a proportion ε of those individuals were efficaciously immunised, entering the R compartment. The remaining proportion $1 - p\varepsilon$ enter the susceptible compartment at birth. We also assume that the birth rate equals the death rate and hence the population size is constant. Note that the R equation does not appear in the S or I equations, and since birth and death rates are equal, $R = 1 - S - I$ therefore we can characterize the system entirely in terms of S and I. This system has two equilibria:

$$E_1 = (S_1, I_1) = (1 - p\varepsilon, 0) \tag{4}$$

$$E_2 = (S_2, I_2) = \left(\frac{\gamma + \mu}{\beta}, \frac{\mu}{\gamma + \mu} \left(1 - p\varepsilon - \frac{\gamma + \mu}{\beta} \right) \right) \tag{5}$$

It is possible to show that the elimination threshold–the proportion of individuals who should be vaccinated in order to eliminate the infection–is given by

$$p_{crit} = \frac{1}{\varepsilon} \left(1 - \frac{1}{R_0} \right) \tag{6}$$

where the basic reproductive ratio $R_0 = \beta/(\gamma + \mu)$ is interpreted as the average number of secondary infections produced by a single infected individuals in an otherwise susceptible population [1]. (In the absence of vaccination, when $R_0 < 1$, the *disease-free equilibrium* E_1 is stable, but when $R_0 > 1$, the disease-free state loses stability and the system converges instead to the *endemic equilibrium* E_2.) When $p \geq p_{crit}$, the disease-free state E_1 is globally asymptotically stable, hence the infection is eliminated [31]. However, when $p < p_{crit}$, E_2 is globally asymptotically stable and the infection is endemic [31].

2 Coupled behaviour-disease systems

The SIR model, equations (1)-(3), represents a world where vaccine coverage is fixed at a specified level p. This is probably applicable where a decision-maker can assume that all eligible individuals will receive a vaccine. However, as we have noted in the first few paragraphs of this paper, vaccine refusal is an increasing problem. Therefore, we cannot always take it for granted that p will be fixed at sufficiently high to eliminate an infection from a population.

Multiple factors influence vaccine decision-making. However, several lines of evidence indicate that individuals are more likely to get vaccinated if (1) they perceive a risk of becoming infected (either due to an ongoing outbreak, possible future outbreaks, or due to a personal history of infection), (2) they perceive a risk of serious complications due to infection, and/or (3) they believe that the vaccine is safe and effective [25, 15, 13, 10, 44]. Indeed, we might have predicted the first factor from the SIR model: once $p = p_{crit}$ is obtained, the infection has been eliminated. In that case, any small real or perceived risk of suffering an adverse effect from the vaccine appears large compared to zero risk of being infected, thus the vaccine becomes undesirable and vaccine coverage can fall back below p_{crit}. Hence, we have a situation where individuals influence disease prevalence through their decision to become vaccinated, but disease prevalence in turn influences vaccine decision-making through individuals' desire to avoid becoming infected. We can therefore conceptualize this as a coupled behaviour-disease system, where disease dynamics and behavioural dynamics are combined into a single coupled system (Figure 1). An increase in vaccine coverage reduces infection prevalence (negative feedback), whereas an increase in infection prevalence boosts perception of infection risk and therefore boosts vaccine coverage (positive feedback), hence together they form a negative feedback loop leading to a stable state of endemic infection and intermediate vaccine coverage. Similar approaches to coupling human and natural systems have been taken up by ecologists and environmental scientists studying terrestrial and other ecosystems [34, 32, 2, 8, 30].

The importance of this interaction between infection and behaviour was not lost on the mathematical epidemiologists of the late twentieth century. Perhaps the earliest work to incorporate behaviour into epidemic dynamics was by Capasso and Serio, who proposed a model where the infection incidence term βSI in the SIR compartmental model is modified to take into account behavioural reactions to changing infection incidence during an outbreak [14]. Year later, the HIV/AIDS pandemic stimulated research on modelling the dynamics of core groups in infection transmission models, where recruitment into the core group depends on infection prevalence [27]. Economists and epidemiologists studied the problem from the perspective conflicts between individual interest and socially optimal approaches starting in the 1980s and 1990s as well [21, 11]. Subsequently, models of coupled

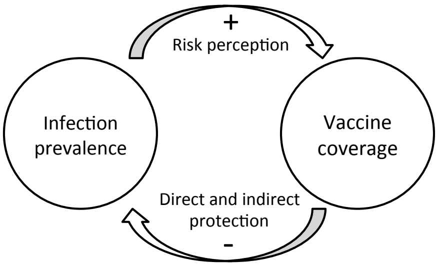


Fig. 1 Schematic diagram of a coupled behaviour-disease system. Increasing vaccine coverage reduces infection prevalence, which in turn causes a drop in vaccine coverage if the population becomes complacent due to lack of infections. The result is a negative feedback loop.

behaviour-disease interactions started becoming popular starting in the mid-2000s [7, 3, 26, 18, 19] (see Refs. [23, 6, 35, 45, 22, 46, 43] for reviews).

A game theoretical treatment of vaccine refusal provides a clear example of how adding adaptive human behaviour changes the predictions of epidemic models. For instance, following the approach of Ref. [5] for equations (1)–(3), it is possible to find the Nash equilibrium vaccine coverage p^* at which the payoff for an individual to vaccinate equals the payoff for an individual not to vaccinate. This turns out to be the vaccine coverage that should be exhibited by a population of rational, self-interested agents [5]. The expression is

$$p^* = \frac{1}{\varepsilon}\left(1 - \frac{1}{R_0(1 - r_v/r_i)}\right) \tag{7}$$

where r_v is the perceived risk of vaccine side effects and r_i is the perceived risk of infection complications. By comparing this expression to equation (6) for the elimination threshold, it is clear that $p^* < p_{crit}$ when $0 < r_v < r_i$. Due to the free-rider effect, it should therefore be impossible to eliminate an infection under a voluntary vaccination policy in a population of rational, self-interested agents [21, 5].

However, individuals are not rational self-interested agents when it pertains to vaccinating decisions [25, 15, 13, 10, 44]. For instance, peer imitation is an important feature of vaccinating behaviour that can be incorporated into epidemic models [17]. In the remainder of this paper we use a model that accounts for more realistic processes including imitation (social learning), social norms, and use of rule-of-thumb (heuristics) to determine infection risks [37]. The SIR equations are modified by replacing constant vaccine coverage p by a dynamic vaccine coverage x, where x is determined by a differential equation capturing how individuals learn vaccine opinions from others. A perfectly efficacious vaccine is assumed ($\varepsilon = 1$) which is a good approximation to the actual effectiveness for multi-course doses of most common pediatric vaccines. The model equations are:

$$\frac{dS}{dt} = \mu(1-x) - \beta SI - \mu S, \tag{8}$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I, \tag{9}$$

$$\frac{dR}{dt} = \mu x + \gamma I - \mu R, \tag{10}$$

$$\frac{dx}{dt} = \kappa x(1-x)\left[-\omega + I + \delta(2x-1)\right]. \tag{11}$$

In these equations, x is the proportion of the population favouring vaccination; $\omega \equiv r_v/mr_i$ controls the relative effects of the perceived risk of vaccine complications r_v, the perceived risk of infection complications r_i, and a proportionality constant m determining the perceived probability of becoming infected as a function of current infection prevalence $I(t)$ (the ‘rule of thumb’ for determining individual risk of becoming infected); δ is the strength of social norms; κ represents the social learning rate; and other parameters and variables are as in equations (1)–(3). In this model, individuals are either vaccinators or non-vaccinators and sample other individuals at a specified rate. If the other person being sampled is playing a different strategy and is receiving a higher utility, the given individual will change to that strategy with a probability proportional to the expected gain in utility (see Ref. [37] for details). When I is higher, more individuals will switch to a vaccinator strategy by imitating others. But when ω is higher due to higher perceived vaccine risk, or lower perceived risk of infection or infection complications, then more individuals will switch to a non-vaccinator strategy. The social norms term $\delta(2x-1)$ moves the population in the direction of whichever strategy is more popular, and thus captures peer pressure. We may remove the R equation since R does not appear in the other equations, hence dynamics can be described completely through (S,I,x).

The coupled behaviour-disease model, equations (8)–(11), exhibits a broad range of behaviour including 5 equilibria: a disease-free equilibrium where no one gets vaccinated, (1,0,0); a disease-free equilibrium where everyone gets vaccinated, (0,0,1); a disease-free equilibrium where part of the population are vaccinators; an endemic equilibrium where no one gets vaccinated; and an endemic equilibrium where part of the population are vaccinators. The model also exhibits stable limit cycles where x and I oscillate (Figure 2). The model is characterized in Ref. [37].

3 Challenges and opportunities for statistics in coupled behaviour-disease modelling

3.1 Parameterization and validation

Parameterizing and validating coupled behaviour-disease models present unique challenges on account of both their larger dimensionality and their coupling. Even with rich data on the epidemiological and sociological layers of the system in separation from one another (Figure 1), one is faced with the additional challenge of obtaining data on the coupling between the two layers–an aspect often ignored in

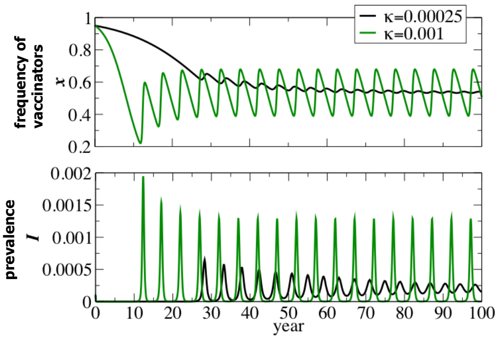


Fig. 2 Example dynamics of the coupled behaviour-disease model in equations (8)–(11). When κ is high, rapid social learning destabilizes the non-trivial equilibrium, causing infection prevalence and vaccine opinion to oscillats. Other parameters: $1/\mu = 50$ yrs., $1/\gamma = 10$ days, $\mathscr{R}_0 = 10$, $\kappa = 0.001$. Figure reproduced from Ref. [3].

traditional epidemiological and sociological studies. Accordingly, statistical inference [29], probabilistic uncertainty analysis [24], and model selection analyses such as use of information criteria [4, 37] are even more important for coupled behaviour-disease models than for sociological models or disease dynamic models in the absence of coupling. Information criteria can be particularly helpful because higher dimensionality and relative lack of data create hazards of over-fitting.

When the sample size n of a dataset is small compared to the number of parameters K being used to fit a model ($n/K < 40$), a variant of the Aikaike information criterion known as the corrected Aikaike information criterion (AICc) may be used (AICc = AIC + $2K(K+1)/(n-K-1)$) [12]. Using AICc, the baseline model in equations (8)–(11) has been compared to variant models lacking either the social learning mechanism (such that individuals switch opinions immediately as soon as the utility becomes more favourable, without learning the new opinion from peers); feedback from infection prevalence (such that infection prevalence is not a part of the utility function); or both mechanisms. The baseline model and its three variants were compared under five different forms for the possible time evolution of relative risk perception during a vaccine scare, $\omega = \omega(t)$ (see Figure 3, left-hand side). The 5 $\times$ 4 = 20 models were fitted using maximum likelihood to vaccine coverage and case notification data from the whole cell pertussis vaccine scare in the United Kingdom in the 1970s-80s (Figure 3) [4]. Comparing the AICc for these 20 models reveals interesting findings. Firstly, adding both social learning and prevalence feedback (i.e., using the baseline model) improved the AICc score and resulted in a better fit for most of the risk evolution curves (Figure 3, first column). (The comparison is worse under the bottom risk evolution curve, but this may be expected since an arbitrarily good AICc score can be obtained by adding enough degrees of freedom to the phenomenological curves that describes risk evolution.) Secondly, the variant model with infection prevalence feedback but no social learning (Figure 3, third column) exhibited highly unstable dynamics that both yields poor AICc scores and does not resemble vaccine coverage time series in any known system. This variant can be taken as a representative of *Homo economicus*–the idea that humans adopt Nash equilibria irrespective of social influences, while the baseline model including social learning could be taken as representative of *Homo socialis*–humans as social animals. Hence, this information theoretic exercise supports the notion that both infection prevalence feedback and social learning are important parts of explaining vaccine refusal in coupled behaviour-disease systems.

Model validation in coupled behavior-disease models can take the form of retrospectively testing of predictive power, for instance. For the pertussis vaccine scare, equations (8)–(11) were also fitted to the early years of the vaccine scare to see whether the model could predict the later years, and it was found that the first seven years of data provided enough information to predict the last ten years of the time series with good accuracy, despite the simplicity of the model [4]. However, the model did not show predictive power in retrospective analysis for the measles-mumps-rubella (MMR) autism vaccine scare in the UK during the 1990s-2000s. This might be due to the fact that measles dynamics were too irregular and stochastic throughout most of the MMR vaccine scare and thus a deterministic differential equation model might not be the right model to use in that situation. This is in contrast to the pertussis vaccine scare where large ‘deterministic’ outbreaks occurred.

3.2 Applications of statistical learning

The previous section described the need for data on both sociological and epidemiological subsystems. However, acquiring data for social subsystems–or sometimes even epidemiological subsystems–can be a challenge. The advent of digital data from sources like online social media has provided an alternative data source that can complement existing methods such as social surveys and case notifications [39]. Digital social data have been used not only to study online sentiment relating to vaccinating behaviour [40] but also to predict the outbreaks themselves, such as through

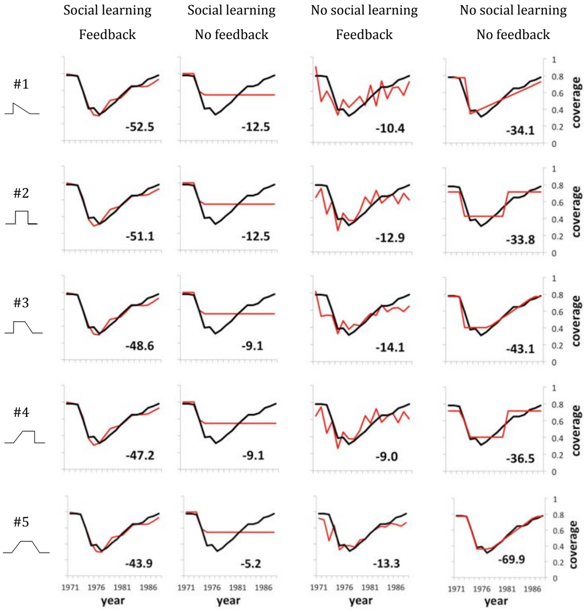


Fig. 3 Aikaike information criterion (AICc) scores and model fit of the coupled behaviour-disease model compared to variants for the UK whole pertussis vaccine scare. Pertussis vaccine coverage in the UK showed a steep decline over 5 years, from ∼80% to ∼30%, before commencing a slow return trajectory to high coverage levels (black lines). The red lines show best-fitting models for the baseline model (first column) and three variant models (second to fourth columns), for 5 risk evolution curves (rows, with form of curve shown on left). The numerical value in each subpanel is the AICc value for the fit: more negative AICc values correspond to better scores, i.e., the model exhibits a better balance of explanatory power with as few parameters as possible. Figure reproduced from Ref. [4].

symptom searches on the Internet [39]. Accordingly, it can help investigators obtain data on social dynamics, disease dynamics, and their coupled dynamics.

However, the amount of digital data is staggering compared to the size of most conventional epidemiological datasets and it cannot be manually processed. Hence, methods such as machine learning are required to analyze the data [42]. A particularly common type of machine learning is statistical learning, in which a computer is used to construct a probabilistic model of a dataset that exhibits statistical regularities. The statistical learning algorithm may use a training set in order to improve its probabilistic models. In Ref. [38], a statistical learning algorithm called a linear support vector machine (SVM) is used to study the 2014-15 California measles outbreak, in which vaccine refusal played a considerable role. The algorithm classified tweets about MMR vaccines into 'pro-vaccine', 'anti-vaccine', and 'other' categories. The number of pro-vaccine tweets were taken to correspond to x in equations (8)–(11) (see Ref. [38] for discussion of limitations of this assumption).

When the perceived vaccine risk ω increases sufficiently, equations (8)–(11) exhibit a tipping point beyond which vaccine uptake falls dramatically and the disease becomes endemic again [38]. The authors hypothesized that California was approaching this tipping point in the years before the relatively small Disneyland outbreak, and then receded from the tipping point afterward as vaccination became popular again. The approach and recession from a tipping point can be detected far in advance through changes in statistical indicators such as the lag-1 autocorrelation, coefficient of variation, and variance of a time series [41, 8]. The authors show that three empirical datasets based on SVM-classified tweets generally show expected trends, as predicted by equations (1)–(3) (Figure 4) [38].

This research suggests that vaccinating behaviour in coupled behaviour-disease systems can be classified as a critical phenomenon, and may exhibit early warning signals before widespread changes in behaviour such as the occurrence of large-scale vaccine scares. Interestingly, the coefficient of variation of the anti-vaccine time series of tweets shows a decline before the tipping point, instead of an increase as shown in all other time series tested and as might be expected from other research on tipping points in related systems [38]. The model predicts the same decline for the coefficient of variation of antivaccinators, however, illustrating the importance of using mechanistic models when interpreting statistical indicators.

4 Summary and Discussion

These examples illustrate the statistical challenges that emerge when parameterizing and validating coupled behaviour-disease models, as well as the synergies and opportunities that may arise when statistical and mechanistic approaches are used in conjunction. In the example of the model selection exercise, we saw how comparing the AICc scores of different models supported the notion that vaccinating behaviour is closer to the *Homo socialis* description than the *Homo economicus* description. In the example of using statistical learning to analyze tipping points, we saw how algorithms like linear support vector machines can be used to analyze vast amounts of online social media data to look for early warning signals of tipping points. The need for using mechanistic models to help interpret statistical patterns was shown by the decrease in coefficient of variation near a tipping point for anti-vaccinators, instead of the increase that is more commonly expected.

This research also suggests a more general approach by which mechanistic models can help us to make sense out of the bewildering complexity of social data. Dynamics generally simplify near tipping points, such that different types of complex nonlinear systems with highly divergent dynamics generally exhibit only the same restricted set of possible dynamics near a tipping point [41]. Hence, looking for evidence of tipping points in social media data is one possible way to begin moving from current predominantly descriptive statistical approaches to social media data, to statistically-informed mechanistic theories of social interactions.

These cases are only two selection-biased examples from a vast array of published work on how statistics and mathematics can be used together to study coupled behaviour-disease systems. For instance, a further opportunity not addressed here is the use of stochastic models, which maybe be particularly relevant close to the disease elimination threshold, or when rare but scary events perceived to be associated with vaccines or diseases occur. Moreover, new data sources such as online social media are already generating new statistical methodologies, and will likely continue to do so in the future. In conclusion, mechanistic approaches to coupled behaviour-disease dynamics of vaccine refusal can benefit from close attention to use of rel-

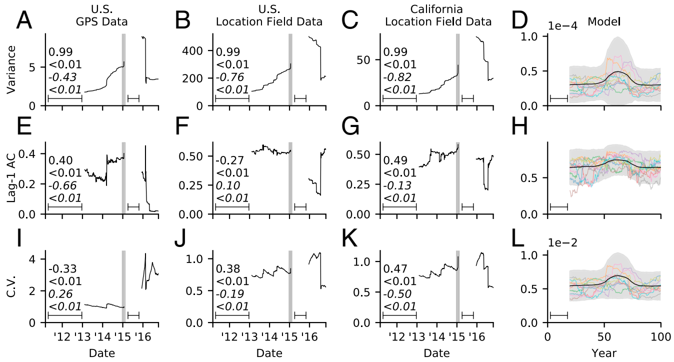


Fig. 4 Critical slowing down in pro-vaccine tweets near a tipping point, before and after Disneyland measles outbreak. (A-D) Variance, (E-H) lag-1 AC, and (I-L) coefficient of variation for (A, E, and I) US GPS-derived data, (B, F, and J) US Location Field-derived data, (C, G, and K) California Location Field-derived data data, and (D, H, and l) model predictions. The residual time series was used for variance and lag-1 AC. Kendall tau rank correlation coefficients are displayed before (regular font) and after (italic) the Disneyland peak with p values denoted by $<$. Window width used to compute rolling averages is indicated by line interval. Shaded region indicates outbreak time period. Model panels show indicators averaged across 500 stochastic model realizations (black), 2 SDs (shaded), and 10 example realizations (colored lines). Figure reproduced from Ref. [38].

evant empirical data for parameterization and validation, analyzed with both traditional and state-of-the-art statistical methods. Such empirically-driven modelling may help us tackle problems of vaccine refusal around the world, and perhaps even speed the global eradication of some vaccine-preventable infections.

Acknowledgements The author is grateful to Madhur Anand for fruitful discussions and to Piero Manfredi for helpful comments on the manuscript.

References

1. R. M. Anderson and R. M. May. *Infectious Diseases of Humans*. Oxford University Press, Oxford, 1991.
2. L.A. Barlow, J. Cecile, C.T. Bauch, and M. Anand. Modelling interactions between forest pest invasions and human decisions regarding firewood transport restrictions. *PLoS One*, 9(4):e90511, 2014.
3. C. T. Bauch. Imitation dynamics predict vaccinating behaviour. *Proc. R. Soc. Lond. B.*, 272:1669–1675, 2005.
4. C. T. Bauch and S. Bhattacharyya. Evolutionary game theory and social learning can determine how vaccine scares unfold. *PLoS Comp Biol*, 8(4):e1002452, 2012.
5. C.T. Bauch and D.J.D. Earn. Vaccination and the theory of games. *Proc. Natl. Acad. Sci. USA*, 101(36):13391–13394, 2004.
6. C.T. Bauch and A.P. Galvani. Social factors in epidemiology. *Science*, 342:47–49, 2013.
7. C.T. Bauch, A.P. Galvani, and D.J.D. Earn. Group interest versus self-interest in smallpox vaccination policy. *Proc. Natl. Acad. Sci. USA*, 100(18):10564–67, 2003.
8. C.T. Bauch, R. Sigdel, J. Pharaon, and M. Anand. Early warning signals of regime shifts in coupled human–environment systems. *Proc. Natl. Acad. Sci. USA*, 113:14560–67, 2016.
9. P. Bonanni. Demographic impact of vaccination: a review. *Vaccine*, 17:S120–S125, 1998.
10. S. Brien, J.C. Kwong, and D.L. Buckeridge. The determinants of 2009 pandemic A/H1N1 influenza vaccination: a systematic review. *Vaccine*, 30(7):1255–1264, 2012.
11. D.L. Brito and E. Sheshinski. Externalities and compulsory vaccinations. *Journal of Public Economics*, 45:69–90, 1991.
12. K.P. Burnham and D.R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.
13. H.S. Canning et al. Health care worker beliefs about influenza vaccine and reasons for non-vaccination: a cross-sectional survey. *Journal of Clinical Nursing*, 14(8):922–925, 2005.
14. V. Capasso and G. Serio. A generalization of the Kermack-McKendrick deterministic epidemic model. *Mathematical Biosciences*, 42(1-2):43–61, 1978.
15. G. Chapman and E. Coups. Predictors of influenza vaccine acceptance among healthy adults. *Preventive Medicine*, 29(4):249–262, 1999.
16. C.A. De Quadros, J.K. Andrus, et al. Feasibility of global measles eradication after interruption of transmission in the americas. *Expert review of vaccines*, 7(3):355–362, 2008.
17. A. d'Onofrio, P. Manfredi, and P. Poletti. The impact of vaccine side effects on the natural history of immunization programmes: an imitation-game approach. *J. Theor. Biol*, 273:63–71, 2011.
18. A. d'Onofrio, P. Manfredi, and E. Salinelli. Vaccinating behaviour, information, and the dynamics of sir vaccine preventable diseases. *Theor. Pop. Biol.*, 71(3):301–317, 2007.
19. Alberto d'Onofrio and Piero Manfredi. Information-related changes in contact patterns may trigger oscillations in the endemic prevalence of infectious diseases. *Journal of Theoretical Biology*, 256(3):473–478, 2009.

20. F. Fenner, D. A. Henderson, I. Arita, et al. Smallpox and its eradication. World Health Organization, Geneva, 1988.
21. P.E.M. Fine and J.A. Clarkson. Individual versus public priorities in the determination of optimal vaccination policies. *Am. J. Epidemiol.*, 124:1012–1020, 1986.
22. S. Funk, S. Bansal, C.T. Bauch, et al. Nine challenges in incorporating the dynamics of behaviour in infectious diseases models. *Epidemics*, 10:21–25, 2015.
23. S. Funk, M. Salathé, and V.A.A. Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *J. R. Soc. Interface*, page rsif20100142, 2010.
24. J.A. Gilbert, L.A. Meyers, A.P. Galvani, et al. Probabilistic uncertainty analysis of epidemiological modeling to guide public health intervention policy. *Epidemics*, 6:37–45, 2014.
25. K.P. Goldstein, T.J. Philipson, H. Joo, and R.S. Daum. The effect of epidemic measles on immunization rates. *JAMA*, 276(1):56–58, 1996.
26. T. Gross, C.J. Dommar D'Lima, and B. Blasius. Epidemic dynamics on an adaptive network. *Phys. Rev. Lett.*, 96(20):208701, 2006.
27. K.P. Hadeler and C. Castillo-Chávez. A core group model for disease transmission. *Mathematical biosciences*, 128(1-2):41–55, 1995.
28. N.A. Halsey and D.A. Salmon. Measles at Disneyland, a problem for all ages. *Annals of internal medicine*, 162(9):655–656, 2015.
29. D. He, E.L. Ionides, and A.A. King. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *J. Roy. Soc. Interface*, 2009.
30. K.A. Henderson, C.T. Bauch, and M. Anand. Alternative stable states and the sustainability of forests, grasslands, and agriculture. *Proc. Natl. Acad. Sci. USA*, 113(51):14552–14559, 2016.
31. H.W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.
32. C. Innes, M. Anand, and C.T. Bauch. The impact of human-environment interactions on the stability of forest-grassland mosaic ecosystems. *Scientific reports*, 3:2689, 2013.
33. A.S. Jegede. What led to the Nigerian boycott of the polio vaccination campaign? *PLoS Medicine*, 4(3):e73, 2007.
34. J. Liu, T. Dietz, S.R. Carpenter, et al. Complexity of coupled human and natural systems. *science*, 317(5844):1513–1516, 2007.
35. P. Manfredi and A. D'Onofrio. *Modeling the interplay between human behavior and the spread of infectious diseases*. Springer Science & Business Media, 2013.
36. C.J.L. Murray, A.D. Lopez, World Health Organization, et al. The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary. *Global burden of disease and injury series*, 1996.
37. T. Oraby, V. Thampi, and C.T. Bauch. The influence of social norms on the dynamics of vaccinating behaviour for paediatric infectious diseases. *Proc. R. Soc. B*, 281:20133172, 2014.
38. A.D. Pananos, T.M. Bury, C. Wang, et al. Critical dynamics in population vaccinating behavior. *Proc. Natl. Acad. Sci. USA*, page 201704093, 2017.
39. M. Salathe, L. Bengtsson, T.J. Bodnar, et al. Digital epidemiology. *PLoS Computational Biology*, 8(7):e1002616, 2012.
40. M. Salathé and S. Khandelwal. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology*, 7(10):e1002199, 2011.
41. M. Scheffer, J. Bascompte, W.A. Brock, et al. Early-warning signals for critical transitions. *Nature*, 461(7260):53, 2009.
42. B. Schölkopf and A.J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
43. F. Verelst, L. Willem, and P. Beutels. Behavioural change models for infectious disease transmission: a systematic review (2010–2015). *Journal of The Royal Society Interface*, 13(125):20160820, 2016.
44. P. Verger et al. Vaccine hesitancy among general practitioners and its determinants during controversies: a national cross-sectional survey in France. *EBioMedicine*, 2(8):891–897, 2015.
45. Z. Wang, M.A. Andrews, Z. Wu, et al. Coupled disease–behavior dynamics on complex networks: A review. *Physics of life reviews*, 15:1–29, 2015.
46. Z. Wang, C.T. Bauch, S. Bhattacharyya, et al. Statistical physics of vaccination. *Physics Reports*, 664:1–113, 2016.

3. Specialized Sessions

Bayesian Nonparametric Learning

Bayesian nonparametric covariate driven clustering

Un modello bayesiano nonparametrico per clustering in presenza di covariate

Raffaele Argiento, Ilaria Bianchini, Alessandra Guglielmi and Ettore Lanzarone

Abstract In this paper we introduce a Bayesian model for clustering individuals with covariates. This model combines the joint distribution of data in the sample, given the parameter and covariates, with a prior for this parameter. Here, the partition of the sample subjects is the parameter, and the prior we assume encourages two subjects to co-cluster when they have similar covariates. Cluster estimates are based on the posterior distribution of the random partition, given data. As an application, we fit our model to a dataset on gap times between recurrent blood donations from AVIS (Italian Volunteer Blood-donors Association), the largest provider of blood donations in Italy.

Abstract *Introduciamo un modello per il clustering di individui in presenza di covariate. Il modello combina la distribuzione congiunta dei dati, condizionatamente al parametero e alle covariate, con una prior per il parametro stesso, secondo l'approccio bayesiano. Qui il parametro è la partizione dei soggetti nel campione. La prior che introduciamo incoraggia due soggetti a stare nello stesso gruppo se hanno covariate simili. Applicheremo il nostro modello ad un dataset che riguarda le donazioni di sangue ripetute nel tempo.*

Key words: Bayesian nonparametrics, clustering, normalized completely random measures, regression models

Raffaele Argiento
Dipartimento ESOMAS, Università di Torino, e-mail: raffaele.argiento@unito.it

Ilaria Bianchini, Alessandra Guglielmi
Dipartimento di Matematica, Politecnico di Milano, e-mail: alessandra.guglielmi@polimi.it

Ettore Lanzarone
CNR-IMATI, Milano, e-mail: ettore.lanzarone@cnr.it

1 Introduction

In this paper, we introduce a Bayesian model for clustering individuals with covariates. Typically, in the Bayesian framework, there are two approaches for clustering. The first one assumes data to be independently distributed according to a mixture of parametric densities (possibly depending on covariates), with k components; k may be finite, with a prior on k, or infinite ($k = +\infty$), corresponding to a Bayesian nonparametric mixture. The second approach assumes a prior directly on the partition of sample subjects into clusters. Dirichlet process mixtures (DPMs), popularized by [4], are an example of Bayesian nonparametric models, where the weights of the infinite mixture are constructed using the stick-breaking representation (see [14]). Advantages of DPMs over finite mixtures with a prior on the number of components are the existence of generic Markov chain Monte Carlo (MCMC) algorithms for posterior inference to adapt to various applications and elegant mathematical properties of the nonparametric model; see [11] for a discussion on both models. Any DPM induces a random partition of the subject labels $\{1,2,\ldots,n\}$ through the values of the parameters $(\theta_1,\ldots,\theta_n)$ identifying the mixture component the observations are sampled from; in fact, since the mixing measure is almost surely discrete, there are ties in $(\theta_1,\ldots,\theta_n)$ with positive probability. Two subjects i and l share the same cluster if and only if $\theta_i = \theta_l$. In general, this relationship holds for any Bayesian nonparametric mixture model where the mixing measure is an almost surely discrete random probability measure.

In this paper, we follow the second Bayesian approach to clustering, which is more direct, since the random parameter of the model is the subject of the inference itself, i.e. the partition of the sample subjects. To define the model, we assign the joint conditional distribution of data in the sample, given the random partition, and the prior for this parameter. Corresponding cluster estimates are summaries of the posterior distribution of the random partition, given data. In particular, our prior depends on covariates, and we encourage, a priori, two subjects to co-cluster when they have similar covariates.

Our model is a generalization of the PPMx model proposed in [12], a product partition model with covariates information, extending the product partition model by [7]. This latter assumes a prior probability mass function for the random partition $\rho_n = \{A_1,\ldots,A_{k_n}\}$ proportional to the product of functions defined over the clusters, which are called cohesion. In [12], as well as in its generalizations, the cohesion function is restricted to be the one induced by the Dirichlet process, namely $c(S_j) = \kappa(n_j - 1)!$, where κ is a positive constant and n_j is the size of cluster A_j. However, this cohesion inherits "the rich-gets-richer" property of the Dirichlet process, i.e. sampled posterior partitions consist of a small number of large clusters, where new observations are more likely to join already-large clusters with probability proportional to the cardinality of the cluster.

To overcome this limitation, we introduce a more general class of PPMx models, with cohesion induced by normalized completely random measures (NormCRMs); see [13]. We perturb the product partition expression of the prior of the random partition via a similarity function g which depends on all the covariates associated to subjects in each cluster. Such g can be any non-negative function of some similarity measure guaranteeing that the prior probability that two items belong to the same cluster increases if their similarity increases. Note that we call our model "nonparametric", even though the random partition parameter has finite dimension; however its dimension is huge and increases with sample size. We are also able to build a general MCMC sampler to perform posterior analysis that does not depend on the specific choice of similarity. We test our model on a simulated dataset, and on a dataset on gap times between recurrent blood donations from AVIS (Italian Volunteer Blood-donors Association), the largest provider of blood donations in Italy. For the latter application, the problem is to find suitable methods to cluster recurrent event data, and predict a new recurrent event, using covariates describing personal characteristics of the sample individuals. In this paper, we model the sequence of gap times between recurrent events (blood donations) since donors are expected to donate blood not before a fixed amount of time imposed by the law. According to AVIS standard practice, the gap time between donations is the quantity that can be influenced for a better planning of the overall daily blood supply.

2 Bayesian covariate driven clustering

In a regression context, let $\boldsymbol{y}_i, \boldsymbol{x}_i$, $i = 1,\dots,n$ be the vector of responses and covariates for subject i, with $dim(\boldsymbol{y}_i) = n_i$; we assume $n_i = 1$ for all i in this section for greater clarity. We denote by $\boldsymbol{y}_j^*$ (and $\boldsymbol{x}_j^*$) the set of all responses y_i (and covariate vectors $\boldsymbol{x}_i$) in cluster A_j; the notation will be used later in the paper. We start from a family of regression models $f(\cdot;\boldsymbol{x},\theta)$, $\theta \in \Theta \subset \mathbb{R}^l$, and specify a hierarchical model that encourages subjects with similar covariates to be in the same cluster, using a data dependent prior for the random partition of data. We assume that data are independent across groups, conditionally on covariates and the cluster specific parameters; these are i.i.d from a base distribution P_0. Covariates enter both in the likelihood and the prior in our model. Concretely, we assume:

$$Y_1,\dots Y_n|\mathbf{x}_1,\dots,\mathbf{x}_n,\theta_1^*,\dots,\theta_{k_n}^*,\rho_n \sim \prod_{j=1}^{k_n} f(\mathbf{y}_j^*|\mathbf{x}_j^*,\theta_j^*) \tag{1}$$

$$\theta_1^*,\dots,\theta_{k_n}^*|\rho_n \overset{\text{iid}}{\sim} P_0 \tag{2}$$

$$p(\rho_n = \{A_1,\dots,A_{k_n}\}|\mathbf{x}_1,\dots,\mathbf{x}_n) \propto \int_0^{+\infty} D(u,n) \prod_{j=1}^{k_n} c(u,n_j) g(\mathbf{x}_j^*) du \tag{3}$$

where n_j denotes the size of cluster A_j, $g(\mathbf{x}_j^*)$ is the similarity function on cluster A_j such that $g(\emptyset) = 1$, and P_0 is a diffuse probability on the parameter space. Here $D(u,n)$ and $c(u,n_j)$ are defined as:

$$D(u,n) = \frac{u^{n-1}}{\Gamma(n)} \exp\{-\kappa \int_0^{+\infty} \left(1 - \mathrm{e}^{-us}\right) \rho(s) ds\}$$

where $\rho(ds) = \dfrac{1}{\Gamma(1-\sigma)} s^{-1-\sigma} \mathrm{e}^{-s} \mathbb{1}_{(0,+\infty)}(s) ds$ and

$$c(u,n_j) = \int_0^{+\infty} \kappa s^{n_j} \mathrm{e}^{-us} \rho(s) ds = \frac{\kappa\, \Gamma(n_j-\sigma)}{\Gamma(1-\sigma)} \frac{1}{(1+u)^{n_j-\sigma}}. \tag{4}$$

The intensity $\rho(ds)$, the positive parameter κ and the probability P_0 define a specific class of normalized completely random measures, called normalized generalized gamma process (NGG). Parameter σ has a deep influence on the clustering behavior. In particular, the discount parameter σ affects the variance: the larger it is, the more disperse is the distribution on the number of clusters. This feature mitigates "the rich-gets-richer" effect, typical of the Dirichlet process, leading to more homogeneous clusters. For more details on the behavior of σ in NGG's, see for instance [3], [10] and [2].

The likelihood specification in (1) may be any model, from simple regression models to the more complex models for gap times of recurrent events as in the AVIS application. The prior (3) is a perturbation of a prior for ρ_n, called product partition model (PPM) and introduced in [7]. When $g \equiv 1$, i.e. there are no extra information from covariates, the prior mass of each cluster depends only on its size through $c(u,n_j)$; when g is a proper function, the higher is the value of $g(\mathbf{x}_j^*)$, i.e. the more similar are covariates in cluster j, the higher is the prior probability mass of that cluster. This interpretation is justified since the prior $p(\rho_n|\mathbf{x}_1,\ldots,\mathbf{x}_n)$ in (3) can be equivalently written as

$$p(\rho_n|\mathbf{x}_1,\ldots,\mathbf{x}_n,u) \propto M(u) \prod_{j=1}^{k_n} c(u,n_j) g(\mathbf{x}_j^*) \tag{5}$$

for some prior density on the auxiliary variable $u > 0$. In other words, our model is an extension on the PPMx model, namely, it is a mixture of PPMx models (5).

It is quite natural to let the similarity to be a non-increasing function of the distance among covariates in the cluster, namely

$$\mathscr{D}_{A_j} = \sum_{i \in A_j} d(\mathbf{x}_i, \mathbf{c}_{A_j}) \tag{6}$$

where $\mathbf{c}_{A_j}$ is the centroid of the set of covariates in cluster j and d is a suitable distance function that we discuss later. Moreover, we assume $g(\mathscr{D}_{A_j}) := 1$ if the size of the set A_j is 1, i.e. $|A_j| = 1$.

The choice of the similarity is crucial, since this function controls how covariates affect the clustering. For this reason, we propose a list of similarity functions that proved to work reasonably well in practice; among those, here we list:

$g_A(\mathbf{x}_j^*;\lambda) = \mathrm{e}^{-t^{\alpha}}$, for $\alpha > 0$ ($\alpha = 0.5, 1, 2$), with $t = \lambda \mathscr{D}_{A_j}$;

$g_C(\mathbf{x}_j^*;\lambda)$ equals to $\mathrm{e}^{-t\log t}$ if $t \geq \frac{1}{e}$, or to $\frac{\mathrm{e}^{1/e-1}}{t}$ if $t < \frac{1}{e}$, where $t = \lambda \mathscr{D}_{A_j}$.

Here $\lambda > 0$ is a tuning parameter. The similarity function g_A is intuitive, i.e. its behaviour for $t \to +\infty$ is exponential. As far as the expression of g_C is concerned, we have proposed the expression $\mathrm{e}^{-t\log t}$ in such a way that, for large t, we contrast the asymptotic behavior of the cohesion function (4) induced by the NGG process. In fact, our model works well if the prior is not completely driven by covariates, because otherwise we could lose all the advantages of a Bayesian model-based clustering approach (e.g., uncertainty quantification, prediction, sharing information across clusters).

When the similarity is g_A, if we choose a very small λ, we concentrate the values of $\lambda\mathscr{D}_A$ around the origin, and hence we obtain similar values for $g_A(\cdot)$: in this case, the effect of covariate information on the prior of ρ_n will be very mild, since the range of values that the similarity can assume is very limited. A similar argument is valid for large values of λ. In conclusion, we calibrate λ such that g_A is evaluated in the range, say, $(0,3)$, for this particular choice of similarity.

In the applications we consider later, covariates will always be continuous or binary; categorical or ordinal covariates are translated into dummies. Hence, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors of covariates, $\mathbf{x}_j = (\mathbf{x}_j^c, \mathbf{x}_j^b)$, where $\mathbf{x}_j^c$ is the sub-vector of all the continuous covariates and $\mathbf{x}_j^b$ is the sub-vector of all binary covariates, we define the function d in (6) as

$$d(\mathbf{x}_1,\mathbf{x}_2) = d^c(\mathbf{x}_1^c,\mathbf{x}_2^c) + d^b(\mathbf{x}_1^b,\mathbf{x}_2^b), \tag{7}$$

where d^c is the Malahanobis distance between vectors, i.e. the Euclidean distance between *standardized* vectors of covariates, and d^b is the Hamming distance between vectors of binary covariates. The choice of the distance in (7) is not unique, but alternatives are among the subject of current research.

The way we define g does not increase the complexity of the algorithm for posterior inference. Indeed, we are able to devise a general MCMC sampler to perform posterior analysis that does not depend on the specific choice of similarity. The full-conditionals of the Gibbs sampler are relatively easy to implement in this case, since our algorithm generalizes the augmented marginal algorithm for mixture models in [9] and [5].

3 Simulated data

We apply model (1)-(3) in the regression context. Here $f(\mathbf{y}_j^*|\mathbf{x}_j^*,\theta_j^*)$ is the Gaussian regression model. We simulated a dataset of points $(y_i, x_{i1},\ldots,x_{ip})$ for $i=1,\ldots,n$, with $n=200$ and $p=4$. The last 2 covariates are binary and were generated from the Bernoulli distribution, while the first 2 were generated from Gaussian densities. The responses y_i's were generated from a linear regression model with linear predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}^0 := (\beta_0^0,\beta_1^0,\beta_2^0,\beta_3^0,\beta_4^0)$ and variance $\sigma_e^2 = 0.5$. We have generated 3 different groups by generating both covariates and responses from distributions with different parameters.

We run the Gibbs sampler algorithm to obtain 5,000 final iterations from the full posterior distribution, with initial burn-in of 2,000 and thinning of 10 iterations.

A-posteriori we classified all datapoints according to the optimal partition, under the different similarity functions; results are summarized in Table 1. By optimal partition we mean the realization, in the MCMC chain, of the random partition ρ_n which minimizes posterior expected value of the Binder's loss function with equal missclassification weights [8]. Observe that there are no missclassified data using similarity g_A, while 4% of data are missclassified using g_C, while the missclassification error increases to 16% if we do not assume covariate information ($g \equiv 1$).

Table 1 Missclassification rates for the simulated dataset

missclassif rate	g_A	g_C	$g \equiv 1$
	0%	4%	16%

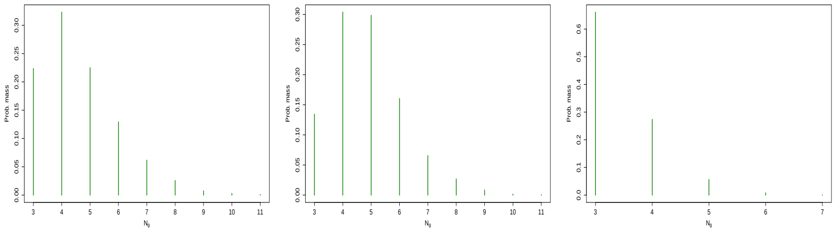


Fig. 1 Posterior distribution of K_n under g_A (left), g_C (center) and $g \equiv 1$ (right).

We computed the posterior distribution of K_n, the number of clusters, in the three cases; see Figure 1. Figure 2 displays the predictive distribution corresponding to covariates $\boldsymbol{x}_1$ of the first subject. The green vertical line corresponds to the actual observation y_1. It is clear that in the last case, i.e. when we do not include covariate information in the prior for the random partition, the predictive law is not able to distinguish to which of the three groups the subject belongs (thus, we have three peaks in the law). In cases A and C the predictive law exhibits only one main peak: the covariate information helps, in this case, in selecting the right group for the observation. This is also proved by the missclassification table above.

We underline that our prior encourages subjects with the same covariates to be in the same cluster, so that the posterior will generally allocate these subjects in the same cluster as well. On the other hand, if two subjects have very different covariates, our prior would classify them to different clusters, even if their responses are similar. However, the likelihood, i.e. the conditional distribution of data given the parameter, could correct the prior probability, if this is the case, and could allocate two subjects with different covariates to the same cluster.

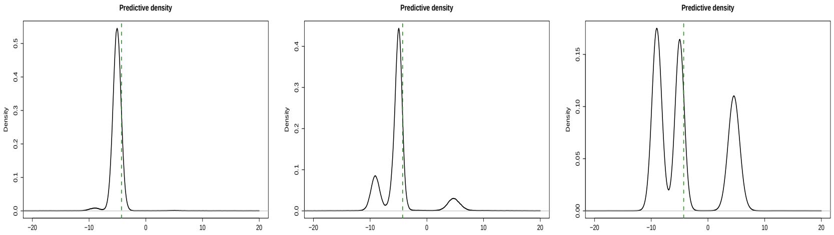


Fig. 2 Predictive distribution of Y_1 under g_A (left), g_C (center) and $g \equiv 1$ (right); vertical lines denote the true value

4 Blood donation data

Our data concern new donors of whole blood donating in a fixed time window in the main building of AVIS in Milano. Data are recurrent donation times, with extra information summarized in a set of covariates, collected by AVIS physicians. The last gap times are administratively censored for almost all the donors, except those having their last donation exactly on that date. The dataset contains 17198 donations, made by 3333 donors.

The statistical focus here is the clustering of donors according to the trajectories of gap times. Figure 3 reports the histogram of this variable (in the log-scale) for men and women. The skewness of these histograms can be explained since, according to

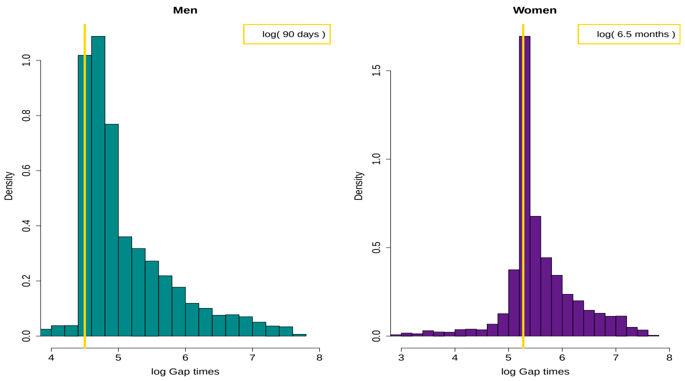


Fig. 3 Histogram of the logarithm of the observed gap-times divided according to gender, male (left) and women (right).

the Italian law, the maximum number of whole blood donations is 4 per year for men and 2 for women, with a minimum of 90 days between a donation and the next one. Note that the minimum for men is around 4.5 ($e^{4.5} \simeq 90$ days). For women, the distribution has a mode approximately in 5.3 in the log scale: this means 200 days, that corresponds to about 6 month and a half. Observe that donors may donate before the minimum imposed by law, under good donor's health conditions and the physician's consent.

We model gap times of successive donations as a regression model for recurrent gap times with two linear predictor terms, involving fixed-time and time varying covariates. The distribution of each gap time, in the log scale, is assumed to be skew-normal (see Figure 3); using parameterization in [6], we model the logarithm of the t-th gap time of donor i as Gaussian distributed. Cluster specific parameters are the intercept, the skewness parameter, and the variance of the response. We assume the prior for the random partition as in (5). Among donor's covariates, we include gender, blood type and RH factor, age, body mass index (BMI) and other information.

Preliminary analysis shows that, a posteriori, age and BMI (time-varying) have an effect on the gap time, as well as gender and RH factor. Details on the cluster estimate will be given during the talk.

References

1. Argiento, R., Guglielmi, A., Hsiao, C. K., Ruggeri, F., Wang, C.: Modeling the association between clusters of SNPs and disease responses. In: Müller, Mitra (eds.) Nonparametric Bayesian Inference in Biostatistics, pp. 115-134. Springer, New York (2015)
2. Argiento, R., Guglielmi, A., Pievatolo, A.; Bayesian density estimation and model selection using nonparametric hierarchical mixtures. Computational Statistics & Data Analysis, **54**, 816-832 (2010)
3. Argiento, R., Bianchini, I., Guglielmi, A.: Posterior sampling from ε-approximation of normalized completely random measure mixtures. Electronic Journal of Statistics, **10**, 3516-3547 (2016)
4. Escobar, M. D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American statistical association, **90**, 577–588 (1995)
5. Favaro, S., Teh, Y. W.: MCMC for normalized random measure mixture models. Statistical Science, **28**, 335–359 (2013)
6. Frühwirth-Schnatter, S., Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. Biostatistics, 11(2), 317-336.
7. Hartigan, J. A.: Partition models. Communications in statistics-Theory and methods, **19**, 2745–2756 (1990)
8. Lau, J. W., Green, P. J.: Bayesian model-based clustering procedures. Journal of Computational and Graphical Statistics, **16**, 526–558 (2007)
9. Lijoi, A., Prünster, I.: Models beyond the Dirichlet process. In: Hiort, Holmes, Müller, Walker (eds.). Bayesian nonparametrics, pp. 80–136. Cambridge University Press, Cambridge (2010)
10. Lijoi, A., Mena, R. H., Prünster, I.: Controlling the reinforcement in Bayesian nonparametric mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **69**, 715-740 (2007)
11. Miller, J. W., Harrison, M. T.: Mixture models with a prior on the number of components. Journal of the American Statistical Association, 1-17, 10.1080/01621459.2016.1255636 (2017)
12. Müller, P., Quintana, F., Rosner, G. L.: A product partition model with regression on covariates. Journal of Computational and Graphical Statistics, **20**, 260–278 (2011)

13. Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. Annals of Statistics, **31**, 560–585 (2003)
14. Sethuraman, J.: A constructive definition of Dirichlet priors. Statistica sinica, **4**, 639–650 (1994).

A Comparative overview of some recent Bayesian nonparametric approaches for the size of a population

Una rassegna comparativa su alcuni recenti approcci nonparametrici bayesiani per la stima della numerosità di una popolazione

Luca Tardella and Danilo Alunni Fegatelli

Abstract We review some recent approaches that have been used to address the difficult problem of estimating the unknown size of a finite population. We start from illustrating what types of inferential difficulties one should expect when no parametric assumption is made on the class of distributions for the distribution of counts of the number of multiple occurrences of the same unit when the observed counts are modelled in terms of Poisson mixtures. We then consider the problem from the species sampling model perspective where each unit is represented by the distinct species and a sequence of exchangeable unit label observations are available. We discuss the implementation of the alternative approaches with real datasets and we compare their performance with simulated data.

Abstract *In questo lavoro passiamo in rassegna alcuni recenti approcci bayesiani nonparametrici per stimare la numerosità incognita di una popolazione finita. Iniziamo dall'illustrare le difficoltà inferenziali che si devono affrontare quando nessuna assunzione parametrica restringe la classe di distribuzioni che regola il tasso medio di conteggio non negativo delle occorrenze multiple delle unità distinte nel campione. Consideriamo quindi il problema dalla prospettiva dei modelli per species sampling dove le unità corrispondono alle specie distinte e si assume una successione scambiabile etichette di specie osservabili. Si effettuano analisi comparative sull'implementazione dei diversi approcci su dati reali e sulla performance con simulati*

Key words: Poisson Mixture counts, Finite Population Size estimation, Species Sampling, Bayesian Nonparametrics

Luca Tardella
Sapienza Università di Roma, Piazzale Aldo Moro,5 (00185) Roma, e-mail: luca.tardella@uniroma1.it

Danilo Alunni Fegatelli
Sapienza Università di Roma, Piazzale Aldo Moro,5 (00185) Roma, e-mail: danilo.alunnifegatelli@uniroma1.it

1 Introduction

In this paper we consider the estimation of the total size of a finite population based on a single sample of individual detections. There are many instances in which such problem is of interest starting from the complete enumeration of elusive populations (Böhning and van der Heijden, 2009), software debugging to get the total number of errors (Lloyd et al., 1999) and the species richness problem in ecology (Bunge and Fitzpatrick, 1993; Chao and Bunge, 2002; Wang and Lindsay, 2005; Chao and Chiu, 2016) to measure and preserve biodiversity. We assume there are N distinct units in the population labelled as $i = 1,...,N$ which can be encountered (or detected) C_i times where each C_i is a non-negative integer. Indeed, only those units which are encountered at least once ($C_i > 0$) during the sampling stage are actually detected. Hence, if we denote with M the maximum number of multiple encounters (count) observed for a single unit, the positive frequencies of frequencies statistics $(f_1,...f_k,...,f_M)$, where f_k is the number of distinct units i which have been encountered exactly k times (i.e. for which $C_i = k$), provide a possibly incomplete enumeration $n = \sum_{k=1}^{M} f_k$ of the total population size N since $N = f_0 + n \geq n$. In fact, there are $f_0 > 0$ undetected units for which $C_i = 0$. There have been several attempts in the literature to address the problem of estimating N starting from modelling of the frequencies of frequencies distribution. In fact, this distribution is often determined by the nonparametric modelling of the individual encounter count data. One of the most flexible and general models commonly used in this setting is a mixture of Poisson distributions where the mixing distribution can be arbitrary. Indeed this setting has been dealt with both from the classical side with likelihood-based estimates Norris and Pollock (1998); Wang and Lindsay (2005) or Abundance-based Coverage Estimator (ACE), lower bounds and their variants (Chao and Lee, 1992; Mao, 2006; Mao et al., 2013) and from the Bayesian perspective Barger and Bunge (2010); Guindani et al. (2014). An alternative approach can be based on modelling the sequence of single encounters which is well known in the literature as species sampling model where it is assumed an exchangeable sequence of labels $X_1,...,X_j,...,X_s$ where a label uniquely and perfectly identifies each distinct species (to be understood as a distinct unit of the population) with $s = \sum_{k=1}^{M} k f_k$. In fact, in the species sampling terminology the word population size can be misleading since the total population size N of our original formulation corresponds to the total number of distinct species and is one of the main inferential objectives, whereas the total number of encounters of the different species corresponds to the number s of sequentially observed labels of the species sampling units. Exchangeability of labels ensures that the frequencies statistics $(f_1,...f_k,...,f_M)$ are the relevant statistics for inferring the population structure, including the probability of a new discovery and hence a possible assessment of the total number of distinct species N. Moreover, in most of the recent contribution in the Bayesian species sampling modelling (Lijoi et al., 2007; Arbel et al., 2017) the underlying population size, denoted with N in our setting, is indeed assumed to be infinite since the relative abundances of the population of species correspond to the random atoms of an almost surely discrete random probability measure belonging to the so-called Gibbs-type class. Noticeable exceptions of species sampling models assuming a finite population structure are considered in Gnedin (2010), Cerquetti (2011), Bissiri et al. (2013) and Zhou et al. (2017).

2 Alternative Bayesian Nonparametric approaches

In this section we briefly review the main features of some recent alternative approaches used to infer on the characteristics of a population which have individuals that have varying probability to be encountered in a sampling stage. More specifically we will consider the Dirichlet process mixture approach of Guindani et al. (2014) and the moment based approach in Alunni Fegatelli (2013). We also find it interesting to provide a comparative analysis with nonparametric Bayesian species sampling approach based on a two-parameter (α, β) Poisson-Dirichlet random measure as in Lijoi et al. (2007) with $0 \leq \alpha < 1$ and $\beta \geq -\alpha$. Indeed the comparability with the latter approach should take into account the structurally different underlying assumption on the population size although, in practice, the alternative approaches can be used to analyze the same real datasets. However, we will also consider a more appropriate comparison with a structurally different two-parameter (α, β) Poisson-Dirichlet random measure with $\alpha < 0$ and $\beta = -N\alpha$ for which the random measure has a finite support on N distinct units.

2.1 Dirichlet process mixture of Poisson counts

Guindani et al. (2014) propose to analyse observed positive counts of unique proteomic and genomic units with a semiparametric mixture of Poisson distributions in the presence of overdispersion and uncertainty on the true number of unique proteins or genes in a specific tissue (population). They assume the following hierarchical model: for a fixed population size N, any population unit $i = 1, ..., N$ is possibly detected according to $C_i|\lambda_i \sim Pois(\lambda_i)$, $\lambda_i|F \sim F$ and $F \sim DP(F_0, \tau)$ i.e. a Dirichlet process prior on an almost surely random discrete distribution F on the individual Poisson rate parameter λ_i. The Dirichlet process prior requires the specification of an expected distribution F_0 for λ and a positive total mass parameter τ regulating the concentration of the expected relative abundances corresponding to each unit of the population. They propose the use of a $Gamma(a, b)$ distribution for F_0. Indeed N is the main unknown parameter of interest and a prior distribution is needed. They acknowledge that the choice of the prior on N has a relevant impact and requires careful consideration. They starts arguing that in lack of genuine expert prior information a prior centered around the number of observed sequences n can provide a reasonable default choice. However, for simulation study purposes they implement their model with a uniform prior over a compact support.

2.2 Gibbs-type prior and nonparametric Bayesian species sampling

Lijoi et al. (2007) use a Bayesian nonparametric approach to evaluate the probability of discovering new species in the population conditionally on the number s of species recorded in a sample. The discovery probability represents a natural tool for a quantitative assessment of concepts such as species richness and sample coverage that is the proportion of distinct species present in the observed sample. In particular, they provide a way of estimating the proportion of yet unobserved species which is the complementary sample coverage fraction. However, we must point out from the outset that the species sampling setting and terminology should be carefully rephrased and understood within the original context described in Section 1. Indeed, in the species sampling model of Lijoi et al. (2007) an exchangeable sequence of s observable labels $X_1,...,X_j,...,X_s$ are sampled and the corresponding number n of distinct labels $X_1^*,...,X_n^*$ allows to compute the counts $C_{i,s} = \sum_{j=1}^{s} I(X_j = X_i^*)$ and those counts are sufficient statistics for inferring the sample coverage $1 - U_s = \sum_i \pi_i I(C_{i,s} = 0)$ conditionally on the observed labels. π_i's are the probability masses attached to each distinct label X_i^* which are in turn assumed to be random according to a Gibbs-type prior which selects a.s. discrete distributions with a countable support of distinct points corresponding to a countable subset of labels. In Favaro et al. (2012) and Arbel et al. (2017) an empirical Bayes approach is used to infer on the underlying parameters of the Gibbs-type prior and derive point estimate and interval estimate of the discovery probability of a new species in the Poisson-Dirichlet case. In fact, one could try to relate this discovery probability to the fraction of yet unseen species which can then be turned into an estimate of the total population size N using the relation $E[n] = N(1 - U_s)$. However, this could be rigorously justified only if the number of point masses, i.e. N is assumed to be finite almost surely which happens in the presence of Gibbs-type prior of fixed type $\alpha < 0$ according to Gnedin and Pitman (2005). However in this case one can more directly derive a fully Bayesian inference based on the conditional (on a fixed N) probability of the observed counts and the underlying mixing measure for the finite number of species N. To our knowledge such approach has not been considered in the literature. Indeed a recent attempt in the same direction has been put forward by Zhou et al. (2017) although with no emphasis on the estimation of N.

2.3 Moment-based mixtures of truncated Poisson counts

In Alunni Fegatelli (2013) and Alunni Fegatelli and Tardella (2018) a Bayesian nonparametric approach is proposed. It starts from highlighting that when a finite sample of counts are observed from a mixture of Poisson distributions with unconstrained mixing F for the Poisson rate parameter the sample basically carries information on the mixing F only for a finite number of features. More precisely, if M is the maximum number of observed counts, it depends only on the first M moments of a suitable finite measure Q representing a one-to-one reparameterization of F with the remaining moments of Q being completely unidentified by the sampling distribution (see details in Alunni Fegatelli and Tardella (2018)). In fact, the corresponding likelihood for N and the first M moments of Q, $(m_1(Q),....,m_M(Q))$ is

$$L(N,F;\mathbf{n}) = L(N,Q;\mathbf{n}) = L(N,m_1(Q),....,m_M(Q)) \propto \binom{N}{n} \prod_{k=0}^{T} \left[\frac{m_k(Q)}{k!}\right]^{n_k}$$

This yields the idea of working around a suitable moment-based approximation of the former likelihood which relies on a suitable truncation of the support of the rate parameter in a bounded interval $[0,u]$ and can then be used to derive a suitable default prior in terms of the Jeffreys rule for $(m_1,...,m_M)$ conditionally on N and u. A suitable Rissanen prior on the integer valued population size parameter N and a possible ad hoc choice of the prior on the truncation u complete the specification of the Bayesian model. Indeed it must be remarked that in Barger and Bunge (2010) alternative improper priors are derived as default priors from the reference and the Jeffreys prior approaches. The authors also provide justification for independent prior distributions for the parameter of interest N and the nuisance parameters of the stochastic abundance distribution.

3 Numerical Illustration

A simulation study was used to investigate on the frequentist performance of the three Bayesian nonparametric procedure. We considered the same 12 simulation settings proposed in Wang (2010) where the distribution of the species abundance varied from gamma, gamma mixture, lognormal and lognormal mixture, to discrete distributions, with the expected coverage of the sampled species ranging from 0.20 to 0.90. The corresponding results labelled s_1 through s_{12} are shown in Figure 1 where on the top row shows the average point estimates resulting from 100 simulated datasets for each simulation setting. In the other two rows the root mean square error and the coverage of equal tail 0.95 interval estimates are reported. Differently from the original work, where the estimates were evaluated only for $N = 1000$, we considered also a true population size of 10000. For a fairer comparison with respect to the Poisson-Dirichlet species sampling distribution we have also considered simulation settings s_{13}, s_{14} and s_{15} using a Poisson-Dirichlet structure with parameters $\alpha = -2$ and $\beta = -N\alpha$ and with observed sample coverage equal to 0.3, 0.5 and 0.8 respectively. For both values of N there wasn't a procedure resulting better than others for each simulation setting. However, overall, point and interval estimates for the moment-based method seemed to be the most stable, robust and with a smaller average (over all simulation settings) mean square error. Poor adaptivity of the Poisson-Dirichlet model might be explained by the fact that it indeed incorporates a smaller number of free parameters. We again stress on the fact that simulations were based on finite values of N. Hence, comparisons with the model proposed in Lijoi et al.

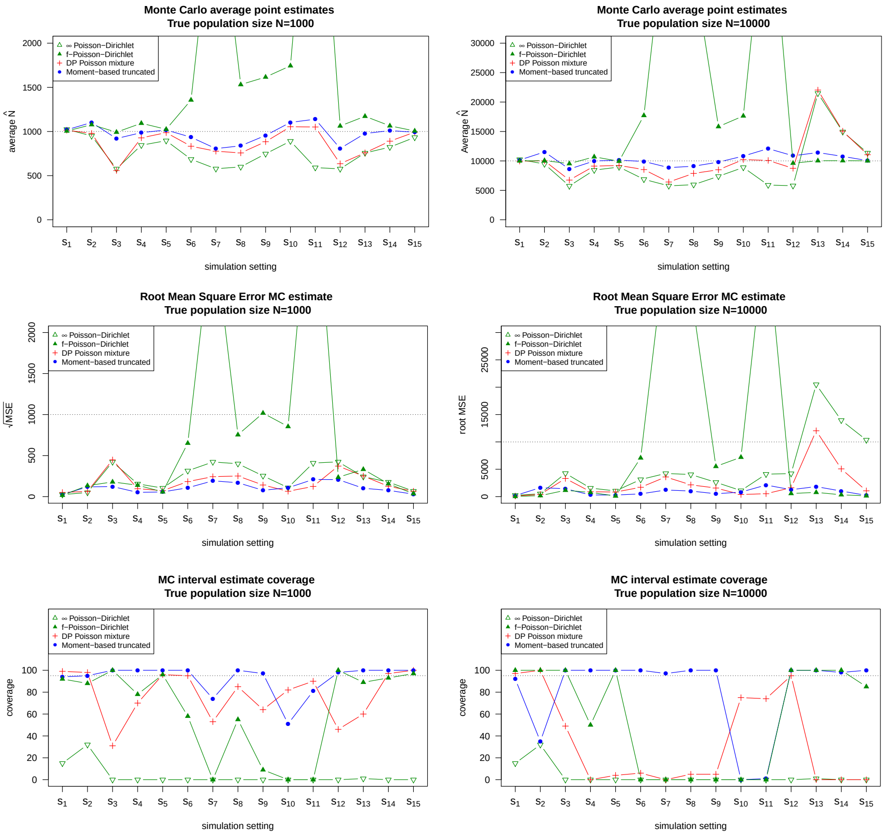


Fig. 1 Comparative performance of 4 alternative Bayesian methods

(2007) are admittedly unfair since in their approach they consider an infinite number of species. However one must also take into account that the comparison is of interest since that approach can be used in real applications where the population size cannot be indeed assumed to unbounded.

4 Concluding remarks

Particular attention to the performance of alternative methods in an inferential context where inference can be challenging and non standard asymptotics is expected. In this framework Bayesian posterior inference can be more sensitive to the prior input and this should be properly taken into account in the absence of genuine prior information. In this sense we believe that our simulation study conducted under alternative simulation settings as those proposed in the frequentist analysis of Wang (2010) could provide some practical suggestion for practitioners. Indeed we have also highlighted some possible drawbacks in using Bayesian nonparametric methods for species sampling based on Gibbs-type prior relying on the assumption of infinitely many species. A more extensive simulation study should be carried out to understand at what extent Bayesian nonparametric methods are sensible and numerically robust when the size of the underlying population grows. To our knowledge there is lack of theoretical understanding of this asymptotic behaviour even in the classical frequentist framework Wang (2010).

References

D. Alunni Fegatelli. *New methods for capture-recapture modelling with behavioural response and individual heterogeneity*. PhD thesis, Dipartimento di Scienze Statistiche, Sapienza Università di Roma, 2013.

Danilo Alunni Fegatelli and Luca Tardella. Moment-based bayesian poisson mixtures for inferring unobserved units. *arXiv preprint arXiv:https://arxiv.org/submit/2299462*, 2018.

J. Arbel, S. Favaro, B. Nipoti, and Y. W. Teh. Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*, 27:839–858, 2017.

Kathryn Barger and John Bunge. Objective Bayesian estimation for the number of species. *Bayesian Analysis*, 5(4):765–786, 2010.

P. G. Bissiri, A. Ongaro, and S. G. Walker. Species sampling models: consistency for the number of species. *Biometrika*, 100(3):771–777, 2013.

Dankmar Böhning and Peter G. M. van der Heijden. A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Ann. Appl. Stat.*, 3(2):595–610, 2009.

J. Bunge and M. Fitzpatrick. Estimating the number of species: A review. *Journal of the American Statistical Association*, 88:364–373, 1993.

Annalisa Cerquetti. Reparametrizing the two-parameter gnedin-fisher partition model in a bayesian perspective. pages 4678–4683, 8 2011.

Anne Chao and John Bunge. Estimating the number of species in a stochastic abundance model. *Biometrics*, 58(3):531–539, 2002.

Anne Chao and Chun-Huo Chiu. *Species Richness: Estimation and Comparison*, pages 1–26. American Cancer Society, 2016.

Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.*, 87(417):210–217, 1992.

S. Favaro, A. Lijoi, and I. Prunster. A new estimator of the discovery probability. *Biometrics*, 68(4):1188–1196, Dec 2012.

A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 12):83–102, 244–245, 2005.

Alexander Gnedin. A species sampling model with finitely many types. *Electron. Commun. Probab.*, 15:79–88, 2010.

Michele Guindani, Nuno Sepúlveda, Carlos Daniel Paulino, and Peter Müller. A Bayesian Semi-parametric Approach for the Differential Analysis of Sequence Counts Data. *Journal of the Royal Statistical Society Series C*, 63(3):385–404, 2014.

Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.

C. J. Lloyd, P. S. F. Yip, and Kin Sun Chan. Estimating the number of faults: efficiency of removal, recapture, and seeding. *IEEE Transactions on Reliability*, 48(4):369–376, Dec 1999.

Chang Xuan Mao. Inference on the number of species through geometric lower bounds. *Journal of the American Statistical Association*, 101(476):1663–1670, 2006.

Chang Xuan Mao, Nan Yang, and Jinhua Zhong. On population size estimators in the Poisson mixture model. *Biometrics*, 69(3):758–765, 2013.

James L. Norris and Kenneth H. Pollock. Non-parametric mle for poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics*, 5(4):391–402, Dec 1998.

Ji-Ping Wang. Estimating species richness by a Poisson-compound gamma model. *Biometrika*, 97(3):727–740, 2010. With supplementary data available online.

Ji-Ping Z. Wang and Bruce G. Lindsay. A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, 100(471):942–959, 2005.

Mingyuan Zhou, Stefano Favaro, and Stephen G Walker. Frequency of frequencies distributions and size-dependent exchangeable random partitions. *Journal of the American Statistical Association*, 112(520):1623–1635, 2017.

Logit stick-breaking priors for partially exchangeable count data

Distribuzioni a priori stick-breaking logistiche per dati di conteggio parzialmente scambiabili

Tommaso Rigon

Abstract Recently, Rigon and Durante (2018) discussed a Bayesian nonparametric dependent mixture model, which is based on a predictor-dependent stick-breaking construction. They provided theoretical support and proposed a variety of algorithms for posterior inference, including a Gibbs sampler. Their results rely on a formal representation of the stick-breaking construction, which has an appealing interpretation in terms of continuation-ratio logistic regressions. In this paper we review the contribution of Rigon and Durante (2018), and we extend their proposal to the case of partial exchangeability with count data. As an illustration of this methodology, we analyze the number of epileptic seizures of a single patient in a clinical trial.

Abstract *Recentemente, Rigon e Durante (2018) hanno discusso un modello di mistura bayesiano nonparametrico basato su una costruzione di tipo stick-breaking e dipendente da covariate. Gli autori hanno fornito sostegno teorico e hanno introdotto vari algoritmi per condurre inferenza a posteriori, incluso un campionamento di tipo Gibbs. I loro risultati si basano su una rappresentazione formale della costruzione stick-breaking, la quale ha un'interessante interpretazione in termini di regressioni logistiche sequenziali. In questo contributo, viene sintetizzata la proposta di Rigon e Durante (2018), e viene estesa la loro proposta nel caso parzialmente scambiabile con dati di conteggio. Per illustrare le loro metodologie, vengono analizzati il numero di attacchi epilettici di un singolo paziente durante un test clinico.*

Key words: Continuation-ratio logistic regression, Partial exchangeability, Poisson mixture model.

Tommaso Rigon
Dip. di Scienze delle Decisioni, Università Bocconi, e-mail: tommaso.rigon@phd.unibocconi.it

1 Introduction

Let $Y_1,\dots,Y_n \in \mathbb{N}$ be a collection of count response variables, each corresponding to a qualitative covariate $x_i \in \{1,\dots,J\}$, for $i = 1,\dots,n$. The observations $y_1,\dots,y_n$ from $Y_1,\dots,Y_n$ can be naturally divided in J distinct groups, given the covariates $x_1,\dots,x_n$. Our goal is to flexibly model the conditional distributions $\mathrm{pr}(Y = y \mid x = j) = p_j(y)$, for $j = 1,\dots,J$, under the assumption that each data point y_i is a conditionally independent draw from

$$(Y_i \mid x_i = j) \overset{\text{ind}}{\sim} p_j, \qquad i = 1,\dots,n, \tag{1}$$

where p_j denotes the probability mass function of the random variable $(Y_i \mid x_i = j)$. Within the Bayesian framework, assumption (1) is known as *partial exchangeability*, and model elicitation is completed by specifying a prior law Q_J for the vector of probability distributions: $(p_1,\dots,p_J) \sim Q_J$. Broadly speaking, the partial exchangeability assumption reflects an idea of homogeneity within the J subsets of observations but not across them. The prior measure Q_J governs dependence between groups, allowing borrowing of information across them. Maximal dependence, i.e. *exchangeability*, is attained if Q_J is such that $p_1 = \cdots = p_J$ almost surely, reflecting the prior belief that observations belong to the same latent population. Conversely, the case of full heterogeneity arises if each random probability distribution p_j is independent on $p_{j'}$ for any $j \neq j'$, implying that the distinct J groups share no information.

A common and flexible formulation for Q_J is given by mixture models of the form $p_j(y) = \int_\Theta K(y;\theta)\mathrm{d}P_j(\theta)$, where $K(y;\theta)$ denotes a known kernel function and $P_j(\theta)$ a random discrete mixing measure which is allowed to change across groups. In this paper we consider the class of predictor–dependent infinite mixture of Poisson distributions

$$p_j(y) = \int_\Theta \mathrm{Pois}(y;\theta)\mathrm{d}P_j(\theta) = \sum_{h=1}^{+\infty} \pi_{hj}\mathrm{Pois}(y;\theta_h), \qquad j = 1,\dots,J, \tag{2}$$

where $\pi_{hj} = \nu_{hj}\prod_{l=1}^{h-1}(1-\nu_{lj})$ are group–dependent mixing probabilities having a stick-breaking representation [11], and $\mathrm{Pois}(y;\theta)$ denotes the probability mass function of a Poisson with mean θ. Additionally, we assume that the atoms θ_h in (2) are independent and identically distributed (iid) draws from a diffuse baseline measure P_0, that is, $\theta_h \sim P_0$ independently for $h = 1,\dots,+\infty$ and independently on the weights π_{hj}. As for the stick-breaking weights ν_{hj}, we let

$$\mathrm{logit}(\nu_{hj}) = \alpha_{hj}, \qquad \text{with} \quad \alpha_h = (\alpha_{h1},\dots,\alpha_{hJ})^\intercal \overset{\text{iid}}{\sim} \mathrm{N}_J(\mu_\alpha,\Sigma_\alpha), \tag{3}$$

independently for every $h = 1,\dots,+\infty$. Specification of equations (2)-(3) can be regarded as a particular instance of the more general logit stick-breaking process (LSBP) of [8, 9], in which the covariate space is finite dimensional and with a Poisson kernel. As such, it inherits all the theoretical and computational properties of LSBP processes, some of which are reviewed in this manuscript.

Let us first consider an equivalent formulation of the logit stick-breaking Poisson mixture model of equations (2)-(3). Leveraging standard hierarchical representations of mixture models, the independent samples $y_1,\dots,y_n$ can be obtained equivalently from the random variable

$$\begin{aligned}(Y_i \mid G_i = h) &\sim \text{Pois}(\theta_h),\\ \text{pr}(G_i = h \mid x_i = j) = \pi_{hj} &= \nu_{hj}\prod_{l=1}^{h-1}(1-\nu_{lj}),\end{aligned} \tag{4}$$

for every unit $i = 1,\dots,n$, where $G_i \in \{1,2,\dots,+\infty\}$ is a categorical random variable denoting the mixture component associated to the i-th unit. Each indicator G_i has probability mass function $p(G_i \mid x_i = j)$ which can be written, after some algebraic manipulation, as

$$p(G_i \mid x_i = j) = \prod_{h=1}^{+\infty} \pi_{hj}^{\mathbb{1}(G_i=h)} = \prod_{h=1}^{+\infty} \nu_{hj}^{\mathbb{1}(G_i=h)}(1-\nu_{hj})^{\mathbb{1}(G_i>h)}, \tag{5}$$

for any $j = 1,\dots,J$. Equation (5) suggests an appealing interpretation of the stick-breaking weights ν_{hj} as the allocation probabilities to component h, conditionally on the event of surviving to the previous $1,\dots,h-1$ components, precisely

$$\nu_{hj} = \text{pr}(G_i = h \mid G_i > h-1, x_i = j), \tag{6}$$

for each $h = 1,\dots,+\infty$ and $j = 1,\dots,J$. This result, together with the prior formulation of equation (3), allows to interpret the stick-breaking construction (4) in terms of continuation–ratio logistic regressions [12]. This connection with the literature on sequential inference for categorical data is common to all the stick-breaking priors [e.g. 1, 8–10] and provides substantial benefits. Indeed, this characterization implies a simple sequential generative process for each membership indicator G_i and facilitates the implementation of a Gibbs sampler for posterior inference.

We briefly recall here the generative mechanism underlying equations (4), as described in [9], for the j-th group of observations. In the first step of the sequential process, each unit of the j-th group is either assigned to the first component $G_i = 1$ with probability ν_{1j} or to one of the subsequents with probability $1-\nu_{1j}$. If $G_i = 1$ the process stops, otherwise we draw another binary indicator, with probability ν_{2j}, to decide whether $G_i = 2$ or $G_i > 2$. The following steps proceed in a similar manner. Thus, we can reformulate each $\mathbb{1}(G_i = h) = \zeta_{ih}$, that is, the assignment indicator of each unit to the h-th component, in terms of binary sequential choices

$$\zeta_{ih} = z_{ih}\prod_{l=1}^{h-1}(1-z_{il}), \quad (z_{ih} \mid x_i = j) \sim \text{Bern}(\nu_{hj}), \tag{7}$$

for each $h = 1,\dots,+\infty$ and $j = 1,\dots,J$, where z_{ih} is a Bernoulli random variable representing the h-th sequential decision.

2 Theoretical properties

Let $(P_1,\dots,P_J)$ denote the vector of dependent random probability measures on $\mathbb{R}^+$ induced by the LSBP in equation (2). Thus, each random probability measure P_j can be represented as

$$P_j(\cdot) = \sum_{h=1}^{+\infty} \pi_{hj}\delta_{\theta_h}(\cdot), \qquad j = 1,\dots,J. \tag{8}$$

The stick-breaking representation of the π_{hj} implies that the random weights π_{hj} sum to 1 almost surely. Although this result is straightforward to derive, it should not taken for granted because of the analogy with [11], which leverages on peculiar characteristics of the Dirichlet process. This property is formalized in Proposition 1.

Proposition 1 (Rigon and Durante (2018)). *Let $(P_1,\dots,P_J)$ be a vector of random probability measures defined as in* (8) *and with stick-breaking weights defined as in* (3). *Then, $\sum_{h=1}^{+\infty}\pi_{hj} = 1$ almost surely for any $j = 1,\dots,J$.*

Proposition 2 provides some insights about the first two moments of the random vector $(P_1,\dots,P_J)$.

Proposition 2 (Rigon and Durante (2018)). *Let $(P_1,\dots,P_J)$ be a vector of random probability measures defined as in* (8)*, with stick-breaking weights defined as in* (3)*. Then, for any measurable set B, and for any $j = 1,\dots,J$ and $j' = 1\dots,J$, it holds*

$$\begin{aligned} \mathrm{E}\{P_j(B)\} &= P_0(B), \\ \mathrm{cov}\{P_j(B),P_{j'}(B)\} &= P_0(B)(1-P_0(B))\frac{\mathrm{E}(\nu_{1j}\nu_{1j'})}{\mathrm{E}(\nu_{1j})+\mathrm{E}(\nu_{1j'})-\mathrm{E}(\nu_{1j}\nu_{1j'})}. \end{aligned}$$

The expectation of $P_j(\cdot)$ coincides with the base measure $P_0(\cdot)$, which can be therefore interpreted as the prior guess for the mixing measure for any $j = 1,\dots,J$. Also, the variance of the random probability $P_j(B)$ can be recovered from the above covariance by letting $j = j'$. Unfortunately, the expectations in Proposition 2 are not available in closed form, although they can be easily computed numerically.

As noted by [9], the prior covariance between pairs of random probabilities is governed by the hyperparameters in specification (3) and it is always positive. From a modeling standpoint, this suggests that full heterogeneity among groups—using the terminology of Section 1—can be approximated for some suitable choice of the hyperparameters but it cannot be attained completely. A similar reasoning holds also for maximal dependence among groups which, again, arises only as a limiting case.

3 Posterior inference via Gibbs sampling

In this section we adapt the Gibbs sampler of [9] to the proposed infinite mixture model of Poisson distributions. Our approach exploits representation (4) and the continuation–ratio characterization of the logit stick-breaking prior. By conditioning on the latent indicators $G_1,\dots,G_n$, the model reduces to a set of standard conjugate updates—one for each mixture component—as long as the prior distribution of the atoms is

$$\theta_h \sim \mathrm{Gamma}(a_\theta, b_\theta), \qquad h = 1,\dots,+\infty.$$

Moreover, exploiting the sequential representation, posterior inference for the stick-breaking parameters α_h in (3) proceeds as in a Bayesian logistic regression in which the latent binary indicators z_{ih} in (7) play the role of the response variables, precisely

$$(z_{ih} \mid x_i) \sim \mathrm{Bern}\left(\{1+\exp(-\psi(x_i)^\intercal \alpha_h)\}^{-1}\right), \tag{9}$$

for each $i = 1,\dots,n$ and $h = 1,\dots,+\infty$, where $\psi(x_i) = \{\mathbb{1}(x_i = 1),\dots,\mathbb{1}(x_i = J)\}^\intercal$, and with $\mathbb{1}(\cdot)$ denoting the indicator function. To perform conjugate inference also for α_h, we adapt a recent Pólya-Gamma data augmentation scheme for logistic regression [7] to our statistical model, which relies on the following integral identity

$$\frac{e^{z_{ih}\psi(x_i)^\intercal\alpha_h}}{1+e^{\psi(x_i)^\intercal\alpha_h}} = \frac{1}{2}\int_{\mathbb{R}^+} f(\omega_{ih})\exp\left\{(z_{ih}-0.5)\psi(x_i)^\intercal\alpha_h - \omega_{ih}(\psi(x_i)^\intercal\alpha_h)^2/2\right\}\mathrm{d}\omega_{ih},$$

for each $i = 1,\dots,n$ and $h = 1,\dots,+\infty$, where $f(\omega_{ih})$ denotes the density function of a Pólya-gamma random variable $\textsc{pg}(1,0)$. Thus, the updating of α_h for any $h = 1,\dots,+\infty$ can be easily accomplished noticing that—given the Pólya-gamma random variables ω_{ih}—the contributions to the log-likelihood are quadratic in α_h and hence conjugate under the Gaussian priors (3). Moreover, the conditional density

$$f(\omega_{ih} \mid \alpha_h) = \frac{\exp[-0.5\{\psi(x_i)^\intercal\alpha_h\}^2\omega_{ih}]f(\omega_{ih})}{[\cosh\{0.5\psi(x_i)^\intercal\alpha_h\}]^{-1}},$$

defined for every $i = 1,\dots,n$ and $h = 1,\dots,+\infty$, is still a Pólya-Gamma random variable—and therefore conjugate—with updated parameters $f(\omega_{ih} \mid \alpha_h) \sim \textsc{pg}(1,\psi(x_i)^\intercal\alpha_h)$. This scheme allows posterior inference under a classical Bayesian linear regression.

Before providing a detailed derivation of the Gibbs sampler, we first describe a truncated version of the vector of random probability measure $(P_1,\dots,P_J)$, which can be regarded as an approximation of the infinite process. In line with [8–10], we develop a Gibbs sampler based on this finite representation, which has key computational benefits. We induce the truncation by letting $\nu_{Hj} = 1$ for some integer $H > 1$ and any $j = 1,\dots,J$, which guarantees that $\sum_{h=1}^{H}\pi_{hj} = 1$ almost surely. According to Theorem 1 in [9], the discrepancy between the two processes is exponentially decreasing in H, and therefore the number of components has not to be very large in practice to accurately approximate the infinite representation. Refer to [9] for a

Algorithm 1: Steps of the Gibbs sampler

```
begin
    [1] Assign each unit $i = 1, \dots, n$ to a mixture component $h = 1, \dots, H$;
    for $i$ from 1 to $n$ do
        Sample $G_i \in (1, \dots, H)$ from the categorical variable with probabilities
            $$\text{pr}(G_i = h \mid -) = \frac{\pi_{hx_i} \text{Pois}(y_i; \theta_h)}{\sum_{q=1}^{H} \pi_{qx_i} \text{Pois}(y_i; \theta_q)},$$
        for every $h = 1, \dots, H$.
    [2] Update the parameters $\alpha_h$, $h = 1, \dots, H-1$;
    for $h$ from 1 to $H-1$ do
        for every $i$ such that $G_i > h-1$ do
            Sample the Pólya-Gamma data $\omega_{ih}$ from $(\omega_{ih} \mid -) \sim \text{PG}(1, \psi(x_i)^\intercal \alpha_h)$.
        Given the Pólya-Gamma data, update $\alpha_h$ from the full conditional
            $$(\alpha_h \mid -) \sim \text{N}_J(\mu_{\alpha_h}, \Sigma_{\alpha_h}),$$
        having $\mu_{\alpha_h} = \Sigma_{\alpha_h}\{\Psi_h^\intercal \kappa_h + \Sigma_\alpha^{-1}\mu_\alpha\}$, $\Sigma_{\alpha_h} = \{\Psi_h^\intercal \Omega_h \Psi_h + \Sigma_\alpha^{-1}\}^{-1}$,
        $\Omega_h = \text{diag}(\omega_{i1}, \dots, \omega_{i\bar{n}_h})$ and $\kappa_h = (z_{i1} - 0.5, \dots, z_{i\bar{n}_h} - 0.5)^\intercal$, with $z_{ih} = 1$ if
        $G_i = h$ and $z_{ih} = 0$ if $G_i > h$.
    [3] Update the kernel parameters $\theta_h$, $h = 1, \dots, H$, in (2), leveraging standard results;
    for $h$ from 1 to $H$ do
        Sample the parameters $\theta_h$ from the full conditional
            $$(\theta_h \mid -) \sim \text{Gamma}\left(a_\theta + \sum_{i:G_i=h} y_i,\ b_\theta + \sum_{i=1}^{n} \mathbb{1}(G_i = h)\right).$$
```

more formal treatment. As a historical remark, the idea of truncating discrete nonparametric priors was firstly given by [6] and later developed by [3]. Theorem 1 in [9] is somehow the analogue of these results, for a class of models beyond exchangeability.

Let Ψ_h denote the $\bar{n}_h \times J$ predictor matrix in (9) having row entries $\psi(x_i)^\intercal$, for only those statistical units i such that and $G_i > h - 1$. The Gibbs sampler for the truncated representation of model (2) alternates between the full conjugate updating steps in Algorithm 1.

4 Illustration

As an illustration of the proposed methodology, we apply the LSBP Poisson mixture model to the `seizure` dataset, which was already analyzed in [13] and is available in the `flexmix` R package [2]. Data are extracted from a clinical trial conducted at the British Columbia's Children's Hospital, aiming to assess the effect of intravenous gamma globulin in reducing the daily frequency of epileptic seizures. Our dataset consists of daily myoclonic seizure counts (`seizures`) for a single subject, comprising a series of $n = 140$ days. After 27 days of baseline observation (`Treatment: No`), the subject received monthly infusions of intravenous gamma globulin (`Treatment: Yes`). The relative frequency of counts are shown in the upper plots of Figure 1, where the two groups—days with treatment and days without treatment—are compared.

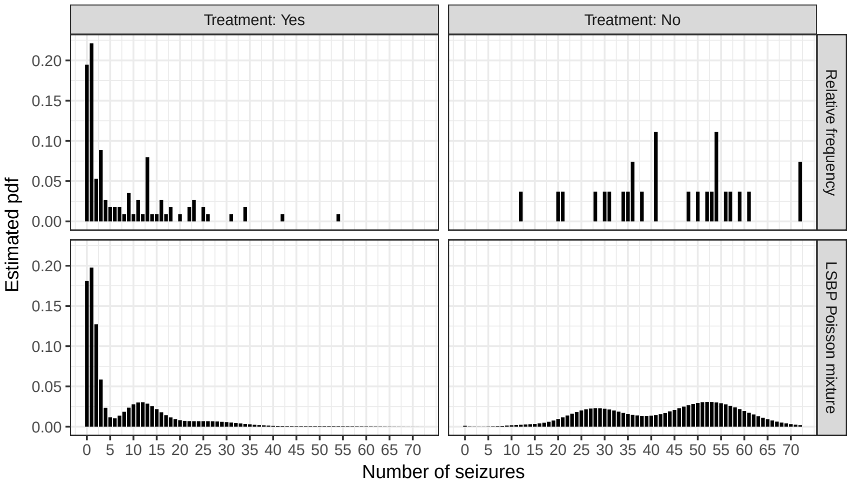


Fig. 1 Upper plots: for the two groups of observations (`Treatment: Yes` and `Treatment: No`), the relative frequencies of the daily number of myoclonic seizure counts are reported. Lower plots: for both groups of observations the (MCMC) posterior expectation of the probability mass function arising from the LSBP model is reported.

As evidenced by the raw frequencies displayed in Figure 1—which seem to present a multimodal structure—and consistent with the discussion in [13], a simple parametric formulation might be overly restrictive for the data at our disposal, thus motivating flexible representations. Additionally, regardless the effectiveness of the treatment, some form of dependence structure among observations from the two groups is expected, since they all refer to the same subject.

Consistent with these considerations, we model the `seizures` counts using the flexible mixture of Poissons described in Section 1. The number of groups is $J = 2$. As prior hyperparameters for the stick-breaking weights in (3), we set $\mu_\alpha = (0,0)$ and $\Sigma_\alpha = \mathrm{diag}(1000,1000)$, expressing the prior belief of a moderate amount of dependence among groups. As for the kernels parameters, we set $a_\theta = b_\theta = 0.05$, inducing a prior centered on 1 with a relatively large variance. Finally, we truncated the infinite mixture model choosing a conservative upper bound $H = 20$ for the number of mixture components. Although other hyperparameters settings are certainly possible, this would require a careful sensitivity analysis, which is beyond the scope of this paper. The Gibbs sampler in Section 3 was run for 50000 iterations, discarding the first 5000 draws as a burn-in period. The traceplots showed a satisfactory mixing and no evidence against convergence.

In the lower plots of Figure 1 we report the MCMC approximation of the posterior expectation for the probability mass function under the proposed LSBP Poisson mixture model, for the two groups. From this simple posterior check, it is apparent that our model is able to capture the main tendencies of the data. In particular, the proposed mixture model effectively resembles the multimodal behavior of the data.

References

[1] Dunson, D. B. and Park, J. H. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–323.

[2] Grün, B. and Leisch, F. (2008). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* **28**, 1–35.

[3] Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

[4] MacEachern, S. N. (1999). Dependent nonparametric processes. In *Proceedings of the Bayesian Section*, Alexandria, VA: American Statistical Association, pp. 50–55.

[5] MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Department of Statistics, Ohio State University.

[6] Muliere, P. and Tardella, L. (1998) Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics* **26**, 283–297.

[7] Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association* **108**, 1339–1349.

[8] Ren, L., Du, L., Carin, L. and Dunson, D. B. (2011). Logistic stick-breaking process. *Journal of Machine Learning Research* **12**, 203–239.

[9] Rigon, T. and Durante, D. (2018). Tractable Bayesian density regression via logit stick-breaking priors. *arXiv:1701.02969.*

[10] Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**, 145–178.

[11] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

[12] Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics & Data Analysis* **11**, 275–295.

[13] Wang, P., Puterman, M. L. and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics* **52**, 381–400.

BDsports - Statistics in Sports

A paired comparison model for the analysis of on-field variables in football matches

Gunther Schauberger and Andreas Groll

Abstract We use on-field variables from football matches in the German Bundesliga and connect them to the sportive success or failure of the single teams in a paired comparison model where each match in a Bundesliga season is treated as a paired comparison between the two competing teams. We propose an extended paired comparison model that extends the classical Bradley-Terry model to ordinal response variables and includes different types of covariates. We apply penalized likelihood estimation and use specific L_1 penalty terms for fusion and selection in order to reduce the complexity of the model and to find clusters of teams with equal covariate effects. The proposed model is a very general one and can easily be applied to other sports data or to data from different research fields. We apply the model to data from the latest season of the German Bundesliga.

Key words: Bundesliga, Paired Comparison, BTLLasso, Penalization

1 Introduction

In modern football, various variables as, for example, the distance a team runs or its percentage of ball possession, are collected throughout a match. However, there is a lack of methods to make use of these on-field variables simultaneously and to connect them with the final result of the match. We propose to treat each football match as a paired comparison between the two competing teams and to analyse the results of football data by an extended paired comparison model.

Gunther Schauberger
Chair of Epidemiology, Technical University Munich, e-mail: gunther.schauberger@tum.de,
Department of Statistics, LMU Munich

Andreas Groll
Faculty of Statistics, Technische Universität Dortmund e-mail: groll@statistik.tu-dortmund.de

Paired comparisons occur if two objects out of a set of objects are compared with respect to an underlying latent trait. In the case of football matches in national leagues all teams from the respective league are considered to be these objects. Football matches can be treated as paired comparisons between two teams where the playing abilities of the teams represent the underlying latent traits that are compared.

Our main goal is to set up a paired comparison model that is able to incorporate so-called on-field variables as covariates. In general, if covariates are to be considered in paired comparison, one has to distinguish between subjects and objects of the paired comparisons. A covariate can either vary across the subjects or the objects of a paired comparison, or, as in our case, both over subjects and objects. In football matches, the teams are the objects while a single match can be considered to be the subject that conducts the comparison between the two objects/teams. If one considers a variable like the percentage of ball possession a team has in a specific match, this variable varies both from team to team and from match to match. Therefore, in our application subject-object-specific covariates are considered. After all, the proposed model could in principle consider all three types of variables simultaneously.

The Bradley-Terry model (Bradley and Terry, 1952) is the standard model for paired comparison data. Assuming a set of objects $\{a_1,\dots,a_m\}$, in its most simple form the Bradley-Terry model is given by

$$P(a_r \succ a_s) = P(Y_{(r,s)} = 1) = \frac{\exp(\gamma_r - \gamma_s)}{1+\exp(\gamma_r - \gamma_s)}. \qquad (1)$$

One models the probability that a certain object a_r dominates or is preferred over another object a_s, $a_r \succ a_s$. The random variable $Y_{(r,s)}$ is defined to be $Y_{(r,s)} = 1$ if a_r dominates a_s and $Y_{(r,s)} = 0$ otherwise. The parameters γ_r represent the attractiveness or strength of the respective objects. In football matches, the random variable $Y_{(r,s)}$ which represents the paired comparison between a_r and a_s needs to have at least $K = 3$ possible categories instead of two because in football one needs to account for the possibility of draws. However, if one distinguishes, for example, clear wins and losses from wins and losses with only one goal difference one could also use $K = 5$ categories. In general, for the case of ordered responses $Y_{(r,s)} \in \{1,\dots,K\}$ the model is extended accordingly to

$$P(Y_{(r,s)} \le k) = \frac{\exp(\theta_k + \gamma_r - \gamma_s)}{1+\exp(\theta_k + \gamma_r - \gamma_s)}, \quad k = 1,\dots,K, \qquad (2)$$

which essentially corresponds to the generalization from a binary logistic regression model to a cumulative logistic regression model. In our application, the strength parameters γ_r represent the playing abilities of the teams.

In general, for the ordinal paired comparison model (2) it is assumed that the response categories have a symmetric interpretation so that $P(Y_{(r,s)} = k) = P(Y_{(s,r)} = K-k+1)$ holds. Therefore, the threshold parameters should be restricted by $\theta_k = -\theta_{K-k}$ and, if K is even, $\theta_{K/2} = 0$ to guarantee for symmetric probabilities. The threshold for the last category is fixed to $\theta_K = \infty$ so that $P(Y_{(r,s)} \le K) = 1$ will hold.

The probability for a single response category can be derived from the difference between two adjacent categories, $P(Y_{(r,s)} = k) = P(Y_{(r,s)} \leq k) - P(Y_{(r,s)} \leq k-1)$. To guarantee for non-negative probabilities of the single response categories one restricts $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_K$.

When football matches are considered as paired comparisons one has to consider that so-called order effects are possible. To some extent, order effects contradict the assumption of symmetric response categories which was described above. When order effects are present, the order of the two objects (i.e. which object is the first-named object and which is the second-named object) is not random and possibly has an influence on the outcome. In football (especially in national leagues) this is obviously the case as the first-named object is the home team and, hence, usually has a home advantage over the (second-named) away team. To include such an order effect in model (2) we extend the model to

$$P(Y_{(r,s)} \leq k) = \frac{\exp(\delta + \theta_k + \gamma_r - \gamma_s)}{1 + \exp(\delta + \theta_k + \gamma_r - \gamma_s)}, \quad k = 1, \ldots, 5, \tag{3}$$

where δ represents an order effect. In football matches, this parameter represents the home effect (or home advantage if positive). It is possible to assume a global home effect δ which is equal for all teams or team-specific home effects δ_r.

2 Bundesliga Data

The main goal of this work is to analyze if (and which) on-field variables that are collected throughout a match are associated to the final result of football matches. In total, our data set contains all the following variables separately for each team and each match:

Distance	Total amount of km run
BallPossession	Percentage of ball possession
TacklingRate	Rate of tacklings won
ShotsonGoal	Total number of shots on goal
Passes	Total number of passes
CompletionRate	Percentage of passes reaching teammates
FoulsSuffered	Number of fouls suffered
Offside	Number of offsides (in attack)

The data were collected from the website of the German football magazin kicker (http://www.kicker.de/). Exemplarily, Table 1 shows the collected data for the opening match of the season 2016/17 between Bayern München and Hamburger SV.

Bayern München		Hamburger SV
Goals	5 : 0	Goals
Shots on goal	23 : 5	Shots on goal
Distance	108.54 : 111.28	Distance
Completion rate	90 : 64	Completion rate
Ball possession	77 : 23	Ball possession
Tackling rate	52 : 48	Tackling rate
Fouls	10 : 12	Fouls
Offside	3 : 0	Offside

Table 1 Illustrating table for original data situation showing data for the opening match in season 2016/17 between Bayern München and Hamburger SV. Source: http://www.kicker.de/

3 A Paired Comparison Model for Football Matches Including On-field Variables

When using a paired comparison model for football matches the standard Bradley-Terry model needs to be extended in several ways. In model (3) we already extended the Bradley-Terry model to handle both an ordinal response (in particular draws) and home effects. Now the model is further extended to incorporate on-field variables, which in the context of paired comparisons are considered as subject-object-specific variables. We propose to use the general model for ordinal response data $Y_{i(r,s)} \in \{1,\dots,K\}$ denoted by

$$\begin{aligned} P(Y_{i(r,s)} \leq k) &= \frac{\exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})}{1 + \exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})} \\ &= \frac{\exp(\delta_r + \theta_k + \beta_{r0} - \beta_{s0} + z_{ir}^T\alpha_r - z_{is}^T\alpha_s)}{1 + \exp(\delta_r + \theta_k + \beta_{r0} - \beta_{s0} + z_{ir}^T\alpha_r - z_{is}^T\alpha_s)} \,. \end{aligned} \tag{4}$$

The model allows for the inclusion of so-called subject-object-specific covariates z_{ir}. It belongs to the general model family proposed by Schauberger and Tutz (2017a) for the inclusion of different types of covariates in paired comparison models. Within this framework, Tutz and Schauberger (2015) present a model including object-specific covariates z_r and Schauberger and Tutz (2017b) present a model including subject-specific covariates z_i. In Schauberger et al. (2017), the presented model is applied to data from the Bundesliga season 2015/16.

The response $Y_{i(r,s)}$ encodes an ordered response with K categories (including a category for draws) for a match between team a_r and team a_s on matchday i where a_r plays at its home ground. The linear predictor of the model contains the following terms:

δ_r team-specific home effects of team a_r

θ_k category-specific threshold parameters
β_{r0} team-specific intercepts
z_{ir} p-dimensional covariate vector that varies over teams and matches
α_r p-dimensional parameter vector that varies over teams.

Instead of fixed abilities γ_r, the teams have abilities $\gamma_{ir} = \beta_{r0} + z_{ir}^T \alpha_r$ which differ for each matchday depending on the covariates of team a_r on matchday i. In its general form, the model has a lot of parameters that need to be estimated. It could, for example, be simplified if both the home effect and the covariate effects were included with global instead of team-specific parameters. For this purpose, we use penalty terms to decide whether the home effect or single covariate effects should be considered with team-specific or global parameters. In particular, the absolute values of all pairwise differences between the team-specific home advantages are penalized using the L_1 penalty term

$$P(\delta_1, \ldots, \delta_m) = \sum_{r<s} |\delta_r - \delta_s| \,. \tag{5}$$

The penalty term enforces the clustering of teams with equal home effects as it is able to set differences between parameters to exactly zero. Therefore, the penalty could for example produce three clusters of teams where each of the clusters has a different home effect. As an extreme case, the penalty leads to one global home effect if all differences are set zero.

Also the team-specific covariate effects are penalized. The respective penalty term penalizes the absolute values of all pairwise differences of the covariate parameters and of the parameters themselves, i.e.

$$J(\alpha_1, \ldots, \alpha_m) = \sum_{j=1}^{p} \sum_{r<s} |\alpha_{rj} - \alpha_{sj}| + \sum_{j=1}^{p} \sum_{r=1}^{m} |\alpha_{rj}| \,. \tag{6}$$

The penalty enforces clustering of teams with respect to certain on-field variables, possibly leading to global effects instead of team-specific effects. Moreover, due to the penalization of the absolute values, covariates can be eliminated completely from the model. For comparability of the penalties and the resulting effects, all covariates have to be transformed to a joint scale.

Finally, both penalty terms are combined and the respective penalized likelihood

$$l_p(\cdot) = l(\cdot) - \lambda \left(P(\delta_1, \ldots, \delta_m) + J(\alpha_1, \ldots, \alpha_m) \right)$$

is maximized, $l(\cdot)$ denoting the (unpenalized) likelihood. The tuning parameter λ is chosen by 10-fold cross-validation with respect to the so-called ranked probability score (RPS) proposed by Gneiting and Raftery (2007). The RPS for ordinal response $y \in \{1, \ldots, K\}$ can be denoted by

$$RPS(y, \hat{\pi}(k)) = \sum_{k=1}^{K} (\hat{\pi}(k) - \mathbb{1}(y \leq k))^2,$$

where $\pi(k)$ represents the cumulative probability $\pi(k) = P(y \leq k)$. In contrast to other possible error measures (e.g. the deviance or the Brier score), it takes the ordinal structure of the response into account.

4 Application to Bundesliga Season 2016/17

We now apply the model to data from the Bundesliga season 2016/17. The data contain each of the 306 macthes of this season on the 34 matchdays. For easier interpretation of the intercepts, the covariates were centered (per team around the team-specific means). Centering of covariates only changes the paths (and interpretation) of the team-specific intercepts. Now, the intercepts represent the ability of a team when every covariate is set to the team-specific mean. Beside that, the paths and the interpretation of the covariate effects are not affected by the centering of the covariates. They represent the effect of a covariate on the ability of a team when the respective covariate deviates from the team-specific mean.

Figure 1 illustrates the parameters' paths for the proposed model, separately for each covariate along the tuning parameter λ. The dashed vertical line indicates the model that was selected by 10-fold cross-validation. In contrast to the home effects and all covariate effects, the team-specific intercepts are not penalized and, consequently, do not show any particular clusters of teams. Bayern München clearly dominated the league in this season which is also represented by a very high team-specific intercept.

The paths of the home and the covariate effects clearly illustrate the clustering effect of the penalty terms. It can be seen that the home effect seems to be equal for all teams. The home effect is positive and, therefore, represents an actual home advantage for all teams as it was expected. The greatest effect of all covariates can be seen for *Distance*. It has a strong positive effect for all teams. The teams gain better results in matches where they had a good running performance. Interestingly, the covariate *BallPossession* has negative effects for all teams. Here, only Darmstadt 98 is separate from the other teams with an even more negative effect while all other teams form a big cluster for this variable. None of the variables is eliminated completely from the model, each variable has effects for at least two of the teams. *TacklingRate* and *ShotsonGoal* have (small) positive effects for all teams. Figure 2 shows the RPS of the cross-validation along the tuning parameter λ.

5 Concluding remarks

This work deals with data from the German Bundesliga from the season 2016/17 and considers several on-field variables in a paired comparison model. We propose a model that is able to make use of the big amount of data that is collected in modern football and to simultaneously connect the corresponding variables to the outcome

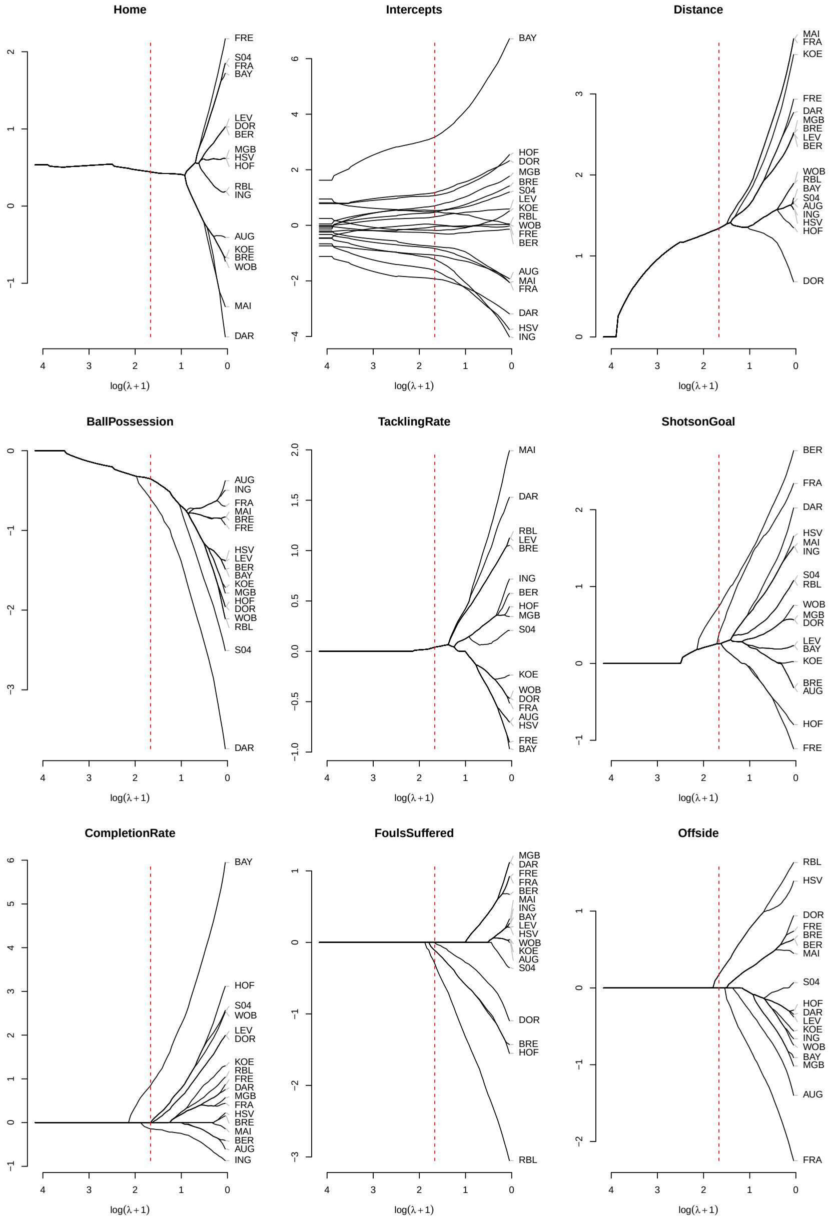


Fig. 1 Parameter paths, separately for home effect, intercepts and all (centered) covariates. Dashed vertical line represents the optimal model according to 10-fold cross-validation.

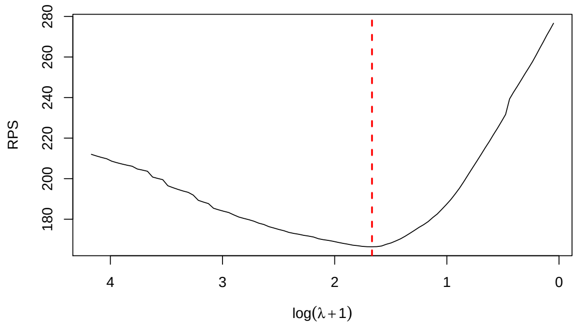


Fig. 2 Ranked probability score (RPS) for cross-validation along tuning parameter λ for model (4). Dashed vertical line represents optimal model according to 10-fold cross-validation.

of the matches. Complex modeling approaches are rather scarce in this area. The model incorporates football matches into the framework of paired comparisons and uses the general model proposed by Schauberger and Tutz (2017a) for the incorporation of different types of variables into paired comparison models.

In contrast to standard paired comparison models, the model offers a much more flexible and less restricted approach. Each team is assigned with individual strengths per matchday, depending on the on-field covariates of the team. This extension of the simple Bradley-Terry model allows for a much better discrimination between the different match outcomes and, therefore, for a better predictive performance.

References

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs, I: The method of pair comparisons. *Biometrika 39*, 324–345.

Gneiting, T. and A. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–376.

Schauberger, G., A. Groll, and G. Tutz (2017). Analysis of the importance of on-field covariates in the German Bundesliga. *Journal of Applied Statistics published online*, 1–18.

Schauberger, G. and G. Tutz (2017a). BTLLasso - A common framework and software package for the inclusion and selection of covariates in Bradley-Terry models. Technical Report 202, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.

Schauberger, G. and G. Tutz (2017b). Subject-specific modelling of paired comparison data - a lasso-type penalty approach. *Statistical Modelling 17*(3), 223–243.

Tutz, G. and G. Schauberger (2015). Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Advances in Statistical Analysis 99*(2), 209–227.

Are the shots predictive for the football results?

I tiri sono predittivi per modellare il numero di reti nel calcio?

Leonardo Egidi, Francesco Pauli, Nicola Torelli

Abstract In modelling football outcomes, scores' data are regularly used for the estimation of the attack and the defence strength of each team. However, these teams' abilities are quite complex and are correlated with many quantities inherent to the game. Additional available information, relevant for their estimation, are shots, both made and conceded. For such a reason, we propose a hierarchical model that incorporates this information in three stages for each game and each team: number of scores, number of shots on target and number of total shots. We fit the model on English Premier League data and obtained predictions for future matches.

Abstract *Nel modellare i risultati calcistici, i dati sui goal sono solitamente utilizzati per stimare le abilità di attacco e difesa di ogni squadra. Tuttavia, queste abilità hanno una natura complessa e sono correlate con molte quantità inerenti al gioco. Un'ulteriore informazione disponibile, rilevante per stimare questi parametri, è data dai tiri, sia quelli realizzati che quelli concessi. A tale scopo proponiamo un modello gerarchico che incorpora questa informazione in tre stadi per ogni partita e ogni squadra: numero di goal, numero di tiri nello specchio e numero di tiri totali. Abbiamo applicato il modello sui dati della Premier League inglese e ottenuto previsioni per partite future.*

Key words: modelling football outcomes, hierarchical model, shot, prediction

1 Introduction

Modelling the outcome of a football match is the subject of much debate, and various models based on different assumptions have been proposed. The basic assump-

Leonardo Egidi, Francesco Pauli, Nicola Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, 'Bruno de Finetti', Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: legidi@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it

tion is that the number of goals scored by the two teams follow two Poisson distributions (Maher, 1982; Baio and Blangiardo, 2010)—possibly with rates parameters accounting for different sources of information, such as bookmakers odds (Egidi et al., 2018)—but many researchers investigated the correlation between them by proposing a more complicated bivariate Poisson distribution (Karlis and Ntzoufras, 2003).

Another typical assumption is the inclusion in the models of some teams' effects to describe the attack and the defence strengths of the competing teams. For this aim, the advent of some dynamic structures (Owen, 2011; Koopman and Lit, 2015) allowed these parameters to vary over the time, in order to specify an intuitive temporal evolution of these teams' skills along the match days and the seasons. The historical match results, possibly along with a set of further covariates, are usually the only data used for the estimation of these abilities. However, the scoring and the defence abilities are strongly correlated with the shots and the shots conceded respectively. For such a reason, including this information into a model designed for predicting the scores could provide relevant benefits both in terms of the realistic description of the game and the prediction of future matches outcomes. As an example of the relevance of the shots on target and the total shots on the statistical prediction of match results for the Italian football league Serie A, see Carpita et al. (2015).

In this paper, we propose a Bayesian hierarchical model consisting of a data-hierarchy in three stages for each game and each team, where the nested quantities are: number of scores, number of shots on target and number of total shots. The number of scores and the number of shots on target follow two binomial distributions respectively, with probability and population treated as further parameters. Intuitively, the total shots also consist of all those attempts—e.g., long distance shots, last minute shots—which may represent a noisy proxy for the attack skills, and for such a reason they represent the last level of the assumed hierarchy.

As far as we know from reviewing the current literature, this proposal represents a novelty also in terms of parameters' interpretation: binomial probabilities associated to the first two levels reflect the conversion rate of the shots on target in goals and the precision rate of all the shot attempts, respectively. In Sect. 2 we introduce the entire model, and we focus on the Gaussian process for the attack and the defence abilities. Moreover, we assess the binomial assumption for the scores through a Pearson's chi-squared test. We present the application on the English Premier League in Sect. 3, along with parameters' estimates and predictions for the test set season. Sect. 4 concludes.

2 A joint model for the shots and the scores

Here, $\boldsymbol{y}_m = (y_{m1}, y_{m2})$ denotes the vector of observed scores, where y_{m1} and y_{m2} are the number of goals scored by the home team and by the away team in the m-th match of the dataset, respectively. Let $\boldsymbol{s}_m = (s_{m1}, s_{m2})$ denote the shots on the target and $\boldsymbol{w}_m = (w_{m1}, w_{m2})$ the total number of shots, respectively. For each m, the data information is represented by the joint vector $(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{w})$. The total number of teams considered across the seasons is $T = 34$. In what follows, the nested indexes $h(m), a(m) = 1, \dots, T$ and $\tau(m)$ identify the home team, the away team and the season τ associated with the m-th game, respectively. The three-stages hierarchical model for the scores and the shots is then specified as follows:

$$\begin{aligned} y_{m1} &\sim \mathsf{Binomial}(s_{m1}, p_{h(m),\tau(m)}) \\ y_{m2} &\sim \mathsf{Binomial}(s_{m2}, q_{a(m),\tau(m)}) \\ s_{m1} &\sim \mathsf{Binomial}(w_{m1}, u_{h(m)}) \\ s_{m2} &\sim \mathsf{Binomial}(w_{m2}, v_{a(m)}) \\ w_{mj}|\theta_{mj} &\sim \mathsf{NegBinomial}(\theta_{mj}, \phi), \;\; j = 1,2. \end{aligned} \tag{1}$$

The *conversion probabilities* p and q are modelled with two inverse logit, depending on the attack and the defence strengths of the competing teams:

$$\begin{aligned} p_{h(m),\tau(m)} &= \text{logit}^{-1}(\mu + att_{h(m),\tau(m)} + def_{a(m),\tau(m)}) \\ p_{a(m),\tau(m)} &= \text{logit}^{-1}(att_{a(m),\tau(m)} + def_{h(m),\tau(m)}). \end{aligned} \tag{2}$$

The attack and defence parameters are assumed to follow two Gaussian processes:

$$\begin{aligned} att_{,\tau} &\sim \mathsf{GP}(\mu_{att}(\tau), k(\tau)), \\ def_{,\tau} &\sim \mathsf{GP}(\mu_{def}(\tau), k(\tau)), \end{aligned} \tag{3}$$

with mean functions $\mu_{att}(\tau) = att_{\cdot,\tau-1}$, $\mu_{def}(\tau) = def_{\cdot,\tau-1}$, and covariance function with generic element $k(\tau)_{i,j} = \exp\{-(\tau_i - \tau_j)^2\} + 0.1$. As outlined in the literature, a 'zero-sum' identifiability constraint within each season is required: for T teams we assume: $\sum_{t=1}^{T} att_{t,\tau} = 0$, $\sum_{t=1}^{T} def_{t,\tau} = 0, \tau = 1, \dots \mathscr{T}$.

The *shots' precision probabilities* u and v and the *shooting rates* θ_{m1}, θ_{m2} are given a Beta distribution with hyperparameters δ, ε and a Gamma distribution with hyperparameters α, β, respectively:

$$u_{h(m)} \sim \mathsf{Beta}(\delta_{h(m)}, \varepsilon_{h(m)}), \; v_{a(m)} \sim \mathsf{Beta}(\delta_{a(m)}, \varepsilon_{a(m)}) \tag{4}$$

$$\theta_{m1} \sim \mathsf{Gamma}(\alpha_{h(m)}, \beta_{h(m)}), \; \theta_{m2} \sim \mathsf{Gamma}(\alpha_{a(m)}, \beta_{a(m)}). \tag{5}$$

The model is completed by the specification of weakly informative priors for the home effect parameter μ, the overdispersion parameter ϕ, and the hyperparameters $\alpha, \beta, \delta, \varepsilon$.

It is of interest to assess the legitimacy of the binomial distribution for the model above. For this purpose, we consider the empirical distribution of the scores conditioned on a given number of shots on target, $y_{\cdot j}|s_{\cdot j} = z$, $z \in \mathbb{N}$, and we check whether this sample may be thought as drawn from a Binomial(n, p), with n and p fixed. For each z, we performed some Pearson χ^2-tests comparing the empirical distribution of the scores conditioned on the z-th shot on target and the hypothesized binomial distribution, with n and p estimated from the data. For both the home and the away scores, for each z the Pearson χ^2 test suggests to not reject the null hypothesis of binomial distribution (all the p-values are always greater than the threshold $\alpha = 0.05$).

3 Application: Premier League from 2007/2008 to 2016/2017

We collected the historical data arising from 10 seasons of the English Premier League (EPL), from 2007/2008 to 2016/2017. The data structure of the model is presented in Table 1 with respect to the first match day in EPL 2007/2008. The goal is fitting the model and deriving the parameters' estimates. Secondly, we make predictions for a set of future matches. Model coding has been written using Stan (Carpenter et al., 2017), precisely the Rstan interface. We strictly followed the software guidelines for monitoring the chains' convergence and speeding up the computational times. The chosen number of Hamiltonian Markov Chain iterations is 2000, with a burnin period of 500.

Table 1 Data structure for the first match day, EPL, 2007/2008 season. Each column reports: match, season, home team, away team, home goals, away goals, home shots, away shots , home shots on target, away shots on target.

Match	Season	$h[m]$	$a[m]$	y_{m1}	y_{m2}	w_{m1}	w_{m2}	s_{m1}	s_{m2}
1	07/08	Aston Villa	Liverpool	1	2	10	17	6	7
2	07/08	Bolton	Newcastle	1	3	13	7	9	5
3	07/08	Derby	Portsmouth	2	2	12	12	5	6
4	07/08	Everton	Wigan	2	1	12	14	8	4
5	07/08	Middlesbrough	Blackburn	1	2	10	4	6	4
6	07/08	Sunderland	Tottenham	1	0	9	6	4	3
7	07/08	West Ham	Man City	0	2	9	14	2	5
8	07/08	Arsenal	Fulham	2	1	19	12	13	9
9	07/08	Chelsea	Birmingham	3	2	19	6	11	4
10	07/08	Man United	Reading	0	0	22	3	9	2

3.1 Parameters' estimates

As usual in Bayesian inference, posterior means ± posterior standard deviations or posterior quantiles are the summaries for describing and visualizing the posterior distribution of the parameters.

The attack and defence abilities are directly connected with the scoring probabilities in (2), and modelled as Gaussian processes in (3) in terms of seasonal evolution. Fig. 1 displays the 50% posterior intervals for the attack (solid red line) and the defence (solid black line) in formula (3) across the seasons. Higher values for the attack are associated with a greater propensity to convert the shots on target in goals; conversely, lower values for the defence correspond to a better ability to not concede goals. These plots may explain something unexpected even for football experts and help in revealing new instances behind events thought as completely unpredictable. For instance, Leicester won the Premier League 2015/2016 with associated initial odds such as 1:5000, and no one, at the beginning of the season, could have predicted that performance. However, there is a surprising trend that emerges clearly: starting from 2007/2008 season, Leicester dramatically improved the propensity to convert the shots in goal and, at the same time, reinforced its defence. The values registered for the attack and the defence are among the highest and the lowest in the EPL respectively. Maybe, the victory of Leicester, despite highly surprising, was less unpredictable than what the experts had thought.

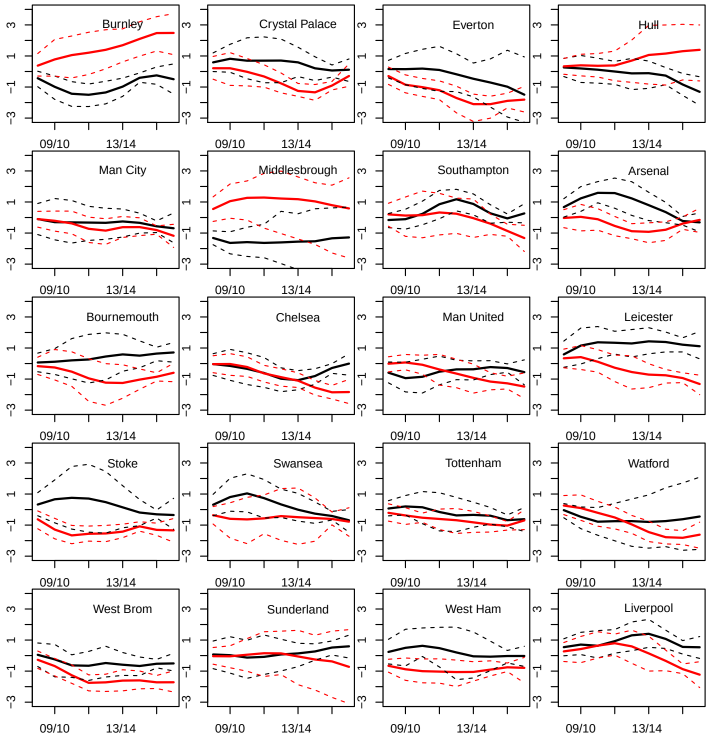


Fig. 1 Posterior estimates for the attack (solid black lines) and defence (solid red lines) across the nine seasons considered, from 2008/2009 to 2016/2017, for the twenty teams belonging to the EPL in the 2016/2017 season.

The power of model (1) is to represent a sort of scores' genesis, able to approximately reproduce the features of the real game. The scores represent the final level, depending on the population of the shots on target, which in turn depends on the population of the total shots. The posterior means ± standard error for the average conversion probabilities p, q and the home precision probabilities u are displayed in Fig. 2 (left panel). For the majority of the teams, the precision probabilities (black bars) are higher than the conversion probabilities (red and blue bars), and this is intuitive in terms of football features: usually, the ratio between the shots on target and the total shots tends to be higher than the ratio between the goals and the shots on target. However, Middlesbrough, one of the three relegated teams at the end of the 2016/2017 season, is associated with the highest precision probability—half of its shots are on the target—but with the lowest conversion—only about one attempt over ten in the targets corresponds to a score. Conversely, the precision probabilities for Leicester almost overlap the conversion probabilities. For what concerns the total shots, Fig. 2 (right panel) displays the average shots rates, where Chelsea, Manchester City, Tottenham and Liverpool register the highest values. For each team, the trend is to kick more when playing at home.

Although a broad analysis of these statistics should benefit from other comparisons and covariates, we imagine these kinds of plots and summaries could be beneficial for football managers or tactic experts, at least in a naive perspective summarized by the quote 'kick less, kick better'.

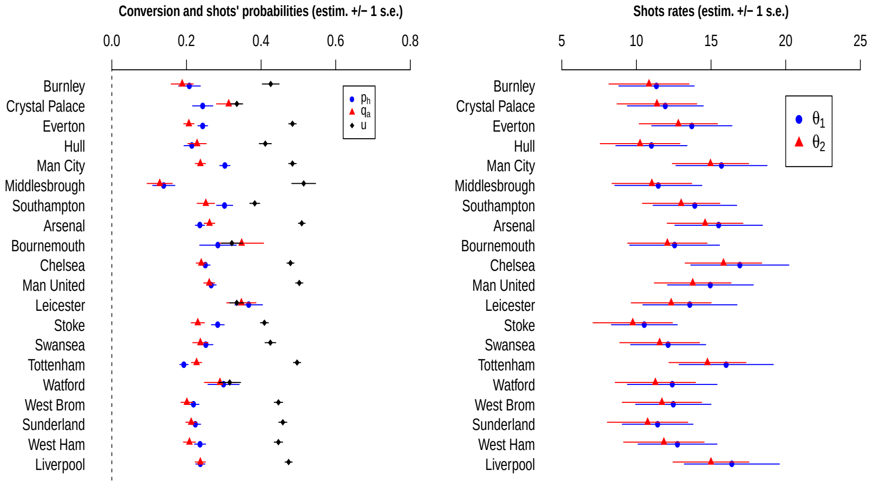


Fig. 2 Average of the posterior estimates for the conversion probabilities p_h and q_a and for the home precision probabilities u (left panel); average of the posterior estimates for the shots rates $\theta_{\cdot 1}, \theta_{\cdot 2}$ (right panel) for the twenty teams belonging to the EPL 2016/2017.

3.2 Prediction and posterior probabilities

Making predictions for future games and seasons is of great appeal for sport statisticians. We used historical data arising from the past seasons for making predictions about the tenth season, the EPL 2016/2017. As usual in a Bayesian framework, the prediction for a new dataset may be performed directly via the posterior predictive distribution for our unknown set of observable values. Fig. 3 displays the posterior 50% credible bars (grey ribbons) for the predicted achieved points for each team for the season 2016/2017, together with the observed final ranks. At a first glance, the model correctly detects Chelsea as EPL champion at the end of the 2016/2017 season, and Middlesbrough and Hull City relegated in Championship. Manchester City, Tottehnham and Arsenal appear to be definitely underestimated, whereas Manchester United and Leicester are quite overestimated. Globally, the predictions reflect the observed pattern.

Table 2 reports the model posterior probabilities being the first, the second and the third relegated team; as may be noticed, Sunderland was pretty unlikely to be relegated in Championship. Conversely, Burnley had an high probability to be relegated, but it performed better than the predictions.

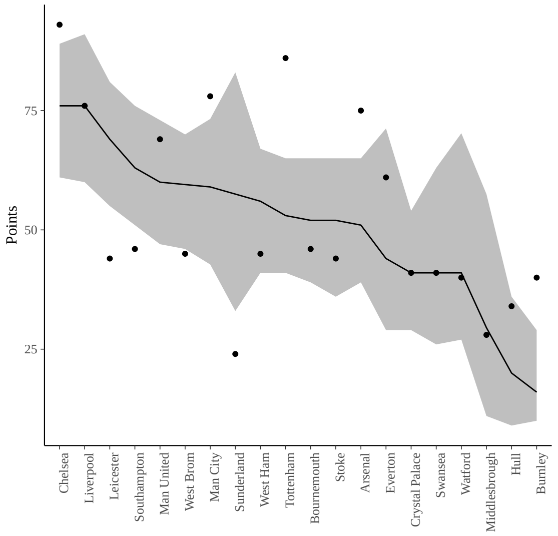


Fig. 3 Posterior 50% credible bars (grey ribbons) for the achieved final points of English Premier League 2016/2017. Black dots are the observed points. Black lines are the posterior medians.

Table 2 Estimated posterior probabilities for each team being the first, the second, and the third relegated team in the Premier League, 2016/2017, together with the observed rank and the number of points achieved (relegated predicted teams by the model are emphasized).

Team	P(1st rel)	P(2nd rel)	P(3d rel)	Actual rank	Points
Burnley	0.096	0.161	0.245	16	40
Hull	0.103	0.198	0.254	18	34
Middlesbrough	0.063	0.117	0.226	19	28
Sunderland	0.055	0.045	0.027	20	24

4 Discussion

We have proposed a Bayesian hierarchical model consisting of a three-stage hierarchy for the scores, the shots on target and the number of total shots. The main novelty is the inclusion of an important latent football feature represented by the kicking ability of each team, modelled both in terms of intensity and precision. Preliminary results on future matches seem to be promising in terms of predictive accuracy. Model comparisons and goodness of fit tools are issues of future interest.

References

Baio, G. and M. Blangiardo: Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics 37*(2), 253–264 (2010)

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell: Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1) (2017)

Carpita, M., M. Sandri, A. Simonetto, and P. Zuccolotto: Discovering the drivers of football match outcomes with data mining. *Quality Technology & Quantitative Management 12*(4), 561–577 (2015)

Egidi, L., F. Pauli, and N. Torelli: Combining historical data and bookmakers' odds in modelling football scores. *arXiv preprint arXiv:1802.08848* (2018)

Karlis, D. and I. Ntzoufras: Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(3), 381–393 (2003)

Koopman, S. J. and R. Lit: A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*(1), 167–186 (2015)

Maher, M. J.: Modelling association football scores. *Statistica Neerlandica 36*(3), 109–118 (1982)

Owen, A.: Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics 22*(2), 99–113 (2011)

Zero-inflated ordinal data models with application to sport (*in*)activity

Modelli per dati ordinali zero-inflazionati con applicazione all'(in)attività sportiva

Maria Iannario and Rosaria Simone

Abstract Traditional models for ordinal data (as CUB models or cumulative models with logit/probit link, among others) present limits in explaining the surplus of zero observations, especially when the zeros may relate to two distinct situations of non-participation/inactivity and infrequent participation, for instance. We consider an extension of standard models: zero-inflated CUB models and zero inflated ordered cumulative (ZIOC) probit/logit models handling the GECUB models and using a double-hurdle combination of a split (logit/probit) model and an ordered probit/logit model, respectively. Both extensions, potentially, relate to different sets of covariates. Finally, models are applied to Sport surveys. Specifically the paper investigates the determinants of sport (*in*)activity: the frequency and the probability of sports participation. It distinguishes between genuine "non-participants" and the ones who do not participate at a time but might do under different circumstances.

Abstract *I modelli tradizionali per i dati ordinali (come modelli* CUB *o cumulativi con link logit/probit, tra gli altri) presentano limiti nello spiegare il surplus di osservazioni nella categoria zero, specialmente quando gli zeri possono riguardare due distinte situazioni di non partecipazione/non attività e/o partecipazione non frequente. Il lavoro propone un'estensione di modelli standard: i modelli* CUB zero inflated *e i modelli ordinal probit/logit con inflazione di zeri (*ZIOC*). I primi costituiscono una revisione dei modelli* GECUB *(modelli* CUB *con effetto shelter), i secondi costituiscono una mistura di modelli (probit/logit) dicotomici e modelli probit/logit ordinali. Entrambe le estensioni possono riferirsi a diversi gruppi di covariate. Infine, i modelli sono applicati a dati rilevati da indagini sullo sport. In particolare, lo studio esplora le determinanti dell'(in)attività sportiva: la frequenza e la probabilità di partecipazione ad attività sportive. Distingue tra veri "non partecipanti"*

Maria Iannario
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22 - Napoli
e-mail: maria.iannaro@unina.it

Rosaria Simone
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22 - Napoli
e-mail: rosaria.simone@unina.it

e coloro che non partecipano al momento dell'indagine, ma potrebbero se in circostanze diverse.

Key words: CUB models, Ordinal logit/probit models, Ordered outcomes, discrete data, sport inactivity, zero-inflated responses

1 Introduction

Excess of zeros is a commonly encountered phenomenon that limits the use of traditional regression models for analysing ordinal data in contexts where respondents express a graduated perception on a specific item or experiments identify levels of increasing assessments.

The situation occurs with ordinal scales in which there is an anchor that represents the absence of the symptom or activity, such as none, never or normal. This level usually tagged *zero* may be scored by respondents certainly not at risk (without symptom or who do not practice any activity/exercise) and respondents with a non-zero probability of risk.

Survey data concerning epidemiological studies or choices, particularly those that refer to an explicit time dimension, may include genuine non-participants whatever the circumstances are, as well as individuals who would decide to participate if the circumstances were different. It is, therefore, likely that these two types of zeros are driven by different behaviour. One example is a study by Harris and Zhao (2007) on the consumer choice problem of tobacco consumption or the analysis of Downward et al. (2011) on sports participation, among others.

Aim of the paper is introducing methodologies that allow users of ordinal scale data to more accurately model the distribution of ordinal outcomes in which some subjects are susceptible to exhibit the response and some are not (i.e. the dependent variable exhibits zero inflation). The study explores the determinants of sport (*in*)activity: the frequency and the probability of sports participation. It distinguishes between genuine "non-participants" and the ones who do not participate at a time but might do under different circumstances. Thus, it includes whether or not to participate in sport and, subsequently, what intensity of participation is undertaken. It is able to distinguish between structured and sampling zeros implementing some results obtained for count data in the ordinal data context.

With respect to the standard models for ordinal data the new methodologies exceed some gaps related to the model of zeros by taking into account the potentially two-fold decision made with respect to participation. Here we propose extensions of standard models: zero-inflated CUB (ZICUB) models and zero inflated ordered cumulative (ZIOC) probit/logit models handling the GECUB models and using a double-hurdle combination of a split (logit/probit) model and an ordered probit/logit models, respectively. Both extensions, potentially, relate to different sets of covariates. The modelling assumption is that different decisions govern the choice to participate and the frequency of participation in sport. The remainder of the paper is as follows.

Section 2 reviews the methods used for the analysis. Section 3 describes the data set and main estimation results with a summary of the main findings and opportunities for further research.

2 Methods

Let Y be a discrete random variable that assumes the ordered values of $0,1,...,J$. Standard ordinal cumulative (Agresti, 2010) or CUB models (Piccolo, 2003) map a single latent variable Y^* to the observable Y, with Y^* related to a set of covariates or consider the response as a weighted mixture of respondents' propensity to adhere to a meditated choice (formally described by a shifted Binomial random variable) and a totally uninformative choice (described by a discrete Uniform distribution) with a possible *shelter* effect (Iannario, 2012; Iannario and Piccolo, 2016), respectively. Here we propose a zero inflated cumulative (ZIOC) model that involves two latent equations with uncorrelated error terms: a logit/probit equation and an ordered logit/probit equation by introducing ZIOL/ZIOP models (subsection 2.1). Or in order to further disentangle the *inflated effect* concentrated at category *zero* we may introduce a variant of GECUB models (subsection 2.2).

2.1 Zero Inflated Cumulative Models

Let r denote a binary variable indicating the split between Regime 0 ($r=0$, for "non participants") and Regime 1 ($r=1$ for "participants"), which is related to the latent variable $r^*=\boldsymbol{x}'\boldsymbol{\beta}+\varepsilon$ where $\boldsymbol{x}$ is a vector of p individual characteristics (covariates) that determine the choice of regimes, $\boldsymbol{\beta}$ is a p-vector of unknown regression parameters, and ε is a random variable with cumulative distribution function $G_\varepsilon(.)$. Accordingly, the probability of an individual being in Regime 1 is given by

$$Pr(r=1|\boldsymbol{x})=Pr(r^*>0|\boldsymbol{x})=G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta}),$$

where we assume $G_\varepsilon(.)$ strictly increasing and symmetric around zero. Standard choices for the distribution function are the logit link function, $G(t)=1/(1+e^{-t})$, corresponding to the logistic distribution, or the probit link function, $G(t)=\Phi(t)$, with Φ the cdf of the standard normal distribution.

Conditional on $r=1$, respondents levels under Regime 1 are represented by $\tilde{Y}(\tilde{Y}=0,1,\dots,J)$, which is generated by a cumulative link model based upon a second underlying latent variable $\tilde{Y}^*$, where

$$\tilde{Y}^*=\boldsymbol{z}'\boldsymbol{\gamma}+u,$$

with $\boldsymbol{z}$ being a vector of covariates with unknown parameters $\boldsymbol{\gamma}$ and $u\sim G_\varepsilon(.)$. The observed ordinal variable $\tilde{Y}$ takes as values the labels 0 if $\tilde{Y}^*\leq 0$, J if $\tilde{Y}^*\geq\alpha_{J-1}$

otherwise,

$$\tilde{Y} = j \qquad \iff \qquad \alpha_{j-1} < \tilde{Y}^* \le \alpha_j \qquad j = 1,2,\dots,J-1\,, \qquad j > 2\,,$$

where $\alpha_j\,(j = 1,\dots,J-1)$ are the intercept values to be estimated in addition to the covariate coefficients $\boldsymbol{\gamma}$. Notice that Regime 1 also allows for zero scores. That is, to observe $Y = 0$ we require either that $r = 0$ (the individual is a non participant) or jointly that $r = 1$ and $\tilde{Y} = 0$ (the individual is a zero consumption participant). To observe a positive score, instead, we require jointly that the individual is a participant $(r = 1)$ and $\tilde{Y}^* > 0$. If we assume that the error terms from the first stage equation and the second stage cumulative outcome equation, that is e and u, are not correlated the probability mass function of the ZIOC model is

$$Pr\,(Y) = \begin{cases} Pr\,(Y = 0 \mid z,\boldsymbol{x}) = Pr\,(r = 0 \mid \boldsymbol{x}) + Pr\,(r = 1 \mid \boldsymbol{x})\,Pr\,\big(\tilde{Y} = 0 \mid \boldsymbol{z}, r = 1\big) \\ Pr\,(Y = j \mid z,\boldsymbol{x}) = Pr\,(r = 1 \mid \boldsymbol{x})\;Pr\,\big(\tilde{Y} = j \mid z, r = 1\big)\;\;(j = 1,\dots,J) \end{cases}$$

$$= \begin{cases} Pr\,(Y = 0 \mid z,\boldsymbol{x}) = [1 - G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta})] + G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta})G_\varepsilon(-\boldsymbol{z}'\boldsymbol{\gamma}) \\ Pr\,(Y = j \mid z,\boldsymbol{x}) = G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta})[G_\varepsilon(\alpha_j - \boldsymbol{z}'\boldsymbol{\gamma}) - G_\varepsilon(\alpha_{j-1} - \boldsymbol{z}'\boldsymbol{\gamma})]\,(j = 1,\dots,J-1) \\ Pr\,(Y = J \mid z,\boldsymbol{x}) = G_\varepsilon(\boldsymbol{x}'\boldsymbol{\beta})[1 - G_\varepsilon(\alpha_{J-1} - \boldsymbol{z}'\boldsymbol{\gamma}). \end{cases}$$

In this way, the probability of a zero score has been inflated as it is a combination of the probability of zero consumption from the cumulative model framework and the probability of non-participation from the split logit/probit model. Notice that the choice of distribution function G_ε allows to consider the Zero Inflated Ordinal Probit (ZIOP) as in Harris and Zao (2007) or Zero Inflated Ordinal Logit (ZIOL) models. Once the full set of probabilities has been specified, and given an *iid* sample $(i = 1,\dots,n)$ from the population on $(Y_i, \boldsymbol{x}_i, z_i)$, the parameters of the full model $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\alpha}')'$ may be estimated using the maximum likelihood (ML) methods. The log-likelihood function is $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=0}^{J} I[Y_i = j] \log Pr\,(Y_i = j|\boldsymbol{x}_i, z_i, \boldsymbol{\theta})$, where $I[Y_i = j]$ is the indicator function of $(Y_i = j)$.

2.2 Zero Inflated CUB Models

Let $\check{Y} \sim CUB_{she=1}(\check{\pi}, \xi, \delta; J+1)$ be a CUB distributed random variable over $J+1$ categories and *shelter* at $c = 1$:

$$Pr\,\big(\check{Y} = j \mid \check{\pi}, \xi, \delta\big) = \delta\,D_j^{(c)} + (1-\delta)\,\big[\check{\pi}\,b_j(\xi) + (1-\check{\pi})\,h_j\big],\;\; j = 1,2,\dots,J+1,$$

where $h_j = \dfrac{1}{J+1}$ is the discrete Uniform distribution over the given support and $b_j(\xi)$ denotes the shifted Binomial distribution with parameter $1-\xi$.

Then, a ZICUB model for the response variable $Y \in \{0,\dots,J\}$ is specified by setting $Y = \check{Y} - 1$. In this way

$$P_r(Y) = \begin{cases} Pr(Y=0|\boldsymbol{\theta}) = \delta + (1-\delta)\left[\pi b_1(\xi_i) + (1-\pi)\dfrac{1}{J+1}\right] \\ Pr(Y=j|\boldsymbol{\theta}) = (1-\delta)\left[\pi b_{j+1}(\xi) + (1-\pi)\dfrac{1}{J+1}\right], \quad j=1,\dots,J. \end{cases}$$

In addition to examine the effects of risk factors on the response variable it may be proposed the inclusion of covariates on the parameters through canonical logit link:

$$logit(\delta_i) = \boldsymbol{\omega}'\boldsymbol{x}_i;\;\; logit(\pi_i) = \boldsymbol{\eta}'\boldsymbol{z}_i;\;\; logit(\xi_i) = \boldsymbol{\zeta}'\boldsymbol{w}_i\,.$$

Here, $\boldsymbol{\theta} = (\boldsymbol{\omega}', \boldsymbol{\eta}', \boldsymbol{\zeta}')'$ is the parameter vector characterizing the distribution of $(Y_1, Y_2, \dots, Y_n)$ with $\boldsymbol{\omega}', \boldsymbol{\eta}', \boldsymbol{\zeta}'$ denoting the parameter vector for the *shelter*, uncertainty and feeling components, respectively, and $\boldsymbol{x}_i \in \boldsymbol{X}, \boldsymbol{z}_i \in \boldsymbol{Z}$ and $\boldsymbol{w}_i \in \boldsymbol{W}$ being the selected covariates for the i-th subject of the three components. The zero-inflated variant of GECUB models also assumes that some zeros are observed due to a specific structure in the data.

Here, given an observed random sample $(Y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{w}_i)$, for $i = 1, 2, \dots, n$, the log-likelihood function is $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=0}^{J} I[y_i = j] \log P_r(Y_i = j|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{w}_i, \boldsymbol{\theta})$, where $I[Y_i = j]$ is the indicator function of $(Y_i = j)$.

3 Data and application

The determinant of sport (*in*)activity will be discussed on the basis of two case studies involving data collected in 2016 and 2017, respectively, through a web link related to the BDsports project (http://bodai.unibs.it/bdsports/). The case studies have been selected to highlight pitfalls and advantages of the two main proposals and to allow the distinction between genuine inactive respondents and the ones who do not play sport at a time of surveys.

In the last decade the modelling of sports participation decision has increased in complexity. The sports participation variable is measured in different ways; relatively few studies consider the time spent on sports participation or the frequency of such participation as we done in this paper. The dependent variable under investigation for 2016 is a rating on a 7 point scale whereas for 2017 is a rating on 11 categories (see Figure 1), asking each respondent the time dedicated to sport practice on weekly basis: from ‘0 = Rarely practiced any sport/not practiced any sport at all’, ‘1 = Less than one hour’ up to ‘6 = More than 7 hours’ (up to ‘10’ for 2017). Notice that the two surveys are about two different main topics (sport preferences and habits for the first survey, on the exercise addiction for the second ones); however both of them present a rating question on sport activity. Because the dependent variable is ordered rather than continuous and because, as noted in the Introduction, ‘zero’ participation could measure never participated (genuine inactive) or not recently/rarely participated a Zero-inflated ordered (ZIOP) estimator and ZICUB models are employed for 2016 whereas a ZIOL and ZICUB models for 2017. The modelling assumption is that different decisions govern the choice to participate and the frequency of participation in sport. Hurdle models are not considered by following what it is in Downward et al. (2011).

Evidence in the literature reveals that the probability of sports activity decreases with *age* (Barber and Havitz, 2001; Downward and Rasciute, 2010; among others) with less difference in gender among the older adults (Bauman et al. 2009). *Gender*, in fact, is the other covariate that has a highly important influence on sports activity. There is evidence about the fact that men, in general, not only participate in sport more than women (Downward and Rasciute, 2010; Eberth and Smith, 2010; Hovemann and Wicker, 2009; Lera-López and Rapún-Gárate, 2007) but they also show a higher frequency of participation (Barber and Havitz, 2001; Eberth and Smith, 2010). These differences may be attributed to biological factors, and cultural and social influences (Humphreys and Ruseski, 2010). Another determinant of sport (*in*)activity is the smoking habit; it has (with alcohol consumption) negative effects on sport practice especially in relation to age (Perretti et al. 2002).

Thus, in our analysis the selected covariates for 2016 are *gender, age* and the *smoking* habit. Results concerning a sample of $n = 647$ respondents are in Table 1 with the thresholds (cutpoints) on the underlying scale for ZIOP model $\hat{\alpha}_1 = -4.683$, $\hat{\alpha}_2 = -0.918$, $\hat{\alpha}_3 = -0.751$, $\hat{\alpha}_4 = -0.188$, $\hat{\alpha}_5 = -0.026$, $\hat{\alpha}_6 = 1.382$.

As revealed by estimation results of ZIOP model (Table 1) and mentioned in the literature sport activity reduces with older age; furthermore, smokers are generally inactive as well as women. Both estimated models confirm the effect of *age* for the inflation in the zero category that in ZICUB models is explicitly due to *smoking* habits. The best performance in terms of BIC index is for ZICUB model (bold in Table).

Similar results have been obtained for the analysis of the second survey (2017) where *gender, age* and the dichotomous response to the question “do you practice any sport or physical activity?” are considered ($n = 554$). Results are in Table 2 with the thresholds (cutpoints) on the underlying scale for ZIOL model $\hat{\alpha}_1 = -6.975$,

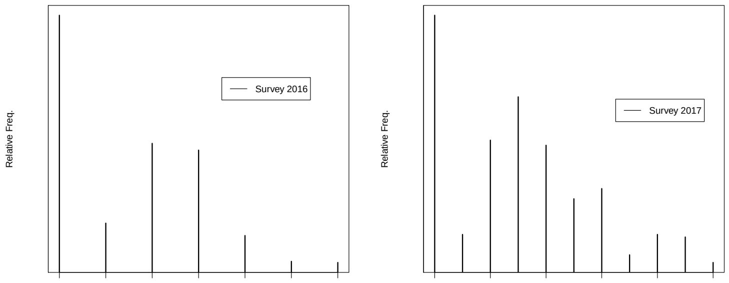


Fig. 1 Frequency distribution of the time spent on sports participation ($J = 0, 1, \dots, 6$; Survey 2016 left side) ($j = 0, 1, \dots, 10$; Survey 2017 right side)

Table 1 Regression results for ZICUB and ZIOP models

Models	Covariates	Parameters	Estimates	StdErr
ZICUB	*Constant*	$\hat{\omega}_0$	−0.871	0.115
	Smoke	$\hat{\omega}_1$	0.556	0.201
	Constant	$\hat{\eta}_0$	2.773	0.715
	Age	$\hat{\eta}_1$	−0.057	0.022
	Constant	$\hat{\zeta}_0$	0.283	0.063
	Woman	$\hat{\zeta}_1$	0.163	0.085
	$\ell(\boldsymbol{\theta})$	−1056.233	*BIC*	**2172.876**
ZIOP	Age	$\hat{\beta}_1$	0.611	0.087
	Constant	$\hat{\gamma}_0$	−0.309	0.112
	Woman	$\hat{\gamma}_1$	−0.345	0.104
	Smoke	$\hat{\gamma}_2$	0.014	0.005
	$\ell(\boldsymbol{\theta})$	−1067.021	*BIC*	2177.189

$\hat{\alpha}_2 = -1.128$, $\hat{\alpha}_3 = -0.825$, $\hat{\alpha}_4 = -0.571$, $\hat{\alpha}_5 = -0.543$, $\hat{\alpha}_6 = -0.074$, $\hat{\alpha}_7 = 0.011$, $\hat{\alpha}_8 = 0.300$, $\hat{\alpha}_9 = 1.034$.

Here it possible to notice the different impact of *age* on the uncertainty component for ZICUB model; for ZIOL model, instead, it has been confirmed the increasing inactivity for older respondents. Furthermore, to be woman and the answer to no sport/physical activity practiced represent requisites which express sport inactivity. In ZICUB model the effect of *gender* influences the feeling component by explaining woman inactivity, especially for "no practice at all" respondents. Generally this last model presents a better performance (BIC index in bold); here the inflation in zero consistently increases with the response "no practice". Data and the *R* code for the implementation of the methods are available upon request from Authors.

Finally, by testing different covariates and models to explain sport (*in*)activity it turns out that *age, gender, smoking* habit and *no sport/physical activity* exercised affect the occurrence of sedentary behaviour: given all these drivers it is possible to analyse the effect of age for zero inflation in ZIOC models, and smoking habits and no sport/physical activity practised for ZICUB models. Both implemented methods confirm the main results of the literature. Although the choice between the two zero-inflated approaches is generally based on the aim of the study, the evaluation in terms of fitting results and the interpretation of covariates may address the selection. Moreover, it is important to highlight some computational drawbacks related to the performance of ZIOC models.

Generally assessing the nature of the zero scores is becoming a more and more relevant issue demanding for the use of both the proposals. They can be used to estimate the proportion of zeros coming from each regime, and to evaluate how the split changes with observed characteristics.

Simulation studies will be planned to further validate and compare the efficacy of the proposals.

Table 2 Regression results for ZICUB and ZIOL models

Models	Covariates	Parameters	Estimates	Std Err
ZICUB	*Constant*	$\hat{\omega}_0$	−5.281	2.357
	No practice at all	$\hat{\omega}_1$	7.202	2.382
	Constant	$\hat{\eta}_0$	−2.895	0.860
	Age	$\hat{\eta}_1$	0.112	0.035
	Constant	$\hat{\zeta}_0$	0.347	0.130
	Woman	$\hat{\zeta}_1$	1.175	0.282
	No practice at all	$\hat{\zeta}_2$	0.212	0.084
	$\ell(\boldsymbol{\theta})$	−1085.695	*BIC*	**2215.61**
ZIOL	Age	$\hat{\beta}_1$	3.595	0.764
	Constant	$\hat{\gamma}_0$	−0.429	0.398
	Woman	$\hat{\gamma}_1$	−5.177	0.529
	No practice at all	$\hat{\gamma}_2$	−0.060	0.012
	$\ell(\boldsymbol{\theta})$	−1078.085	*BIC*	2244.611

References

1. Agresti A.: *Analysis of Ordinal Categorical Data*, 2^{nd} Ed., J.Wiley & Sons, Hoboken (2010).
2. Barber, N., Havitz, M.E.: Canadian participation rates in ten sport and fitness activities. *Journal of Sport Management*, **15**, 51–76 (2001).
3. Bauman, A., Sallis, J., Dzewaltowski, D., Owen, N.: Toward a better understanding of the influences on physical activity. *American Journal of Preventive Medicine*, **23** (2S), 5–14 (2002).
4. Downward, P., Lera-López, F., Rasciute, S.: The Zero-Inflated ordered probit approach to modelling sports participation, *Economic Modelling*, **28**, 2469-2477 (2011).
5. Downward, P., Rasciute, S.: The relative demands for sports and leisure in England. *European Sport Management Quarterly*, **10** (2), 189–214 (2010).
6. Eberth, B., Smith, M.: Modelling the participation decision and duration of sporting activity in Scotland. *Economic Modelling*, **27** (4), 822–834 (2010).
7. Harris, N.M., Zhao, X.: A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *Journal of Econometrics*, **141** (2), 1073–1099 (2007).
8. Hovemann, G., Wicker, P.: Determinants of sport participation in the European Union. *European Journal for Sport and Society*, **6** (1), 51–59 (2009).
9. Humphreys, B., Ruseski, J.E.: The economic choice of participation and time spent in physical activity and sport in Canada, *Working Paper* No 201014. Department of Economics, University of Alberta (2010).
10. Iannario, M.: Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications*, **21**, 1–22 (2012).
11. Iannario, M., Piccolo, D.: A generalized framework for modelling ordinal data. *Statistical Methods and Applications*, **25**, 163–189 (2016).
12. Lera-López, F., Rapún-Gárate, M.: The demand for sport: sport consumption and participation models. *Journal of Sport Management*, **21**, 103–122 (2007).
13. Peretti-Watel P., Beck, F., Legleye, S.: Beyond the U-curve: the relationship between sport and alcohol, cigarette and cannabis use in adolescents. *Addiction*, **97**, 707–716 (2002).
14. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85–104 (2003).

Being young and becoming adult in the Third Millennium: definition issues and processes analysis

Do Social Media Data predict changes in young adults' employment status? Evidence from Italy

I Social Media data possono predire variazioni dello stato occupazione giovanile? Un caso studio in Italia

Andrea Bonanomi and Emiliano Sironi

Abstract This study addresses the conditions in which young Italian people find themselves during the active job search via Social Media Data. Focusing especially on NEETs, the aims of the study are: 1) to predict the changes in employment status from traditional information by using the longitudinal representative survey Rapporto Giovani; 2) to identify the target population inferring from their online digital traces; 3) to predict changes in employment status from social media data. We tried to predict the employment status transitions based on the digital behaviour using a new Facebook application, *LikeYouth*, that gathers information regarding Facebook profile and Likes on Facebook Pages of each user.

Abstract *Questo lavoro affronta le condizioni dei giovani italiani durante ricerca attiva di lavoro, mediante l'utilizzo di social media data. Concentrandoci in particolare sui NEET, gli obiettivi dello studio sono: 1) prevedere i cambiamenti nello stato occupazionale dalle informazioni tradizionali tratte dall'indagine longitudinale rappresentativa "Rapporto Giovani"; 2) identificare la popolazione target inferendola dalle tracce digitali online; 3) prevedere i cambiamenti nello stato occupazionale dai social media data. Abbiamo cercato di prevedere i cambiamenti nello stato di occupazione in base al comportamento digitale mediante l'uso di una nuova applicazione Facebook, LikeYouth, che raccoglie informazioni sul profilo di Facebook e i "like" alle pagine di Facebook di ciascun utente.*

Key words: Social Media Data, Employment Status, Rapporto Giovani, LikeYouth.

[1] Andrea Bonanomi, Università Cattolica del Sacro Cuore di Milano; andrea.bonanomi@unicatt.it
Emiliano Sironi, Università Cattolica del Sacro Cuore di Milano; emilano.sironi@unicatt.it

1 Introduction

Transition to adulthood evolved over the past decades. While in archaic societies most transitions took place by discrete “leaps” and were signed by “rites of passage”, modern societies present a different approach: transitions are represented increasingly as being individual. Since nowadays the transition to adulthood has become longer, the family of origin has assumed more value and power in influencing young people. Italians leave home on average at a later age with respect to young people in other European countries. The peculiarities of the Italian situation can be explained from both cultural and structural standpoints. On one hand, the presence of strong inter-generational ties is coherent with longer stays in the family of origin. On the other hand, the unfavourable labour market and the welfare system that is not generous to young generations tend to discourage individual autonomy and an active job search (Alfieri et *al*, 2015). According to the Organisation for Economic Co-operation and Development, on April 2016, the rate of youth unemployment in Italy was equal to 37.7%. This study aims at improving the current understanding of the conditions in which young Italians find themselves in during the acquisition of autonomy and in the active job search via tradition survey and tradition type of analysis, and via Social Media Data and Digital Behaviours. We place the focal point on the *Not in Employment, Education, or Training* population, hereafter NEET and on the grade of Autonomy that Young Italians have (i.e. in the decision of leaving home). NEETs are at risk for social marginalization. The failure to tap into the economic aspirations limits not only their income and skill development, but also their likelihood of later employability, autonomy, and life planning. In light of this picture, addressing the transition from and to the NEET condition in Italy is becoming an emerging issue in order to reduce the risks of social exclusion and of entering poverty for a large part of Italian young adults. The aim of the paper is to focus on the dynamic to enter or exit the NEET condition through the study of its determinants. To integrate demographic data from social surveys, with social media information, the specific aims of the study are threefold:

1) to predict the changes in employment status from information obtained by the longitudinal representative survey Youth Report (Rapporto Giovani, Istituto Toniolo, 2016).

2) to identify the target population inferring from their online digital traces and to uncover digital behaviours of the community of interest easily accessible from online social platforms, which can then be used as indicators of the most privileged communication channels for unemployment or educational advertising campaigns;

3) to predict changes in employment status from social media data. For this last aim, we tried to predict the changes in employment status in NEET population based on the digital behaviour (especially on Facebook's *likes*) using a new Facebook application, *LikeYouth*. Having obtained the participants' informed consent, the application gathers information regarding their public Facebook profile and their *likes* on Facebook Pages.

2 Employment status change: an explicative model. Data and Methods

To address the determinants of the changes in the employment status of Italian Young Adults, we use data from the database of the "Rapporto Giovani" survey renewed in 2015 by the Toniolo Institute for Advanced Studies with the involvement of the CARIPLO Foundation and IPSOS LTD as executive partners. The initial sample consists of 9,358 individuals aged between 18 and 29 years. The individuals were chosen with a stratified sampling technique and they are representative of young adults residing in Italy (for gender, age, geographical origin, education, marital status, etc.). For describing the evolution from and to the NEET status the survey has been repeated interviewing the same subjects in 2016 (on 6,172 cases) and in 2015 (on 3,034 cases). We implemented two binary logistic regression models to investigate respectively the determinants of 1) the transition from the condition of NEET to the condition of Not NEET between 2015 and 2017 and of 2) the transition from the condition of Not NEET to the condition of NEET in the same time window. In the first model, the dependent variable is a categorical indicator taking value 1 if the respondent, who was not engaged either in education or in training at the time of the first interview (2015), left the NEET condition over the following 24 months. Conversely, the second model focuses on the sub sample of individuals that were in a Not NEET condition at the time of the first interview: the dependent variable is a categorical indicator taking value 1 if the respondent entered the NEET condition over the following 24 months. The variables, measured at the time of the first interview, used in the empirical analysis to predict the condition of NEET are listed as follows: *Gender, Age, Geographical Area Respondent's education, Parents' education, Quality of relationship with parents, Support, Autonomy.* This group of variables has been included in the model in line with previous studies (Alfieri et al., 2015) devoted to investigating the long run predictors for the NEET condition. In addition, we add a *Life Satisfaction,* derived from Diener et al. (1985). Empirical results displayed in Table 1 show interesting and informative results with respect to process of transition who leads individuals in or out of the condition of NEET. Estimates not necessarily show opposite patterns, underlining different drivers for the two transitional processes examined. The first model focuses on the transition from a condition of NEET to the condition of Not NEET, displaying the significance of age groups and of the gender. In more detail, younger individuals are more likely to leave the NEET condition, being constant the other variables, and to enter in labor market or in the education system. Females seem to be more penalized in leaving the condition of NEET instead.

With respect to the effect of geographical heterogeneity, as expected, it looks clear that individuals belonging to the northern regions have greater opportunity to leaving the condition of NEET, due to better conditions in labor market and to an improved macroeconomic context. If the geographical context is fully explanatory in predicting the permanence or less in the NEET condition, a key determinant seems to be the education level of the respondents: more educated individuals are more likely to exit from the NEET conditions. These results fully support the importance of education in promoting the social mobility, also in the Italian context.

With respect to the second model, the main results show that individuals that live in the south of Italy experience a higher risk of becoming NEET. These results show a profile that is complementary to what we observed in the first model. Once again, the respondents' education plays a decisive role in reducing the risk of becoming NEET: lower educated individuals seem to be more exposed to a change in their employment status. Finally, we observe a relevant result concerning the role of the life satisfaction index, only in predicting the entrance to the NEET condition. Individuals that are less satisfied with their life, even if employed or students at the time of the first interview are more likely to fall in the NEET condition in the following 24 months. Surprisingly, the education level of the parents does not have a significant impact in modifying the respondents' employment status, such as the quality of the relationship with the parents, the autonomy and the support indices.

Table 1: Logistic regression: prediction of the transition from and to the condition of NEET

Variables		***NEET → Not NEET***	***Not NEET → NEET***
Age	< 25	ref.	ref.
	≥ 25	-0.424**	-0.013
Gender	Males	ref.	ref.
	Females	-0.519**	0.081
Geographical	North	ref.	ref.
Area	Centre	-0.516*	-0.102
	South	-0.561*	0.584*
Respondents'	Lower	ref.	ref.
Education	Intermediate	0.651*	-0.639**
	Higher	1.351*	-0.662**
Parents'	Lower	ref.	ref.
Education	Intermediate	0.301	0.108
	Higher	0.265	-0.160
Relationship with parents		0.027	0.001
Autonomy		0.063	0.037
Support		0.177	0.070
Life satisfaction		0.006	-0.231*
Intercept		-1.624**	-2.045*
Observations		1,101	4,784

** p<0.05, ** p<0.01, * p<0.001*

3 *LikeYouth* Project: data collection and method

The information gathered for this project originates from two different sources; from the representative survey of the "Rapporto Giovani" and from the *LikeYouth* application. The sample consists of 9,358 individuals aged between 18 and 32 years (M = 25.7, SD = 4.7). Every subject, at the end of the survey, was invited to access to *LikeYouth*. The population gathering from *LikeYouth* consists of 1,858 individuals. The comparison between subjects accessing *LikeYouth* and subjects not accessing *LikeYouth* does not show significant differences respect to the most important demographic variables (i.e: Total Sample: mean age 25±4.7, 50.8% male, 42.0% resident in the North of Italy, 18.0% with a high education level, 80.8% single, 19.9% NEET; Participants of *LikeYouth*: mean age 25.8± 4.4, 48.9% male, 39.7% resident in the North of Italy, 18.7% with an high education level, 80.4\% single, 21.9% NEET). No statistical concordance index shows a significant difference also for psychometric indicators. Participants of *LikeYouth* may, therefore, be considered a representative sample of the Italian youth population.

To this extend we designed a generic experimentation schema, aiming to assess and compare the predictive power of the digital behaviors. We postulated the study as a supervised classification process, automatically identifying the employment status and the potentiality that a specific person belongs to a category or not (i.e. NEET or NOT NEET), inferring only from their digital data. Additionally, we trained a predictive model for each of some demographic attributes. We employed a widely used ensemble learning method for classification was employed, namely a Random Forest (RF) classifier (Breiman, 2001). We evaluate our performance with the weighted area under of the receiver-operating characteristic (AUROC). The weighted AUROC statistic (Li and Fine, 2010) was preferred over the commonly used *accuracy* metric, since the former takes into account the effect of unbalanced labels, which holds true for most of our attributes. All the prediction scores reported through this paper are in terms of the weighted AUROC.

The entire feature space, **X**, is randomly shuffled and split in two mutually exclusive sets, the training set *Tr*, (80% of **X**) and the testing set *Ts* (20% of **X**). Furthermore, for each target variable we estimated the relative rank (i.e. depth) of each feature, as emerged from the "Gini" impurity function, assessing in this way the relative importance of a specific feature to the predictability of the target variable. We use as predictors the Facebook Pages the users visited and the respective Categories, as defined by the Facebook metadata. This information is represented as a sparse user-page and user-category matrix, the entries of which were set equal to the raw counts of visits and 0 otherwise.

3.1 Results of Prediction of NEET Status via Social Media Data

1,858 young people from the "Rapporto Giovani" visited, during their "history" on Facebook, more than 330,000 different Facebook pages, grouped (from Facebook information) in 155 categories. The assignment of the category is determined, at the time of the making the page, by the administrator of the page itself. It is not, therefore, an objective assignment, but a subjective decision of the developer of the page. Thus, a new taxonomic categorization of the initial categories was made, creating 12 macro categories for a quantitative analysis of the results, and about thirty categories for a more qualitative interpretation of the same. On average, every subject has liked almost 400 Facebook pages, belonging to about 50 different categories. For the first area and community of interest, the NEETs, we are able to predict whether an unknown person is potential NEET or not with accuracy 63%. The model is trained on the population originating from the "Rapporto Giovani" survey for which there has been a manual labelling of the NEET status characterisation by a field expert. For a more qualitative and interpretative analysis of data, we chose to consider only a subset of the Facebook pages, that is those pages that have been visited by at least 20 young people in our sample and that had at least 100,000 *likes*, that is, they were very popular and well-known pages in Facebook. 2,422 pages satisfied the two conditions. They were the predictors of the models in the second part of the analysis. Different data analysis techniques have been tested: Cohen's *K*, Phi Correlation Coefficient for dichotomous variables, Random Forest with Breiman's algorithm, CART Decision Tree. For each Facebook page, each technique produced a score of "importance". With a meta-analysis approach, the scores were weighted and normalized. At the end of this procedure, for each Facebook page, a normalized index was created in the closed subset [–1; +1]. Scores close to –1 are predictors of the condition of Not NEET, scores close to +1 are predictors of the NEET condition. Top indicators of NEET status are Facebook pages or categories related to consumer goods, in particular food, beverage, beauty and health (in particular baby and kids goods), TV Channel, Retail and Consumer Merchandise. The use of Facebook appears as a leisure tool and not much as utility or service. Top indicators of Not-NEET status are Facebook pages/categories related to travel, culture, humour and satire, performance art, news media and politics. Further models are being developed for particular cohorts of the population of interest (male/female, under/over 24 years). Table 2 shows some particularly predictive pages of the two-different status.

Table 2: Most predictive Facebook pages of NEET or Not-NEET Status

Cohorts	***NEET***	***Not-NEET***
Males	Worky.biz	Expo2015
	Verissimo	Repubblica.it
Females	Prenatal	MatteoRenziCheFacose
	Lidl Italia	Grey's Anatomy
Under 24 years	Just Cavalli	Dr. House
	Girella	Barack Obama
Over 24 years	Humana Italia	Il Milanese Imbruttito
	ScontieBuoniAcquisto.it	Report

In general, NEETs visit discount pages, promotions, prize competitions, offers, which can be explained for various reasons. Certainly, NEETs have more free time than workers and students, so they can spend their time searching this kind of pages on Facebook. Moreover, they are pushed by their weaker economic situation, to ensure themselves an acceptable quality of life. Not NEETs have a more active and conscious digital behaviour, and typically they like pages of culture, information, satire, so not closely linked to advertising and media exposure, while NEETs visit more pages linked to the Consumer Merchandise world, and therefore linked to massive advertising campaigns on Facebook.

3.2 Results of Prediction of the Change of Employment Status via Social Media Data

Since we were interested in the change of employment status in three years (from 2015 to 2017), we considered two types of changes: from the NEET Status to the Not-NEET Status and vice versa, from Not-NEET to NEET. We considered only subjects who, in Rapporto Giovani Survey, have responded at least to two surveys, so to have a longitudinal view, and we divided the population of *LikeYouth* in two subgroups, the NEET and the Not-NEET. In the first groups (N=300), we considered the subjects who have changed the employment status, from the NEET condition to the Not-NEET one. In the second groups (N=1,171) we considered the subjects who have changed the employment status, from the Not-NEET condition to the NEET one. Table 3 shows some particularly predictive pages of the two different changes of status and the AUROC statistics. In the first subgroups, the percentage of the subjects have changed their status is equal to 36.3%, in the second one only 7.2% of the subjects have changed their status. The AUROC statistics are equal, respectively, to 0.57 and 0.55. The information of the most predictive pages of change of status can be particularly useful in order to target campaigns directly within these pages, so to reach the target set with high probability of success. The low values of AUROC statistics are due, probably, by the reduced sample size of these two models, which require significantly larger sample sizes, but they can justify further studies and investment in this project.

Table 3: Most predictive Facebook pages of Change of Status

	NEET → Not-NEET (N=300, % of Change of Status = 36.3%)	***Not-NEET → NEET (N=1171, % of Change of Status = 7.2%)***
	Accessorize Italy	Chi l'ha visto Rai Tre
	Global Test Market	Omino Bianco
	H&M	Patrick Dempsey
	CESARE CREMONINI	PayPal
	Il Piccolo Principe	Kiko Milano
	Le frasi più belle di Luciano Ligabue	Universitari Esauriti
	Starbucks	Federica Panicucci official page
	Radio Capital	Perlana
	Caparezza	Eminem
	Grey's Anatomy Le frasi più belle	Ficarra e Picone pagina ufficiale
	Gli Autogol	Scontiebuoniacquisto.it
AUROC	*0.57*	*0.55*

4 Conclusions

The study of the employment status change was conducted by two different approaches. In the traditional one, by using two Logistic Regression Models to determine the antecedents of the changes in the employment status (from NEET to Not-NEET and vice versa) with the database of the "Rapporto Giovani", it emerges that a key determinant seems to be the education level of the respondents: more educated individuals are more likely to exit from the NEET conditions. These results support the importance of education in promoting the social mobility. Moreover, we observe a relevant inverse result concerning the role of the life satisfaction index, only in predicting the entrance to the NEET condition. The second approach is an innovative one, by using social media data as predictors of the changes, with the *LikeYouth* database, a new Facebook App. Since taking advantage of the Facebook platform popularity, it can reach out to the population in need with relatively limited economic and temporal requirements. The challenges emerging from this approach are related to the engagement strategies. The analysis conducted by LikeYouth on 1,858 young people of the "Rapporto Giovani" panel gave comforting results, both on the sample's representativeness and on the predictive and classifying performance of modelling used on Social Media Data. This analysis also showed the passivity of the young NEETs, which, compared to other peers, appear to be more passive, less entrepreneurial and less oriented to cultural, educational and information interests. However, especially in the males' cohort, a part of young people visits pages related to labour market and job search portals. Therefore, they begin to implement possible strategies to get out of their status of inactivity. It is important overturning the vision that society and politics often have on the digital social networking tool. They could use as a sort of social activator. The extreme flexibility of social networks has to be exploited in its full potential.

References

1. Alfieri, S., Sironi, E., Marta, E., Rosina, A., & Marzana, D. (2015). Young Italian NEETs (Not in Employment, Education, or Training) and the influence of their family background. Europe's journal of psychology, 11(2), 311-322. doi:10.5964/ejop.v11i2.901
2. Breiman, L. (2001). Random forests. Machine learning, 45(1):5-32. doi:10.1023/A:1010933404324
3. Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. Journal of personality assessment, 49(1), 71-75. doi:10.1207/s15327752jpa4901_13
4. Istituto Toniolo (2016). La condizione giovanile in Italia. Rapporto Giovani. Il Mulino
5. Li, J. and Fine, J.P. (2010). Weighted area under the receiver operating characteristic curve and its application to gene selection. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(4):673-692. doi.org/10.1111/j.1467-9876.2010.00713.x

Parenthood: an advanced step in the transition to adulthood

Diventare genitori: una fase avanzata della transizione allo stato adulto

Cinzia Castagnaro, Antonella Guarneri and Eleonora Meli

Abstract Parenthood constitutes a crucial stage and one of the most challenging phases of the life cycle. Timing of the transition to adulthood has changed across cohorts according to two different patterns, housing independence and family formation. Postponement of childbearing appears a rational choice for women to assure higher educational level and labour market participation, but it refers also to a value and sociocultural change.

A remarkable postponement is taking place in Italy: TFR rapidly lowers as can be seen by fertility trends in cohort analysis. By cohorts, the average number of children per woman in Italy decreases: 2.5 in the early 1920s (mean age at childbearing 28.6 years), 2 of the post-war generations (mean age 27.0) and the estimate of 1.44 of the 1976s (mean age about 31). At the same time, the proportion of women without children doubled from 1950 (11.1%) to 1976 (21.8%, estimate). Women without children between 18 and 49 years are about 5.5 million, almost half of the women of this age group.

The aim of this paper is to explore the transition to adulthood, analysing reproductive behaviours. The approach adopted is by cohort parity. A special focus is devoted to characterize childless people.

Abstract *La genitorialità costituisce una tappa cruciale e una delle fasi più determinanti del ciclo di vita. Il momento in cui si realizza la transizione allo stato adulto ha subito forti variazioni attraverso le generazioni considerando distintamente due differenti aspetti, l'indipendenza abitativa e la formazione della famiglia. La posticipazione della maternità appare come una scelta razionale per le*

[1] Cinzia Castagnaro, Istat, cicastag@istat.it
Antonella Guarneri, Istat, guarneri@istat.it
Eleonora Meli, Istat, elmeli@istat.it

donne per poter raggiungere un alto livello d'istruzione e una maggiore partecipazione nel mercato del lavoro, ma si può inserire anche in un più ampio cambiamento socioculturale e del sistema dei valori.

Contestualmente alla posticipazione della maternità, in Italia il tasso di fecondità totale sta rapidamente diminuendo, così come visibile dall'analisi per coorte. Il numero medio di figli per donna è passato dai 2,5 delle donne nate nei primi anni Venti (con un'età media al parto pari a 28,6) ai 2 delle generazioni successive alla seconda guerra mondiale (età media al parto di 27 anni) per arrivare a 1,44 figli per la generazione del 1976 (età media di 31 anni). Allo stesso tempo la proporzione di donne senza figli è raddoppiata dalla generazione del 1950 (11,1%) a quella del 1976 (stimata pari a 21,8%). Le donne senza figli tra 18 e 49 anni sono circa 5,5 milioni, quasi la metà delle donne di questa fascia di età.

L'obiettivo del lavoro è esplorare la transizione allo stato adulto, analizzando i comportamenti riproduttivi. L'approccio adottato è prevalentemente quello per coorte e un'attenzione particolare sarà dedicata alla caratterizzazione dei childless.

Key words: parenthood, cohort, postponement, childless

1 Introduction

Parenthood constitutes a crucial stage and one of the most challenging phases of the life cycle.

Timing of the transition to adulthood has changed across cohorts according to two different patterns, housing independence and family formation. The most recent cohorts are characterized by increasing school leaving age and delayed entry into the labour market and these delays show a strong effect in the postponement of the steps towards family formation. Postponement of childbearing appears as a rational choice to assure higher educational level (van de Kaa, 1987), labour market participation, but it refers also to a change of values, as an issue linked to gender equity, housing conditions and economic uncertainty.

Referring to childbearing, when a pronounced postponement is taking place, as it is now happening in Italy, the average number of children per woman is rapidly lowered. Because of the impact of changing in the timing of births, a clearer picture of fertility emerges when analysing the cohort trends.

During the latest fifty years women enter in motherhood at increasingly later ages. The delay in childbearing (Kohler, Billari, Ortega, 2002, Sobotka 2004) has an effect for many women to remain childless, delay childbearing at ages when women's fecundity is already in decline (te Velde, Pearson, 2002, te Velde *et al.*, 2012).

2 Aims, data and methods

The aim of this paper is to explore the transition to adulthood, analysing reproductive behaviours. The approach adopted is by cohort parity. A special focus is devoted to characterize childless people.

To describe the general Italian context about fertility most recent data on births are considered. These individual data refer to all births enrolled in the Italian resident population register; trough this register it’s possible to obtain information on births according to a set of demographic variables about the newborn and their parents. Moreover, to adopt a time series perspective it is worthwhile to analyse data contained in the Italian Fertility Database (IFD), a database built and maintained by the Italian Institute of Statistics (Istat). The IFD contains data on births of the resident population since 1952 and it is updated each year with data drawn from other Istat sources. Data are available at NUTS-2 level and by cohort of mothers (the first complete cohort is referred to 1933). These data underwent several rounds of adjustments, specific for different periods, which allowed reconstructing data by birth order.

The cohort approach seems to be particularly suitable to describe differential characteristics of women. The idea is to put in comparison some specific birth cohorts to single out the different attitudes and the main effects of the timing-quantum interaction.

Using data of the last Italian Multipurpose Household Survey “Families, social subjects and life course” (FSS), carried out by Istat in 2016, it is possible to shed light on reproductive behaviours and choices, considering the main individual socio-demographic characteristics of people and the crucial steps of their life cycle.

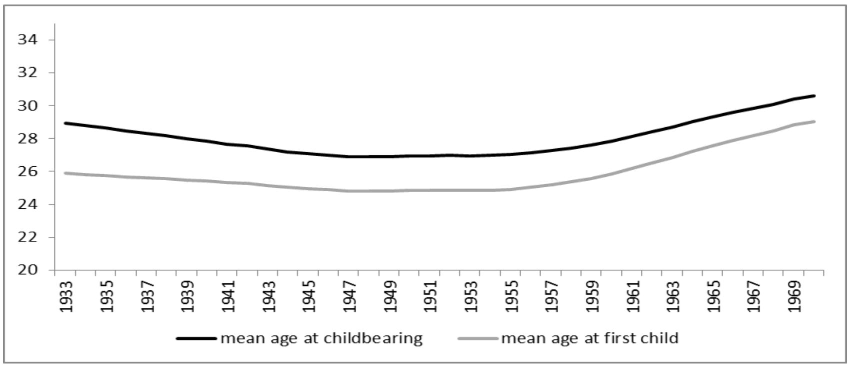
In order to study the factors influencing these different paths and family behaviours, the analysis draws differential profiles of childless and childfree people.

3 The cohort approach: *timing* and *quantum*

The decision of fulfilling the reproductive process is very important in the life history of a parent, as it indicates a long-term project that involves not only the parents but also their descendants. This decision arrives later and later in Italy passing from one cohort to another. This postponement is clearly visible observing the increase in mean age at first child and consequently at childbearing (Figure 1).

The general idea is to verify if the postponement lead to a recovery due to the most recent cohorts compared to the older ones, as observed in some other countries (e.g. Norway, Sweden or Western Europe countries). Theoretically the late-starters in most recent cohorts could achieve slightly more children than the late starters in the earlier cohorts (Berrington, Stone and Beaujouan, 2015). However this recovery does not counterbalance, in the Italian case, the collapse in fertility affecting younger ages in most recent cohorts. Across generations to show a deep change it is not only the timing but also the quantum.

Figure 1: Mean age at first child and mean age at childbearing by mothers' cohort



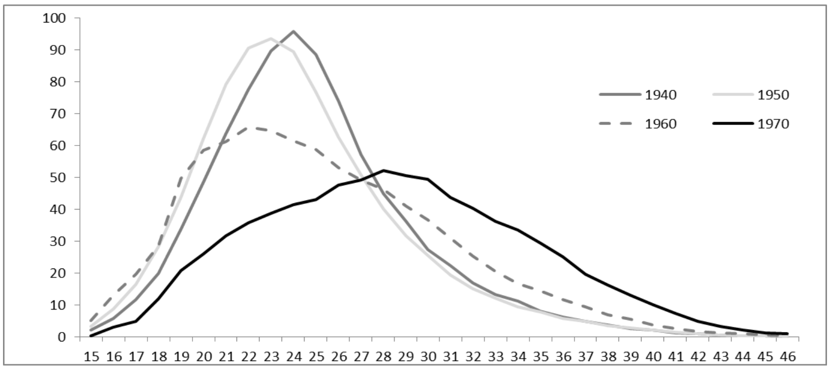
Source: Istat

The average number of children per woman in Italy decreases and the mean age at childbearing shows a U-shaped pattern: from 2.5 in the early 1920s cohort (mean age at childbearing 28.6 years), to 2 of the post-war generations (mean age 27), to the estimate of 1.44 of the 1976 cohort (mean age about 31).

A noticeable decrease in fertility necessarily entails profound modifications in terms of composition of the offspring by birth order. The fertility rates referring to the births of the first order have undergone a relatively limited variation, at least up to the generations of women in the mid-1960s: from 0.89 first children for 1950 women to 0.87 for 1965 cohort. For younger cohorts, more evident changes are observed. As a matter of fact, the proportion of women without children doubled from 1950 (11.1%) to 1976 (21.8%, estimates).

The age profiles display a clear postponement attitude observing 1960 cohort and particularly 1970 cohort (Figure 2).

Figure 2: Fertility rates at first child by mothers' age and cohort



Source: Istat

The contribution of different cohorts led to only 464 thousand births estimated in 2017, over 22,000 less than 2015 and over 100,000 less than 2008 (commonly considered as the first year of the economic crisis). Economic crisis reflected significantly on family's formation. A new phase of decrease in the birth rate that started after 2008 shows a strong contraction of the first children, from 283.922 in 2008 to 227.412 in 2016.

The decline in births is partly due to the so-called "structural" effects induced by significant changes in the female population in the fertile age, conventionally set between 15 and 49 years. The effect of the age structure is responsible for almost three quarters of the difference in births observed between 2008 and 2016. The remaining share depends instead on the decrease in the propensity to have children (from 1.45 children per woman at 1.34).

Women have accentuated the postponement of reproductive experience towards advanced ages; compared to 1995, the year of minimum fertility (1.19 children on average per woman), the mean age at childbearing increases by almost two years, reaching 31.8 years; even the mean age at first child grows to 31 years in 2016 (almost three years more than in 1995).

The delay in childbearing is likely to be the cause for many women to remain childless; as a matter of fact, the postponement at later ages may become a renunciation in having children. As above mentioned, a part of childless women chose childlessness, so it has to be considered on one hand as a personal preference, and on the other hand motherhood postponement as a choice for career possibility or other life goals (Blossfeld, Hiunink, 1991, Andersson 2000, Kneale, Joshi 2008). In order to consider the personal choice in not having children, childlessness has to be studied as childfree condition (Houseknecht, 1979).

4 To be childless or childfree: similarities and peculiarities

The consistent increase in the number of women without children for the younger generations raises many questions about its interpretation and its impact on the future evolution of fertility. Is this an increase due to difficulties in carrying out family projects or adopting a different life model that does not envisage becoming a parent? Using data from the Italian survey "Families, social subjects and life course" it is possible to analyse childless women characteristics (Istat, 2017). Considering women aged between 18 and 49, the childless women are about 5 and a half million in 2016, 45.1% of the women of this age group. Considering men, almost 57.4% of them are childless in the same age group (around 7 million) (Istat, 2017). The younger they are the higher the proportion of people without children: over 55% of people younger than 35 have no children, this proportion decreases first for women (less than one to three is childless over 35 years old, while for men the same proportion is reached over 45 years old).

Among childless people, about 534 thousand declared that having children does not fit into their life project (equal to 2.2% of people 18-49 years old); for men the rate is higher than for women (2.5% vs 1.8%), especially in the 35-44 age group (3.5% among men between 35 and 39 years and 2.9% of 40-44 year-olds). For women the most high is between 40-44 year-olds (2.8% say they do not have and do not want children) (Table 1).

These results show that the phenomenon of women (and couples) without children by choice is very limited in our country and that, conversely, the increase in the share of women without children are mostly due to obstacles to the implementation of family projects; the effect of the postponement that can be transformed into renunciation should not be overlooked with the approach of the most advanced ages of women's reproductive life.

Table 1: Childless and childfree people by gender and age (per 100 people with the same characteristics) - Year 2016

Gender	Age groups	Childless	Childfree
Male	18-24	98.4	2.7
	25-29	88.9	1.5
	30-34	64.6	1.9
	35-39	46.4	3.5
	40-44	32.3	2.9
	45-49	28.1	2.3
	18-49	57.4	2.5
Female	18-24	93.3	1.9
	25-29	73.9	1.6
	30-34	46.8	1.3
	35-39	27.6	1.4
	40-44	24.2	2.8
	45-49	19.8	1.6
	18-49	45.1	1.8
All	18-24	95.9	2.3
	25-29	81.7	1.6
	30-34	55.5	1.6
	35-39	37.1	2.4
	40-44	28.2	2.8
	45-49	24.0	2.0
	18-49	51.3	2.2

Source: Istat

Analysing the individual characteristics, as well known people with higher educational qualifications are less likely to have children (61.3% for males and 55.5% for females) compared to the lower educational level (49.4% for males and 34.0% for females). The highest quota of people that do not have parenthood in their life project is among men and women with university qualifications.

Among unemployed people there is the highest share of childless (64.5%) and childfree (3.2%). Reproductive projects of unemployed people change according to gender showing opposite patterns: the highest quota of childfree is for men (5.0%) whereas the lowest is for women (0.8%); in addition employed women register the highest proportion of childfree (2.5%).

The presence or absence of a partner has a significant impact on reproductive choices. Being single makes more likely to be childless or childfree (respectively 86.7% and 3.8%) as well as having a partner with whom one is not cohabiting (84.8% and 3.0%); these are all elements that characterize people who do not have children and do not intend to have children, because they do not fit into their life projects, to a greater extent if they are men.

Besides the sentimental situation, the role that these people play in their family provides information on the relapse that the phase of the life cycle has on parental planning. People living alone are less likely to be a parent or plan parenthood (respectively 80.0% and 4.9%).

In a gender perspective, 77.9% of single men have no children and 5.5% do not plan fatherhood, because the idea does not fit with their life's plans. On the other hand among women there are higher shares of childless for singles (83.5%) and childfree when they do not have core relationship within a family (9.1% and 4.3% of women in a couple without children).

5 Conclusion and future developments

The childless population shows a high degree of heterogeneity (Tocchioni, 2017). Only a part plans never to have children (Tanturri and Mencarini 2008), and not all of them do not change idea across time (Moore, 2017).

To draw the profiles of different categories of people without children a logistic model will be applied to estimate the probability to be childfree vs childless (not childfree). The analysis will focus on women and men who are at least 35 years old at the time of the interview, considering only people who had already 'crossed' their most fecund period.

Main covariates to be included in the model are: socio-demographic characteristics, type of couple, support networks, economic and health conditions. An important covariate is the education level; as a matter of fact, education plays an important role in the reproductive behaviour.

The phenomenon of postponement of childbirth is even more evident for mother with a high educational level, thus causing an overlap or, sometimes, an inversion of family making steps. These different paths to parenthood could be the focus of further analyses exploiting FSS data 2016 edition.

References

1. Andersson, G.: The impact of labour-force participation on childbearing behaviour: Pro-cyclical fertility in Sweden during the 1980s and the 1990s. European Journal of Population/Revue européenne de démographie 16.4, 293-333 (2000)
2. Berrington, A., Stone, J. and Beaujouan, E.: Educational differences in timing and quantum of childbearing in Britain: A study of cohorts born 1940−1969. Demographic Research, 33 (26), pp. 733-64 (2015)

3. Blossfeld, H, Huinink, J.: Human capital investments or norms of role transition? How women's schooling and career affect the process of family formation. American journal of Sociology 97.1, 143-168 (1991)
4. Houseknecht, S.K.: Timing of the decision to remain voluntarily childless: Evidence for continuous socialization. Psychology of Women Quarterly 4.1, 81-96 (1979)
5. Istat: Natalità e fecondità della popolazione residente". Statistiche Report (2017)
6. Kneale, D., Joshi H.: Postponement and childlessness: Evidence from two British cohorts. Demographic research 19, pp. 1935-1968 (2010)
7. Kohler, H., Billari, F.C., Ortega, J.A.: The emergence of lowest-low fertility in Europe during the 1990s. Population and development review 28.4, 641-680 (2002)
8. Moore, J.: Facets of agency in stories of transforming from childless by choice to mother. Journal of Marriage and Family 79(4), 1144–1159 (2017)
9. Sobotka, T.: Postponement of childbearing and low fertility in Europe. Dutch University Press (2004)
10. Tanturri, M.L., Mencarini, L. Childless or childfree? Paths to voluntary childlessness in Italy. Population and Development Review 34(1), 51–77 (2008)
11. te Velde, E.R., Pearson, P.L.: The variability of female reproductive aging. Human Reproduction Update, 8, 141-154 (2002)
12. te Velde, E., Habbema D., Leridon H., Eijkemans M.: The effect of postponement of first motherhood on permanent involuntary childlessness and total fertility rate in six European countries since the 1970s. Hum Reprod. Apr;27(4), 1179-83 (2012)
13. Tocchioni, V.: Exploring the childless universe: Profiles of women and men without children in Italy". Demographic Research, Volume 38, Article 19, Pages 451-470 (2018)
14. van de Kaa, D.J.: Europe's Second Demographic Transition. Population Bulletin, 42 (1), Washington, The Population Reference Bureau (1987)

Economic Statistics and Big Data

Improvements in Italian CPI/HICP deriving from the use of scanner data

Alessandro Brunetti, Stefania Fatello, Federico Polidoro, Antonella Simone[1]

Abstract Scanner data are a crucial innovative "big data" source to estimate inflation, providing several advantages that derive from the detailed information available about sales and quantities at weekly frequency, GTIN by GTIN, outlet by outlet throughout the entire national territory. In paragraph 1 the paper makes the point about the state of play of the use of scanner data (introduced for grocery products in Italian CPI/HICP in 2018) and evaluates the benefits in terms of accuracy of inflation estimation coming from the improvement of the territorial coverage. The data of a sample of more than 1,700 hyper and supermarkets, for years 2017 and 2018, have been processed in order to calculate price indices differentiated according to outlet location (inside and outside municipal borders of the provincial chief towns) and price indices for the 80 provincial chief towns previously involved, to be compared with the indices calculated for the entire Italian territory (paragraph 2). The results in terms of improved accuracy are analyzed at national and geographical area level (paragraph 3). Perspectives of Italian scanner data project (brought forward by ISTAT) are finally sketched in paragraph 4.

Key words: Scanner data, inflation, accuracy, territorial coverage

1. The use of scanner data to estimate inflation in Italy: the state of play

Starting from January 2018 ISTAT introduced scanner data of grocery products (thus excluding fresh food) in the production process of estimation of inflation. This

[1] Alessandro Brunetti, ISTAT, albrunet@istat.it
Stefania Fatello, ISTAT, fatello@istat.it
Federico Polidoro, ISTAT, polidoro@istat.it
Antonella Simone, ISTAT, ansimone@istat.it

innovation concerns 79 aggregates of product belonging to 5 ECOICOP Divisions (01, 02, 05, 09, 12).

Since the end of 2013 a stable cooperation was established among ISTAT, Association of modern distribution, retail trade chains (RTCs) and Nielsen. Scanner data of grocery products have been collected by ISTAT through Nielsen for years 2014, 2015 and 2016 for about 1400 outlets of the main six RTCs for 37 provinces.

Afterwards, in view of the inclusion of scanner data into price indices calculations, a probabilistic design has been implemented for the selection of the sample of outlets, for which Nielsen provided ISTAT from December 2016. Scanner data for 1.781 outlets (510 hypermarkets and 1.271 supermarkets) of the main 16 RTCs covering the entire national territory are monthly collected by ISTAT on a weekly basis at item code level. Outlets have been stratified according to provinces (107), chains (16) and outlet-types (hypermarket, supermarket) for a total of 867 strata, taking into account only the strata with at least one outlet. Probabilities of selection were assigned to each outlet based on the corresponding turnover value. Table 1 shows the number of the strata, the number of the outlets and the coverage in terms of turnover, at regional and national levels for years 2018. The coverage for the year 2017 is slightly lower because a small RTC has been excluded from the analysis.

Concerning the selection of the sample of items, a static approach that mimics traditional price collection method has been adopted[1]. Specifically, a cut off sample of barcodes (GTINs) has been selected within each outlet/aggregate of products (covering 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover). The products selected in December are kept fixed during the following year. A “thank” of potentially replacing outlets (258) and GTINs (until a coverage of 60% of turnover within each outlet/aggregate) has been detected in order to better manage the possible replacements during 2018.

About 1.370.000 price quotes are collected each week to estimate inflation. For each GTIN, prices are calculated taking into account turnover and quantities (weekly price=weekly turnover/weekly quantities). Monthly prices are calculated with arithmetic mean of weekly prices weighted with quantities.

Scanner data indices of aggregate of products are calculated at outlet level as unweighted Jevons index (geometric mean) of GTINs elementary indices. Provincial scanner data indices of aggregate of products are calculated with weighted arithmetic mean of outlet indices using sampling weights. Finally, for each aggregate of products, scanner data indices and indices referred to other channels of retail trade distribution are aggregated with weighted arithmetic mean using expenditure weights.

To calculate weights for the integration of regional indices of modern and traditional distribution at regional level, data are broken down using regional estimates from National Account (at ECOICOP sub-class level), regional

[1] The static approach to sampling is discussed in EUROSTAT [2017].

expenditure by type of distribution from Ministry of Economic Development and qualitative information on the shopping habits of consumers coming from HBS.

Table 1: Sample size: number of strata, number of outlets and coverage in terms of turnover -Year 2018

	North								Centre				South								
Region	**Piemonte**	**Valle d'aosta**	**Liguria**	**Lombardia**	**Trentino Alto Adige**	**Veneto**	**Friuli Venezia Giulia**	**Emilia Romagna**	**Toscana**	**Umbria**	**Marche**	**Lazio**	**Abruzzo**	**Molise**	**Campania**	**Puglia**	**Basilicata**	**Calabria**	**Sicilia**	**Sardegna**	**ITALY**
Num. of strata	78	4	31	148	12	85	45	84	65	16	43	38	34	11	35	34	6	25	44	29	**867**
Num. of outlets	152	6	62	286	35	161	77	163	142	32	83	106	58	18	88	85	10	51	111	55	**1781**
% market shares (hyper + super)	95,9	76,0	99,8	94,8	99,1	87,0	94,3	98,5	99,9	92,2	97,4	93,2	92,9	96,6	77,5	91,3	67,9	87,2	86,4	98,0	**93,7**

2. The improvements of the territorial coverage of indices and its effect on the accuracy of inflation estimates

Scanner data allow calculating the indexes for the entire national territory using data from outlets of all Italian provinces and located both in the municipal area and outside. With the aim of evaluating the benefits in terms of accuracy of inflation estimation coming from the improvement of the coverage in territorial terms, price indices are calculated by taking into account the outlet location (i.e. inside municipal area of the provincial chief towns: HICP SD MA) and by distinguishing the 80 provincial chief towns previously involved in the consumer price survey from the rest of the provinces whose data now are made available by scanner data (HICP SD 80P).

Table 2 shows the number of outlets used for the calculation of the different indices at national and macro-regional level for years 2018.

Table 2: Number of outlets at national and macro-regional level – Year 2018

Macroregion	**All outlets**		**Outlets in 80 provinces**		**Outlets in municipal area**	
	N° outlets	***% outlets***	***N° outlets***	***% outlets***	***N° outlets***	***% outlets***
North	942	52,9	857	58,8	304	50,6
Centre	363	20,4	286	19,6	133	22,1
South	476	26,7	315	21,6	164	27,3
Italy	**1781**	**100,0**	**1458**	**100,0**	**601**	**100,0**

In order to point out the methodology used for this analysis, it is necessary to start with a short description of the procedure for the aggregation of indices[1].

[1] For a detailed description of the procedures adopted by Istat for the calculation of the consumer price indices, see ISTAT [2012].

Let us introduce the following symbols[1]:

- *n* denotes the *n*-th product aggregate[2] (*n*=1,…,*N*);
- *g* denotes the g-th region (g=1,…,G=20);
- j denotes the j-th province (j=1,…,J(r));
- h denotes the *h*-th outlet (*h*=1,…,*H*(*j*)).

Let:

$$P_{ng,j} = \sum_{h \in j} w_{ngj,h} \cdot P_{ngj,h}$$ be the provincial index of the product aggregate *n*;

$$P_{n,g} = \sum_{j \in g} w_{ng,j} \cdot P_{ng,j}$$ be the regional index of the product aggregate *n*;

$$P_n = \sum_{g} w_{n,g} \cdot P_{n,g}$$ be the national index of the product aggregate *n*;

$$P = \sum_{n} w_n \cdot P_n$$ be the general index at the national level.

where:

$$w_{ngj,h} = \frac{e_{ngj,h}}{\sum_{h \in j} e_{ngj,h}} \;;\; w_{ng,j} = \frac{e_{ng,j}}{\sum_{j \in g} e_{ng,j}} \;;\; w_{n,g} = \frac{e_{n,g}}{\sum_{r} e_{n,g}} \;;\; w_n = \frac{e_n}{\sum_n e_n}$$

and $e_{ngj,h}$ is the expenditure estimate for product aggregate *n* in the outlet *h* of province *j* in region *g*[3].

For the scope of the present analysis, the general index *P* is to be compared with the index $\hat{P}$ which is calculated using:

- Transaction prices of the outlets (*h*') situated inside municipal borders of the provincial chief towns (HICP SD MA);
- Transaction prices of the outlets of the 80 provincial chief towns previously involved in the consumer price survey (HICP SD 80P).

Concerning the first case, the general index *P* can be usefully expressed as the weighted arithmetic mean of provincial product aggregate indices:

$$P = \sum_{nj} \frac{e_{nj}}{\sum_n e_n} \cdot P_{nj} = \sum_{nj} \pi_{nj} \cdot P_{nj}$$

[1] The notation used is adapted from that one suggested in Biggeri L., Giommi A. [1987].

[2] Product aggregates indices are indices calculated at the lower level of aggregation of product-offers.

[3] $e_{ngj,h}$ incorporates the sampling coefficient attached to the outlet *h*.

Accordingly, the impact of the improvement of territorial coverage is calculated as follows:

$$P - \hat{P} = \sum_{nj} \pi_{nj} \cdot \left(P_{nj} - \hat{P}_{nj}\right)$$

where:

$$\hat{P}_{n,j} = \sum_{h' \in j} \frac{e_{nj,h'}}{\sum_{h' \in j} e_{nj,h'}} \cdot P_{nj,h'}$$

The impact can also be decomposed as suggested in Biggeri L., Brunetti A. e Laureti T. [2008]. By indicating with k the product $N \cdot J$, with δ_{nj} the difference between sub-indices $\left(P_{nj} - \hat{P}_{nj}\right)$, with $s_{\pi_{nj}}$ and $s_{\delta_{nj}}$ the standard deviations of π_{nj} and δ_{nj}, with $R_{\pi_{nj},\delta_{nj}}$ the linear correlation coefficient between π_{nj} and δ_{nj}, with $\bar{\delta}_{nj}$ the arithmetic mean of δ_{nj}, we have:

$$P - \hat{P} = k \cdot s_{\pi_{nj}} \cdot s_{\delta_{nj}} \cdot R_{\pi_{nj},\delta_{nj}} + \bar{\delta}_{nj}$$

As for the second case, it is convenient to express the general index P as the weighted arithmetic mean of regional product aggregate indices:

$$P = \sum_{ng} \frac{e_{ng}}{\sum_{n} e_{n}} \cdot P_{ng} = \sum_{ng} \pi_{ng} \cdot P_{ng}$$

Consequently, it is possible to write:

$$P - \hat{P} = \sum_{ng} \pi_{ng} \cdot (P_{ng} - \hat{P}_{ng})$$

where:

$$\hat{P}_{n,g} = \sum_{j' \in g} \frac{e_{ng,j'}}{\sum_{j' \in g} e_{ng,j'}} \cdot P_{ng,j'}$$

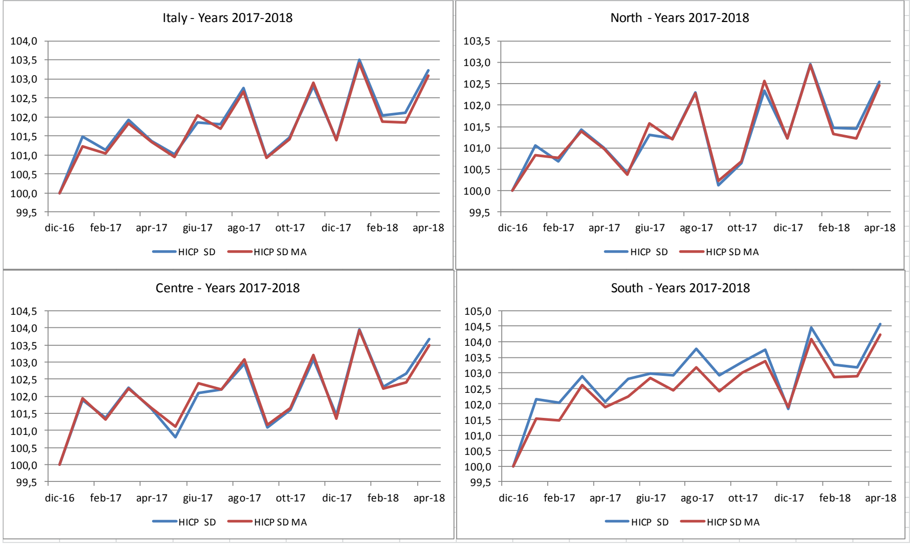
and, with similar notation:

$$P - \hat{P} = k \cdot s_{\pi_{ng}} \cdot s_{\delta_{ng}} \cdot R_{\pi_{ng},\delta_{ng}} + \bar{\delta}_{ng}$$

3. Results

By comparing indices calculated on the whole national territory and the corresponding indicators compiled taking into account only the outlets in the municipal area of the provinces (figure 1) moderate differences emerge in the first months of 2017 and 2018 and in the middle of the first year. However, when the geographical breakdown is considered, the divergences tend to be relatively larger and persistent, especially in the South of Italy (islands included) (table 3). For example, the difference of the indices, calculated on March 2018, shows that the HICP SD MA index of the South is about 0.3 percentage points below the corresponding HICP SD index, while it is 0.24 in the North and 0.15 in the Centre). The main factors explaining this divergence seem to be the relatively higher value of the standard deviation of the differences of sub-indices and the relatively high value of the linear correlation coefficient $R_{\pi_{nj},\delta_{nj}}$ (higher differences in the level of sub-indices tend to have higher weights).

Figure 1: Comparison between HICP SD and HICP SD MA - Years 2017-2018



Italy - Years 2017-2018

North - Years 2017-2018

Centre - Years 2017-2018

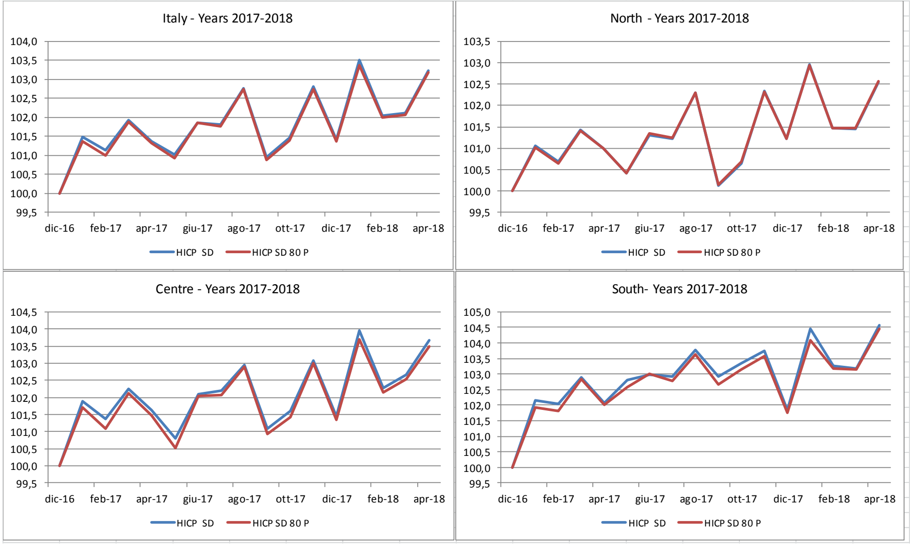
South - Years 2017-2018

Table 3: Decomposition of the difference between HICP SD and HICP SD MA. March 2018.

	Italy	**North**	**Centre**	**South**
K	8.368	3.713	1.738	2.917
$s_{\pi_{nj}}$	0,0002	0,0004	0,0012	0,0005
$s_{\delta_{nj}}$	3,3349	2,9171	2,8525	4,0235
$R_{\pi_{nj},\delta_{nj}}$	0,0139	-0,0088	0,0082	0,0460
$\bar{\delta}_{nj}$	0,1522	0,2732	0,1046	0,0267
HICP SD	100,6830	100,2363	101,2078	101,2760
HICP SD MA	100,4517	100,0011	101,0555	100,9741
HICP SD - HICP SD MA	0,2314	0,2352	0,1523	0,3019

Regarding the comparison between HICP SD and HICP SD 80P, major divergences tend to be concentrated in the South of Italy (figure 2) as well. This result reflects the fact that the share of provinces participating to the survey before the introduction of scanner data, in this part of the country, was relatively low as compare to the North and the Centre.

Figure 2: Comparison between HICP SD and HICP SD 80P - Years 2017-2018



Generally speaking, for what concerns modern retail trade distribution and comparing indices compiled from scanner data source in a time span of 16 months, the improvement in terms of accuracy coming from the coverage of the entire provincial territories are limited to three months at national level with a maximum of three decimal points of differences between grocery index calculated inside the municipal borders and that one compiled with the outlets of the entire municipal areas. For the South the differences between the two indices are more frequent and wider along the time span considered. In all the cases, the level of the indices referred to the entire provincial areas are slightly higher than those ones compiled just within the municipal borders.

If we consider the grocery index compiled taking into consideration the outlets of the 80 provinces that were involved in 2017 in the territorial data collection, the comparison with the grocery index calculated from the data of all the 107 Italian provinces, shows just some local differences (in the South and in the Centre of Italy) but without important consequences in the estimation of the grocery index at national level.

4. Next steps in the development of the ISTAT project on scanner data

Scanner data project (brought forward by ISTAT) is still on the way and it is possible to sketch some further steps.

The first one is the adoption of the so called dynamic approach (abandoning the static one) to the selection of the elementary items (GTINs) to be considered for the compilation of indices. It should be implemented starting from January 2019 and it means the use of all the elementary price quotes of all the GTINs sold monthly in the sample of outlets selected. Next months of 2018 will be dedicated to solve the main crucial issues in sight of this next step, starting from that regarding relaunches and IT environment and procedures. Dynamic approach is the choice towards other National Statistical issues at European level are converging and could represent a further improvement in the accuracy of Italian CPI/HICP.

The following steps concern the extension of the use of scanner data to other retail trade channels (discount, outlets with surface between 100 and 400 square meters) and other goods such as fresh products with variable weight and no grocery products.

Toward these aims the role of the collaboration with the modern distribution and the representative association (Association of Modern Distribution, ADM) keeps its crucial importance.

References

Biggeri L., Brunetti A. e Laureti T. (2008), "*The interpretation of the divergences between CPIs at territorial level: Evidence from Italy*", Invited paper presented at the Joint UNECE/ILO meeting on Consumer Price Indices, May 8-9, Geneva., Proceedings UNECE, Geneva.

Biggeri L., Giommi A. [1987], "*On the accuracy and precision of the consumer price indices. Methods and applications to evaluate the influence of the sampling of households*", Proceedings of the 46th Session of the ISI (International Statistical Institute). Book 2, Tokyo, 134-157.

EUROSTAT [2017], Pratical Guide for Processing Supermarket Scanner Data, available on EUROSTAT website (https://circabc.europa.eu).

ISTAT [2012], Indici dei prezzi al consumo: aspetti generali e metodologia di rilevazione, available on ISTAT website (https://www.istat.it/it/files/2013/04/Indice-dei-prezzi-al-consumo.pdf).

Big data and spatial price comparisons of consumer prices

Big data e confronti spaziali dei prezzi al consumo

Tiziana Laureti and Federico Polidoro

Abstract An accurate measurement of price level differences across regions within a country is essential for assessing inequality in the distribution of real incomes and consumption expenditures. However, systematic attempts to compile sub-national Purchasing Power Parities (PPPs) on a regular basis have been hampered by the labour-intensive analyses required in processing traditional price data. The increasing availability of big data may change the current approach for estimating sub-national PPPs, although only for household consumption expenditures. This paper estimates spatial price indexes using a scanner dataset set up for experimental CPI computations in 2017 and includes information on prices, quantities and quality characteristics of products at barcode level. Our dataset refers to grocery products sold in a random sample of approximately 1,800 outlets across Italy belonging to the most important retail chains (95% of modern retail trade distribution), covering 55.4% of total retail trade distribution for this product category. We use various weighted index formulas for calculating consumption sub-national PPPs at detailed territorial level and at the lowest aggregate level at which no quantity or expenditure weights are usually available. Finally we report preliminary estimates of between-regional spatial price indexes for specific product groups and for Food and Non-Food consumption aggregates.

Riassunto*: Misurare i differenziali di prezzo tra le regioni di un Paese è fondamentale per analizzare la disuguaglianza nella distribuzione dei redditi e dei consumi. Tuttavia, la produzione regolare di Parità del Potere di Acquisto (PPA) a livello sub-nazionale è stata ostacolata dalle difficoltà di utilizzare dati da fonti tradizionali. La crescente disponibilità di big data potrebbe cambiare l'attuale approccio per la stima delle PPA infra-nazionali sebbene solo in riferimento ai*

[1] Tiziana Laureti, Department of Economics, Engineering, Society and Business Organization University of Tuscia, Italy; email:laureti@unitus.it

Federico Polidoro, Living conditions and Consumer prices Unit, Istat; email: polidoro@istat.it

consumi delle famiglie. Il lavoro presenta stime preliminari di indici spaziali regionali utilizzando un dataset di dati scanner costruito per la stima sperimentale degli IPC e include informazioni su prezzi, quantità e caratteristiche dei prodotti a livello di codice a barre per il 2017. Il dataset si riferisce ai prodotti grocery venduti in un campione di circa 1,800 negozi in Italia rappresentativo delle più importanti catene di vendita (95% del fatturato totale della distribuzione moderna e 55,4% del commercio al dettaglio totale di questi prodotti). Si utilizzano diversi metodi con strutture di ponderazione per il calcolo delle PPA infra-nazionali al livello più disaggregato, dove solitamente non sono disponibili informazioni su quantità e spese. Si riportano stime preliminari degli indici spaziali regionali per specifici gruppi di prodotti e per gli aggregati di consumo alimentari e non alimentari.

Key words: regional price levels, spatial price indexes, sub-national PPPs

1 Introduction

An accurate measurement of price level differences across regions within a country is essential in order to better assess regional disparities, thus enabling policy makers to adequately identify and address areas of intervention. Regional values of economic indicators such as Gross Domestic Product (GDP), income and poverty levels, should be adjusted for regional price differences, following the same logic according to which the economic well-being of different countries is compared by taking into account *international purchasing power parities* (PPPs).

At international level, PPPs for countries are compiled by the International Comparison Program (ICP), administered by the World Bank with the collaboration of the OECD, EUROSTAT and other organizations (World Bank, 2013).

Though not as widespread as international comparisons of prices through ICP, there have been research projects and studies on the compilation of spatial price indexes, also referred to as *sub-national purchasing power parities*, carried out by NSOs and individual researchers in various countries including USA, Brazil, India, Indonesia, China, Italy, Australia, New Zealand and the United Kingdom (Biggeri et al, 2017; Laureti and Rao, 2018).

The Italian National Institute of Statistics (Istat) is one of the few National Statistical Offices (NSOs) that carried out official experimental sub-national PPPs computations by using price data from Consumer Price Indexes (CPIs) and ad-hoc surveys and focusing on comparing consumer prices across the 20 Italian regions. Significant price differences emerged in 2010 and in 2008, which encouraged Istat to confirm the project for producing sub-national PPPs on a regular basis (Biggeri et al., 2017). However, systematic attempts to compile sub-national PPPs on a regular basis have been hampered by the labour-intensive analyses required for processing traditional price data, i.e. data used for CPIs, and by the cost involved for carrying out ad-hoc surveys for collecting price data. In this context, the use of big data is both a challenge and a possible solution to more than one difficulty NSOs face when constructing spatial price comparisons worldwide. A range of alternative sources, including scanner data, information obtained from administrative sources or by adopting web-scraping techniques must be investigated for making both temporal and spatial price comparisons. Since the primary use of scanner data is not to measure temporal and spatial price differences, methodological and empirical issues regarding scanner data quality, such as under-coverage for missing outlet types (i.e. hard discount) and products (i.e. perishables and seasonal products) and over-coverage (i.e. business expenditure) should be solved by price statisticians. It is worth noting that, in order to produce sub-national PPPs for the entire household consumption expenditure, other sources should be considered in addition to scanner data.

By focusing on grocery products (for which retail scanner data are currently available), it may be fruitful to use the itemized information contained in scanner data (turnover and quantities for each well-specified item code) in order to compile weighted spatial price indexes at detailed territorial level. When reviewing international practices, it is important to note that even if a fifth of EU countries have been using scanner data for compiling CPIs using different methods, as yet few studies have explored the use of new data sources for compiling spatial indexes (Laureti and Polidoro, 2017). Within the European Multipurpose Price Statistics project, Istat has recently introduced scanner data in the official CPI computation and has been exploring the possibility of using them for compiling sub-national PPPs.

The aim of this paper is to estimate household consumption sub-national PPPs for Italy in 2017 by using a scanner dataset constructed for experimental CPI computation after having carried out data cleaning and trimming outliers processes. This data set refers to a random sample of approximately 1,800 hypermarkets (more than 500) and supermarkets (almost 1,300), concerning the grocery products sold in the most important retail chains (95% of modern retail chain distribution that covers 55.4% of total retail trade distribution for this category of products). Since data are available for the 107 Italian provinces we estimate within-regional and between-regional spatial price indexes for the food and non-food consumption aggregates included in the scanner dataset.

2 Scanner data and spatial price comparisons in Italy

The first and a fairly critical step when compiling household consumption sub-national PPPs is to prepare a long list of goods and services that will be priced in the regions involved in spatial price comparisons. The list used in price comparisons needs to meet and balance the requirements of comparability and representativity. Comparability means that identical products of the same or similar quality should be priced across all the regions so that the PPPs based on these data solely reflect price level differences. Representativity is a concept associated with the relative “importance” of individual products within a group of similar products (called Basic

Heading, BH[1]): ideally, a product’s importance should be determined using information on expenditure.

In this paper we refer to a list of products derived from the dataset described above, that specifically covers 54 grocery product aggregates, belonging to five divisions of the ECOICOP (01, 02, 05, 09, 12). Annual provincial average prices for each item were used which were obtained by aggregating the weekly prices of each GTIN code sold in the supermarkets and hypermarkets of 16 modern distribution chains located in the 107 Italian provinces using turnover weights, thus obtaining a total of 487,094 different products (GTINs).

The identification of the items is based on barcodes (GTINs), which univocally classify the products across the entire national territory. In each outlet (approximately 1,800), items were selected in order to cover up to 60% of the total turnover of the product aggregate. It is worth noting that perishables and seasonal products such as vegetables, fruit and meat were excluded from the scanner data because these products are sold at price per quantity and are not pre-packaged with GTIN codes.

Although this dataset was constructed for CPI compilation, it can also be used for making spatial price comparison among Italian regions bearing in mind that only the best selling products, which are typically consumed in each Italian province and region, may have be included according to the CPI selection procedure. Therefore these products may not be strictly comparable across different provinces and regions. It is useful to note that not all of the listed products must be priced in all of the regions. However, reliable regional price comparisons can be made as long as there is reasonable overlap in the items priced in different regions. We checked this requirement by verifying if product overlaps exhibit a chain structure.

3 Methods

In order to estimate regional spatial indices for products sold in modern distribution chains by using data for the 107 Italian provinces, a two-step procedure similar to the one used in the ICP was adopted whereby provinces are grouped into regions (World Bank, 2013). In the first step, within-regional PPPs are computed by comparing price and quantity data referring to products sold in the various provinces within each region while in the second step, between-regional PPPs are obtained for each region by using deflated price data for each province.

Moreover, as in international practice, sub-national PPP compilation is undertaken at two levels, viz., at BH level and at a more aggregated level (Food and Non-Food products). The methods selected for making multilateral comparisons is based on several axiomatic properties, including two basic properties: transitivity and base region invariance. Transitivity simply means that the PPP between any two regions should be the same whether it is computed directly or indirectly through a

[1] The smallest level of aggregation at which expenditure data are available are known in ICP parlance as basic headings. Although scanner data include expenditure data at item level, we still use the term “basic heading” to indicate a group of similar products which corresponds to a sub-class in the COICOP.

third region. The second requirement is that the PPPs be base region–invariant, which means that the PPPs between any two regions should be the same regardless of the choice of base region.

Aggregation methods at BH level

First step: within-regional PPPs

Let us assume that we are attempting to make a spatial comparison of prices between *R* regions, *r=1,…,R,* with *Mr* provinces in each region *r*. In the first stage of aggregation of price data at item level, which leads to price comparisons at BH level, p_{ijr} and q_{ijr} represent price and quantity of *i-th* item in *j-th* province and in *r-th* region $(i=1,2,...,N;\ j=1,2,...,M_r; r=1,\dots,R)$.

In order to compute within-regional PPPs, we explored using different methods[1], however, due to lack of space, we only illustrate the *Region-product-dummy (RPD) method* which was also used to compute between-regional PPPs. All methods are implemented using R. If product overlaps exhibit a chain structure thus the RPD method exhibits some aspects of spatial chaining.

The RPD is the regional version of the country-product-dummy (CPD) method used in international comparisons. This method suggests that price levels are estimated by regressing logarithms of prices on provinces for each *province* and product dummy variables; the model is given for each BH by:

$$\begin{aligned}\ln p_{ijr} &= \ln PPP_j + \ln P_i + \ln u_{ijr}\\ &= \pi_j + \eta_i + v_{ijr}\\ &= \sum_{j=1}^{M_r} \pi_j D^j + \sum_{i=1}^{N} \eta_i D^i + v_{ijr}\end{aligned} \qquad (1)$$

where D^j is a provincial-dummy variable that takes value equal to 1 if the price observation is from *j-th* province in the *r-th* region; and D^i is a i-dummy variable that takes value equal to 1 if the price observation is for *i-th* commodity. The random disturbance is assumed to satisfy the standard assumptions of a multiple regression model. In order to estimate parameters of this model we impose normalization $\sum_{j=1}^{M_r} \pi_j = 0$ thus treating all regions in a symmetric manner. If $\hat{\pi}_j\ (j=1,2,...,M_r)$ are estimated parameters, the within-regional PPP for the province *j* in region *r* is given by $WR\_PPP_j = e^{\hat{\pi}_j}$. The RPD method based price comparisons are transitive and base-invariant. With the aim of taking into account the economic importance (representativity) of each product expressed by expenditure weights w_{ijr} based on turnover we used a weighted RPD model:

$$\sqrt{w_{ijr}} \ln p_{ijr} = \sum_{j=1}^{M_r} \pi_j \sqrt{w_{ijr}} D^j + \sum_{i=1}^{N} \eta_i \sqrt{w_{ijr}} D^i + \sqrt{w_{ijr}} v_{ijr} \qquad (2)$$

Second step: between-regional PPPs

[1] We used various spatial index formulae, including Fisher based GEKS, Geary-Khamis and CPD (World Bank, 2013; Laureti and Rao, 2018). We found interesting results which suggest a high variability of prices within various regions. However, they cannot be reported here due to lack of space.

In order to use provincial prices adjusted for differences among provinces within the *r-th* region item prices in all the provinces of region *r* are converted by using:

$$\hat{p}_{ijr} = \frac{p_{ijr}}{WR\_PPP_{jr}} \qquad (3)$$

The deflated prices (in log form) were used for estimating a weighted RPD model with regional dummies and weights defined by deflated expenditure for each item in the *r-th* region.

$$\sqrt{w_{ijr}}\ln p_{ijr} = \sum_{k=1}^{R}\pi_k\sqrt{w_{ijr}}D^k + \sum_{i=1}^{N}\eta_i\sqrt{w_{ijr}}D^i + \sqrt{w_{ijr}}v_{ijr} \qquad (4)$$

The between-regional PPP for the region *r* is given by $R\_PPP_r = e^{\hat{\pi}_r}$ and transitive price comparisons based on RPD method are given by:

$$P_{rk}^{RPD} = \frac{\exp(\hat{\pi}_k)}{\exp(\hat{\pi}_r)} \quad for\,all\;\; r,k = 1,2,...,R \qquad (5)$$

Aggregation method for aggregation above basic heading level

The next and final step for compiling regional price comparisons is to aggregate the results from BH level comparisons to higher level aggregates. Let us assume that there are *L* basic headings (*l=1,...,L*) and e_i^r expenditure for *i-th* BH in region *r.* We decided to use the Fisher price index since it has a range of axiomatic and economic theoretic properties. The Fisher index is given by:

$$P_{rk}^{Fisher} = \sqrt{P_{rk}^{Laspeyres}\cdot P_{rk}^{Paasche}} \qquad (6)$$

where

$$P_{rk}^{Laspeyres} = \frac{\sum_{l=1}^{L} p_l^k q_l^r}{\sum_{l=1}^{L} p_l^r q_l^r} = \sum s_i^r\left(\frac{p_l^k}{p_l^r}\right), \qquad P_{rk}^{Paasche} = \frac{\sum_{i=1}^{N} p_l^k q_l^k}{\sum_{i=1}^{N} p_l^r q_l^k} = \left[\sum_l s_l^k\left(\frac{p_l^k}{p_l^r}\right)^{-1}\right]^{-1}$$

with $s_i^r = \frac{e_i^r}{\sum_{l=1}^{L} e_l^r} = \frac{p_l^r\cdot q_l^r}{\sum_{l=1}^{L} p_l^r\cdot q_l^r}$.

As the Fisher binary index in (6) is not transitive, it is possible to use the procedure suggested by Gini (1931), Elteto and Koves (1964) and Szulc (1964) referred to as the GEKS index to generate transitive multilateral price comparisons across different regions. The resulting index is given by:

$$P_{rk}^{GEKS-Fisher} = \prod_{r=1}^{R}\left[P_{rs}^{Fisher}\cdot P_{sk}^{Fisher}\right]^{1/R} \qquad (7)$$

The GEKS-Fisher based formula is used in cross-country comparisons made within the ICP at the World Bank (2015) and the OECD-Eurostat comparisons. In order to obtain a set of R_PPPs that refer to the group of regions (Italy) we standardized the GEKS-Fisher based PPPs (S-GEKS).

4 Results

In order to compute within and between regional spatial price indexes, we ran weighted RPD for all available BHs in scanner data (L=54) using expenditure weights defined by turnover. We then aggregated the results from BH level comparisons to higher level aggregates, i.e. food and non-food products both for within and between-regional PPPs. All results obtained cannot be reported, therefore Table1 illustrates RPPP for 2 BHs while Figure 1 shows aggregated RPPPs.

Table 1: WRPD estimation results for "Pasta products" and "Non-electrical appliances" Lazio=100

	Pasta Products (BH1)				**Non-electrical appliances (BH2)**			
Region	*Coef.*	*std.error*	*p-value*	*RPPPs*	*Coef.*	*std.error*	*p-value*	*RPPPs*
North-Center								
PIEMONTE	0.0187	0.0039	0.000	101.89	-0.0806	0.0079	0.000	92.26
VALLE D'AOSTA	0.0526	0.0039	0.000	105.41	0.0305	0.0081	0.000	103.10
LIGURIA	0.0482	0.0044	0.000	104.94	-0.0269	0.0079	0.001	97.35
LOMBARDIA	0.0264	0.0038	0.000	102.67	-0.0509	0.0079	0.000	95.04
TRENTINO A.A.	0.0716	0.0039	0.000	107.42	0.0051	0.0080	0.523	100.51
VENETO	0.0347	0.0038	0.000	103.53	-0.0309	0.0079	0.000	96.96
FRIULI V.G.	0.0435	0.0038	0.000	104.45	-0.0285	0.0079	0.000	97.19
EMILIA-ROMAGNA	0.0227	0.0041	0.000	102.30	-0.0580	0.0079	0.000	94.37
TOSCANA	-0.0050	0.0039	0.201	99.50	-0.1294	0.0079	0.000	87.86
UMBRIA	-0.0094	0.0039	0.015	99.06	-0.0185	0.0079	0.019	98.17
MARCHE	0.0557	0.0041	0.000	105.73	0.0077	0.0079	0.327	100.77
South and Islands								
ABRUZZO	0.0561	0.0040	0.000	105.77	-0.0163	0.0079	0.039	98.38
MOLISE	0.0471	0.0041	0.000	104.82	0.0142	0.0080	0.076	101.43
CAMPANIA	-0.0097	0.0040	0.014	99.04	0.0167	0.0079	0.035	101.69
PUGLIA	-0.0388	0.0040	0.000	96.20	-0.0267	0.0079	0.001	97.37
BASILICATA	-0.0410	0.0040	0.000	95.98	0.0021	0.0080	0.791	100.21
CALABRIA	-0.0286	0.0040	0.000	97.18	0.0087	0.0080	0.275	100.87
SICILIA	-0.0598	0.0044	0.000	94.19	0.0667	0.0080	0.000	106.89
SARDEGNA	0.0336	0.0046	0.000	103.41	-0.0266	0.0079	0.001	97.37
Obs.	18,007				3,453			
Root MSE	0.09538				0.10105			
AIC	-19261.88				-4447.601			

Regional spatial price indexes for two specific groups of products, that is "Pasta products" (BH1), which belongs to the aggregate Food products, and "Non-electrical appliances" (BH2, e.g. razors, scissors, hairbrushes, toothbrushes, etc.) included in the Non-Food aggregate, confirm large differences in price levels among Italian regions even if BH2 shows a higher territorial heterogeneity than BH1 (range is equal to 19.03 and 13.23 respectively). In the case of BH1, 5 regions located in the South and Islands (out of 8) and 2 Northern- Central regions (out of 11) show lower prices than Lazio while for BH2 higher price indexes are observed in 3 Southern regions and 8 regions in Northern and Central Italy. This different territorial pattern of consumption spatial price indexes is not confirmed when aggregated regional PPPs are computed for Food and Non-food products (Italy=100).

As shown in Figure 1, Southern regions appear to have price levels that are below the national average both for Food and Non-Food products, with the exception of Abruzzo (101.90 and 101.33, respectively), Molise (102.90 and 101.24) and Sardegna (101.93 and 101.57). However, it is worth noting that some

Northern regions also show lower price levels than the national average, such as Emilia-Romagna (98.31 and 98.40), Veneto (99.09 and 98.48) and Piemonte for Food products (99.80). On average, Toscana proved to be the less expensive region for both product aggregates (96.24 and 95.17). These results seem to suggest that when considering the retail modern distribution, the expected relationship in terms of price levels between the North and South of Italy partially changes (for "Pasta products" locally made goods could play a key role for maintaining the traditional price differences) and propose an interesting line for future research on the influence of the various distribution channels when defining sub-national PPPs. Caution is required when interpreting these results since: a) they may be influenced by the characteristics of the modern retail trade which is not uniformly distributed across Italian territory in terms of types of retail chains and market share; b) we excluded two groups of products "Whole Milk" and "Low-Fat Milk" since there were no reliable overlaps among regions enablig spatial price comparisons; c) these results are based on data selected for CPI compilation and hard discounts are excluded.

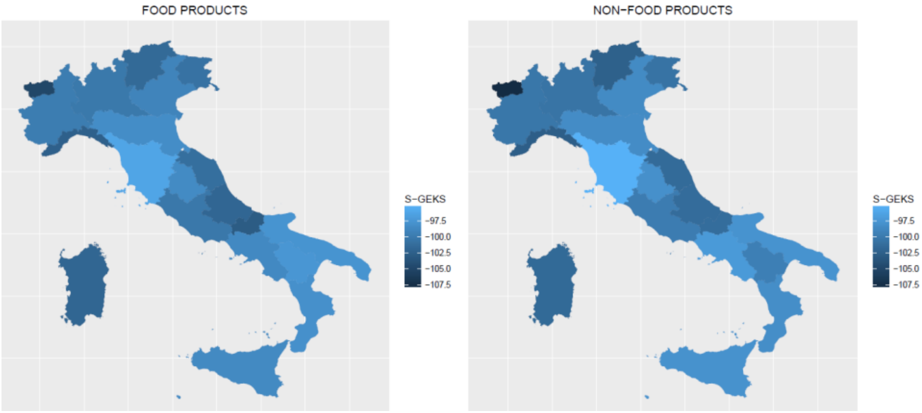


Figure 1: Between-regional PPPs for Food and Non Food products (Italy=100)

References

Biggeri, L., Laureti, T., and Polidoro, F.: Computing sub-national PPPs with CPI data: an empirical analysis on Italian data using country product dummy models. Soc Indic Res, 131(1), pp. 93-121, (2017).
Laureti, T., and Polidoro, F. Testing the use of scanner data for computing sub-national Purchasing Power Parities in Italy, Proceeding of 61st ISI World Statistics Congress, Marrakech, (2017)
Laureti, T., and Rao, D. P.: Measuring Spatial Price Level Differences within a Country: Current Status and Future Developments. Estudios de economía aplicada, 36(1), pp.119-148, (2018).
World Bank: Purchasing Power Parities and the Real Size of the World Economies-A Comprehensive Report of the 2011 International Comparison Program, Washington, DC(2015)

Financial Time Series Analysis

Dynamic component models for forecasting trading volumes

Modelli dinamici a componenti per la previsione dei volumi

Antonio Naimoli and Giuseppe Storti

Abstract We propose a new class of models for high-frequency trading volumes. Namely we consider a component model where the long-run dynamics are based on a Heterogeneous MIDAS polynomial structure based on an additive cascade of MIDAS filters moving at different frequencies. The merits of the proposed approach are illustrated by means of an application to three stocks traded on the XETRA market characterised by different degrees of liquidity.

Abstract *Viene proposta una nuova classe di modelli per volumi azionari ad alta frequenza. In particolare viene proposto un modello a componenti dove le dinamiche di lungo periodo sono basate su una struttura polinomiale di tipo MIDAS costituita da una cascata additiva di filtri MIDAS che si muovono a diverse frequenze. I vantaggi dell'approccio proposto vengono illustrati attraverso una applicazione a tre azioni contrattate sul mercato XETRA e caratterizzate da diversi livelli di liquidità.*

Key words: Intra-daily volume, component models, forecasting.

1 Introduction

Aim of this paper is to propose a novel dynamic component model for high-frequency trading volumes and assess its effectiveness for trading by means of an out-of-sample forecasting exercise. Volumes are indeed a crucial ingredient for the implementation of volume-weighted average price (VWAP) trading strategies.

Antonio Naimoli
Università di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES), Via Giovannni Paolo II, 132, 84084 Fisciano (SA), Italy. e-mail: `anaimoli@unisa.it`

Giuseppe Storti
Università di Salerno, Dipartimento di Scienze Economiche e Statistiche (DISES), Via Giovannni Paolo II, 132, 84084 Fisciano (SA), Italy. e-mail: `storti@unisa.it`

VWAP is one of the most common benchmarks used by institutional investors for judging the execution quality of individual stocks. The VWAP of a stock over a particular time horizon (usually one day) is simply given by the total traded value divided by the total traded volume during that period, i.e. the price of each transaction is weighted by the corresponding traded volume. The aim of using a VWAP trading target is to minimize the price impact of a given order by slicing it into smaller transaction sizes, reducing, in this way, the difference between expected price of a trade and its actual traded price. Investors, spreading the timing of transactions throughout the day, seek to achieve an average execution price as close as possible to the VWAP in order to lower market impact costs. Therefore, in this context, the key for a good strategy relies on accurate predictions of intra-daily volumes, since prices are substantially unpredictable.

The proposed specification, called the Heterogeneous MIDAS Component Multiplicative Error Model (H-MIDAS-CMEM), falls within the class of component MEM models as discussed in Brownlees et al. (2011). The most notable differences with respect to the latter are in the specification of the long-run component that is now modelled as an additive cascade of MIDAS filters moving at different frequencies from which the *heterogeneous* quality of the model comes. This specification is motivated by the empirical regularities arising from the analysis of high-frequency time series of trading volumes. After accounting for intra-day seasonality, treated employing a Fourier Flexible Form, these are typically characterised by two prominent and related features: a slowly moving long-run level and a highly persistent autocorrelation structure. In our model, these features are accounted by the heterogeneous MIDAS specification of the long-run component. Residual short term autocorrelation is then explained by an intra-daily non-periodic component that follows a unit mean reverting GARCH-type process. In addition, from an economic point of view, the cascade structure of the long-run component reproduces the natural heterogeneity of financial markets characterised by different categories of agents operating in the market at different frequencies. This results in a variety of sources separately affecting the variation of the average volume at different speeds. On a statistical ground, the cascade structure has the advantage of increasing model's flexibility since it allows to separately parametrize the dynamic contribution of each of these sources.

The estimation of model parameters is performed by the method of maximum likelihood under the assumption that the innovations are distributed according to the Zero-Augmented Generalized F (ZAF) distribution by Hautsch et al. (2014). The reason for this choice is twofold. First, it delivers a flexible probabilistic model for the conditional distribution of volumes. Second, it allows to control for the relevant proportion of zeros present in our data. In order to assess the relative merits of the proposed approach we have performed a forecasting exercise considering high-frequency trading volume for three stocks traded on the Xetra Market in the German Stock Exchange. The stocks have been selected to reflect different liquidity conditions as measured in terms of the number of non trading intra-daily intervals.

Our results show that the H-MIDAS-CMEM model is able to explain the the salient empirical features of the dynamics of high-frequency volumes. Also, we find that the H-MIDAS-CMEM is able to outperform its main competitors in terms of the usual Mean Squared Error and of the Slicing loss function proposed by Brownlees et al. (2011). Assessing the significance of differences in the predictive performance of models by the Model Confidence Set (MCS) of Hansen et al. (2011), it turns out that the H-MIDAS-CMEM is the only model always included in the set of superior models at different confidence levels.

In the reminder of the paper section 2 describes the proposed H-MIDAS-CMEM model defining its components as intra-daily periodic (subsection 2.1), intra-daily dynamic non-periodic (subsection 2.2) and long-run (subsection 2.3), respectively, while section 3 is dedicated to the out-of-sample forecasting exercise.

2 Model formulation

Let $\{x_{t,i}\}$ be a time series of intra-daily trading volumes. We denote days by $t \in \{1,\dots,T\}$, where each day is divided into I equally spaced intervals indexed by $i \in \{1,\dots,I\}$, then the total number of observations is given by $N = T \times I$.

The empirical regularities of high persistence and clustering of trading activity characterising intra-daily volumes lead us to build a Multiplicative Error Model consisting of multiple components that move at different frequencies. Extending the logic of the Component Multiplicative Error Model (CMEM) by Brownlees et al. (2011) and MIDAS regression models, we propose the H-MIDAS-CMEM which is formulated as

$$x_{t,i} = \tau_t \, g_{t,i} \, \phi_i \, \varepsilon_{t,i}. \tag{1}$$

The multiplicative innovation term $\varepsilon_{t,i}$ is assumed to be conditionally i.i.d., non-negative and to have unit mean and constant variance σ^2, i.e. $\varepsilon_{t,i}|\mathscr{F}_{t,i-1} \sim \mathscr{D}^+(1,\sigma^2)$, where $\mathscr{F}_{t,i-1}$ is the sigma-field generated by the available information until interval $i-1$ of day t. Then, the expectation of $x_{t,i}$, given the information set $\mathscr{F}_{t,i-1}$, is the product of three components characterised by a different dynamic structure. In particular, ϕ_i is an intra-daily periodic component parametrized by a Fourier Flexible Form, which reproduces the approximately U-shaped intra-daily seasonal pattern typically characterising trading activity. The $g_{t,i}$ component represents an intra-daily dynamic non-periodic component, based on a unit mean reverting GARCH-type process, that reproduces autocorrelated and persistent movements around the current long-run level. Finally, τ_t is a lower frequency component given by the sum of MIDAS filters moving at different frequencies. This component is designed to track the dynamics of the long-run level of trading volumes. Furthermore, the use of a time-varying intercept allows to reproduce sudden switches from very low to high trading intensity periods that typically occur in time series of high-frequency trading volumes. The structure of these components is described in more detail in the remainder of this section.

2.1 Intra-daily periodic component

Intra-daily volumes usually exhibit a U-shaped daily seasonal pattern, i.e. the trading activity is higher at the beginning and at the end of the day than around lunch time. To account for the periodic intraday factor we divide volumes $x_{t,i}$ by a seasonal component ϕ_i that is specified via a Fourier Flexible Form as proposed by Gallant (1981)

$$\phi_i = \sum_{q=0}^{Q} a_{0,q}\, \iota^q + \sum_{p=1}^{P} [a_{c,p}\, cos(2\pi p\, \iota) + a_{s,p}\, sin(2\pi p\, \iota)] \tag{2}$$

where $\iota = i/I \in (0,1]$ is a normalized intraday time trend.

Andersen et al. (2000) suggest that the Fourier terms in (2) do not add any significant information for $Q > 2$ and $P > 6$, so the model precision by using $Q = 2$ and $P = 6$ is enough to capture the behaviour of the intra-day periodicities.[1] Thus, assuming a multiplicative impact of intra-day periodicity effects, diurnally adjusted trading volumes are computed as

$$y_{t,i} = x_{t,i}/\phi_i. \tag{3}$$

2.2 Intra-daily dynamic non-periodic component

The intra-daily non-periodic component, unlike the seasonal component, takes distinctive and non-regular dynamics. In order to make the model identifiable, as in Engle et al. (2013), the intra-daily dynamic component follows a unit mean reverting GARCH-type process, namely $g_{t,i}$ has unconditional expectation equal to 1.

Then, the short-run component, in its simplest form, is formulated as

$$g_{t,i} = \omega^* + \alpha_1 \frac{y_{t,i-1}}{\tau_t} + \alpha_0 I(y_{t,i-1} = 0) + \beta_1 g_{t,i-1}, \tag{4}$$

where $\omega^* = (1 - \alpha_1 - (1-\pi)\alpha_0 - \beta_1)$, π is the probability that $y_{t,i} > 0$ and $I(y_{t,i-1} = 0)$ denotes an indicator function which is equal to 1 if the argument is true and to 0 otherwise.

2.3 The low frequency component

The low frequency component is modelled as a linear combination of MIDAS filters of past volumes aggregated at different frequencies. In this framework, a relevant issue is related to the identification of the frequency of the information to be used

[1] This result is confirmed by computing the Bayesian Information Criterion (BIC) for the estimation of P and Q lags.

by the filters, that notoriously acts a smoothing parameter. Therefore, using trading volumes moving at daily and hourly frequencies, the trend component τ_t is defined as

$$\begin{aligned} log\,\tau_t &= m + \theta_d \sum_{k=1}^{K_d} \varphi_k(\omega_{1,d}, \omega_{2,d}) YD_{t-k} \\ &\quad + \theta_h \sum_{k=1}^{K_d} \sum_{j=1}^{H} \varphi_{[j+(k-1)H]}(\omega_{1,h}, \omega_{2,h}) YH_{t-k}^{(H-j+1)}, \end{aligned} \tag{5}$$

where $YD_t = \sum_{i=1}^{I} y_{t,i}$, denotes the daily cumulative volume, with the subscript d referring to the daily frequency parameters. The subscript h is related to the parameters corresponding to the *hourly* frequency. If we let $t/H \in \{1,\ldots,H\times T\}$ denote the hourly frequency, with H being the number of intervals in which the day is divided, the variable $YH_t^{(j)}$ corresponds to the (j)-th hourly cumulative volume of the day t, that is $YH_t^{(j)} = \sum_{i=I\frac{(j-1)}{H}+1}^{I\frac{j}{H}} y_{t,i}$, for $j = 1,\ldots,H$. This multiple frequency specification is compatible with the heterogeneous market assumption of Müller et al. (1993), enforcing the idea that market agents can be divided in different groups characterised by different interests and strategies. Also, as pointed out in Corsi (2009), an additive cascade of linear filters moving at different frequencies allows to reproduce very persistent dynamics such as those typically observed for high-frequency trading volumes.

A common choice for determining $\varphi_k(\omega)$ is the Beta weighting scheme

$$\varphi_k(\omega) = \frac{(k/K)^{\omega_1-1}(1-k/K)^{\omega_2-1}}{\sum_{j=1}^{K}(j/K)^{\omega_1-1}(1-j/K)^{\omega_2-1}}. \tag{6}$$

As discussed in Ghysels et al. (2007), this Beta-specification is very flexible, being able to accommodate increasing, decreasing or hump-shaped weighting schemes. The Beta lag structure in (6) includes two parameters, but in our empirical applications ω_1 is always set equal to 1 such that the weights are monotonically decreasing over the lags. Furthermore, the number of lags K is properly chosen by information criteria to avoid overfitting problems.

The clustering of the trading activity involves a continuous variation of the average volume level and thus the dynamics of trading volumes are typically characterised by sudden transitions from states of very low trading activity to states of intense trading. In order to account for this switching-state behaviour we further extend the proposed modelling approach introducing a time-varying intercept in the formulation of the long-run component. This is specified as a convex combination of two different unknown parameters m_1 and m_2, that is $m_t = \lambda_t m_1 + (1-\lambda_t)m_2$. The combination weights are time-varying, since they change as a function of observable state-variables. The weight function λ_t follows a logistic specification of the type

$$\lambda_t = \frac{1}{1+exp(\gamma(\delta - s_{t-1}))}, \qquad (\gamma,\delta) > 0 \tag{7}$$

where γ and δ are unknown coefficients and s_{t-1} is an appropriately chosen state-variable.[2]

3 Out-of-sample forecasting comparison

High-frequency trading volume data used in our analysis refer to the stocks Deutsche Telekom (DTE), GEA Group (G1A) and Salzgitter (SZG) traded on the Xetra Market in the German Stock Exchange. An important feature of the data is the different number of zeros induced by non-trading intervals, since for DTE proportion of zero observations is 0.03%, for G1A 7.046% and for SZG 15.78%.

The raw tick-by-tick data have been filtered employing the procedure proposed by Brownlees and Gallo (2006), only considering regular trading hours from 9:00 am to 5:30 pm. Tick-by-tick data are aggregated computing intra-daily volumes over 10-minutes intervals, which means 51 observations per day. The data have been seasonally adjusted using the Fourier Flexible Form described in equation (2).

To evaluate the predictive ability of the H-MIDAS-CMEM models and their relative merits with respect to competitors, we perform an out-of-sample forecasting comparison over the period January-December 2007, which includes 251 days. In order to capture the salient features of the data and to safeguard against the presence of structural breaks, the model parameters are recursively estimated every day starting from January 2006 with a 1-year rolling window. Therefore at each step we predict 51 intra-daily volumes before re-estimating the models, for a total of 251 days and 12801 intra-daily observations. The out-of-sample performance of the examined models is evaluated by computing some widely used forecasting loss functions. The significance of differences in forecasting performance is assessed by the Model Confidence Set (MCS) approach (Hansen et al., 2011) which relies on a sequence of statistic tests to construct a set of superior models, in terms of predictive ability, at certain confidence level $(1-\alpha)$.

To compare the out-of-sample predictive performances we use the following loss functions

$$L^{MSE} = \sum_{t=1}^{T} \sum_{i=1}^{I} (x_{t,i} - \hat{x}_{t,i})^2$$

$$L^{Slicing} = -\sum_{t=1}^{T} \sum_{i=1}^{I} (w_{t,i} \, log \, \hat{w}_{t,i})$$

where L^{MSE} is the Mean Squared Error (MSE) of the volumes, while $L^{Slicing}$ is the Slicing loss function developed by Brownlees et al. (2011) to evaluate VWAP trading strategies. The slicing weights $\hat{w}_{t,i}$ are computed under both the static and dynamic VWAP replication strategies. The loss functions for single model shown in the top panel of Table 1 point out that the H-MIDAS-CMEM with fixed and, mainly,

[2] A suitable choice for the state-variable is the daily average of intra-daily volumes $\bar{y}_t$.

Table 1: Out-of-sample loss functions comparison

	Loss functions average values								
	DTE			G1A			SGZ		
	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}
MEM	0.481	3.920	2.767	1.997	3.921	2.765	0.869	3.921	2.764
CMEM	0.477	3.918	2.766	1.966	3.916	2.762	0.858	3.918	2.762
HAR-MEM	0.476	3.918	2.766	1.963	3.916	2.762	0.857	3.918	2.762
MIDAS-MEM	0.477	3.918	2.766	1.977	3.917	2.762	0.858	3.918	2.762
H-MIDAS-CMEM	0.465	3.915	2.764	1.958	3.909	2.757	0.850	3.912	2.758
H-MIDAS-CMEM-TVI	**0.455**	**3.914**	**2.763**	**1.850**	**3.907**	**2.756**	**0.799**	**3.911**	**2.757**

	MCS p-values								
	DTE			G1A			SGZ		
	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}	L^{MSE}	L^{SL}_{stc}	L^{SL}_{dyn}
MEM	0.000	0.000	0.000	0.002	0.000	0.000	0.004	0.000	0.000
CMEM	0.001	0.002	0.000	0.010	0.000	0.002	0.019	0.000	0.000
HAR-MEM	0.001	0.000	0.000	0.014	0.000	0.004	0.025	0.000	0.000
MIDAS-MEM	0.000	0.000	0.000	0.006	0.000	0.002	0.016	0.000	0.000
H-MIDAS-CMEM	0.027	░ 0.059	░ 0.086	▓ 0.263	░ 0.232	▓ 0.334	▓ 0.450	▓ 0.378	▓ 0.592
H-MIDAS-CMEM-TVI	▓ 1.000	▓ 1.000	▓ 1.000	▓ 1.000	▓ 1.000	▓ 1.000	▓ 1.000	▓ 1.000	▓ 1.000

Top panel: loss functions values for Mean Squared Error (L^{MSE}) and Slicing Loss with weights computed under the static (L^{SL}_{stc}) and dynamic (L^{SL}_{dyn}) VWAP replication strategy. In **bold** the best model. Bottom panel: MCS p-values for the examined loss functions. In ░ model ∈ 95% MCS and in ▓ model ∈ 75% MCS.

with time-varying intercept returns the lowest values for both the Mean Squared Error (L^{MSE}) and the Slicing loss using weights computed under the static (L^{SL}_{stc}) and dynamic (L^{SL}_{dyn}) VWAP replication strategy. A lower value of L^{MSE} provides evidence of a greater ability to capture the continuous variation from calms to storms periods, since intra-daily volume series are highly volatile, whereas minimizing the Slicing loss function increases the chances to achieve the VWAP target for a given trading strategy. In order to evaluate if the differences in terms of the considered loss functions are statistically significant, the MCS approach has been used. The MCS results confirm the strength of the H-MIDAS-CMEM, since the model with time-varying intercept is always included into the 75% MCS referring to the set of loss functions employed to measure the predictive ability of the models. For what concerns the H-MIDAS-CMEM with fixed intercept, it falls in the set of the superior models at the 0.75 confidence level for SZG according to the considered loss functions. This also applies to G1A, with the exception of the static Slicing loss entering at the 0.95 level. Finally, for DTE the H-MIDAS-CMEM is out of the MCS for the L^{MSE}, while falling into the 95% MCS for both the Slicing. Furthermore, the benchmark models never fall into the MCS according to the loss functions and the confidence levels considered.

References

Andersen, T. G., T. Bollerslev, and J. Cai (2000). Intraday and interday volatility in the japanese stock market. *Journal of International Financial Markets, Institutions and Money 10*(2), 107–130.

Brownlees, C. T., F. Cipollini, and G. M. Gallo (2011). Intra-daily volume modeling and prediction for algorithmic trading. *Journal of Financial Econometrics 9*(3), 489–518.

Brownlees, C. T. and G. M. Gallo (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis 51*(4), 2232–2245.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 174–196.

Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics 95*(3), 776–797.

Gallant, A. R. (1981). On the bias in flexible functional forms and an essentially unbiased form: the fourier flexible form. *Journal of Econometrics 15*(2), 211–245.

Ghysels, E., A. Sinko, and R. Valkanov (2007). Midas regressions: Further results and new directions. *Econometric Reviews 26*(1), 53–90.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hautsch, N., P. Malec, and M. Schienle (2014). Capturing the zero: A new class of zero-augmented distributions and multiplicative error processes. *Journal of Financial Econometrics 12*(1), 89–121.

Müller, U. A., M. M. Dacorogna, R. D. Davé, O. V. Pictet, R. B. Olsen, and J. R. Ward (1993). Fractals and intrinsic time: A challenge to econometricians. *Unpublished manuscript, Olsen & Associates, Zürich.*

Conditional Quantile-Located VaR

Valore a rischio condizionato ai quantili

Giovanni Bonaccolto, Massimiliano Caporin and Sandra Paterlini

Abstract The Conditional Value-at-Risk (CoVaR) has been proposed by [1] to measure the impact of a company in distress on the Value-at-Risk (VaR) of the financial system. We propose here an extension of the CoVaR, that is, the Conditional Quantile-Located VaR (QL-CoVaR), that better deals with tail events, when spillover effects impact the stability of the entire system. In fact, the QL-CoVaR is estimated by assuming that the financial system and the individual companies simultaneously lie in the left tails of their distributions.

Abstract *Il valore a rischio condizionato (CoVaR) é stato introdotto da [1] per quantificare l'impatto di una societá in fase di stress sul valore a rischio (VaR) del sistema finanziario. Nel presente lavoro proponiamo un'estensione del VaR (QL-CoVaR), che meglio si adatta agli eventi estremi, quando il rischio di contagio impatta sulla stabilitá dell'intera economia. Infatti, il QL-CoVaR é stimato assumendo che il sistema finanziario e le singole societá sono simultaneamente poste sulle code sinistre delle loro distribuzioni.*

Key words: CoVaR, systemic risk, quantile-on-quantile relationships

Giovanni Bonaccolto
University of Enna "Kore", viale delle Olimpiadi, 94100 Enna,
e-mail: giovanni.bonaccolto@unikore.it

Massimiliano Caporin
Department of Statistical Sciences, University of Padova, via C. Battisti 241, 35121 Padova,
e-mail: massimiliano.caporin@unipd.it

Sandra Paterlini
FACT Department–Finance, EBS Business School, Gustav-Stresemann-Ring 3, 65189 Wiesbaden,
Department of Economics and Management, University of Trento, via Inama 5, I-38122 Trento,
e-mail: sandra.paterlini@ebs.edu

1 Methods

We first introduce the Conditional Value-at-Risk (CoVaR) proposed by [1]. Then, we provide the details about the Conditional Quantile-Located Value-at-Risk (QL-CoVaR). Let y_t and $x_{i,t}$ be the returns of the financial system and of the i-th financial company at time t, respectively, for $i = 1,...,N$ and $t = 1,...,T$. Let $Q_\tau(x_{i,t}|\mathbf{I}_{t-1})$ denotes the τ-th quantile of $x_{i,t}$, for $\tau \in (0,1)$, conditional to the information set $\mathbf{I}_{t-1}$, where $\mathbf{I}_{t-1} = (y_{t-1}, x_{i,t-1}, m_{t-1})$ with m_{t-1} being a control variable at time $t-1$. Similarly, $Q_\theta(y_t|\mathbf{I}_{t-1}, x_{i,t})$ is the θ-th quantile of y_t conditional to the information set available at $t-1$ as well as to the return of the i-th company observed at time t, for $\theta \in (0,1)$. For simplicity, we set $Q_\tau(x_{i,t}|\mathbf{I}_{t-1}) \equiv Q_\tau(x_{i,t})$ and $Q_\theta(y_t|\mathbf{I}_{t-1}, x_{i,t}) \equiv Q_\theta^{(i)}(y_t)$; θ and τ take low values, typically in the interval $(0, 0.05)$. The CoVaR introduced by [1] is then estimated from the quantile regression models (see [4]):

$$Q_\tau(x_{i,t}) = \alpha_\tau^{(i)} + \beta_\tau^{(i)} m_{t-1}, \tag{1}$$

$$Q_\theta^{(i)}(y_t) = \delta_\theta^{(i)} + \lambda_\theta^{(i)} x_{i,t} + \gamma_\theta^{(i)} m_{t-1}. \tag{2}$$

Let $\widehat{Q}_\tau(x_{i,\tau}) = \widehat{\alpha}_\tau^{(i)} + \widehat{\beta}_\tau^{(i)} m_{t-1}$ be the estimated τ-th quantile of $x_{i,t}$, it is possible to compute the CoVaR at the distress and at the median state of the conditioning company, respectively, as follows:

$$CoVaR_{t,\theta,\tau}^{(i)} = \widehat{\delta}_\theta^{(i)} + \widehat{\lambda}_\theta^{(i)} \widehat{Q}_\tau(x_{i,t}) + \widehat{\gamma}_\theta^{(i)} m_{t-1}, \tag{3}$$

$$CoVaR_{t,\theta,1/2}^{(i)} = \widehat{\delta}_\theta^{(i)} + \widehat{\lambda}_\theta^{(i)} \widehat{Q}_{1/2}(x_{i,t}) + \widehat{\gamma}_\theta^{(i)} m_{t-1}, \tag{4}$$

and compute the ΔCoVaR to quantify the marginal contribution of the i-th company to the systemic risk (see [1]). Note that $CoVaR_{t,\theta,1/2}^{(i)}$ is always parameterized to the median state of the i-th conditioning company. Hence, we can omit the level 1/2 as subscript of the ΔCoVaR measure as follows:

$$\Delta CoVaR_{t,\theta,\tau}^{(i)} = CoVaR_{t,\theta,\tau}^{(i)} - CoVaR_{t,\theta,1/2}^{(i)} = \widehat{\lambda}_\theta^{(i)} \left[\widehat{Q}_\tau(x_{i,t}) - \widehat{Q}_{1/2}(x_{i,t})\right]. \tag{5}$$

For simplicity, we set $\theta = \tau$ and, then, $\Delta CoVaR_{t,\theta,\tau}^{(i)} \equiv \Delta CoVaR_{t,\tau}^{(i)}$. It is important to highlight that the parameters in (2) and the coefficients in (3) are functions of θ only, neglecting the role of τ. Therefore, the estimation process behind (3) depends on $x_{i,t}$ and not on $Q_\tau(x_{i,t})$. In contrast, we estimate the parameters in (2) assuming that the financial system and the i-th company simultaneously lie in the left tails of their distributions. We then take into account the impact exerted by $x_{i,t}$—in the neighbourhood of its τ-th quantile—on $\widehat{Q}_\theta^{(i)}(y_t)$. This allows us to increase the distress degree in the connections between the individual companies and the system to make our risk measure more sensitive to extreme events. The model we propose is defined as follows:

$$Q^{(i)}_{\theta,\tau}(y_t) = \delta^{(i)}_{\theta,\tau} + \lambda^{(i)}_{\theta,\tau} x_{i,t} + \gamma^{(i)}_{\theta,\tau} m_{t-1}, \tag{6}$$

where the parameters have both θ and τ as subscripts, as they depend on the quantiles levels of both y_t and $x_{i,t}$.

In fact, the unknown parameters in (6) are estimated from the following minimization problem:

$$\underset{\delta^{(i)}_{\theta,\tau},\lambda^{(i)}_{\theta,\tau},\gamma^{(i)}_{\theta,\tau}}{\arg\min} \sum_{t=1}^{T} \rho_\theta \left[y_t - \delta^{(i)}_{\theta,\tau} - \lambda^{(i)}_{\theta,\tau} x_{i,t} - \gamma^{(i)}_{\theta,\tau} m_{t-1} \right] K\left(\frac{\widehat{F}_{t|t-1}(x_{i,t}) - \tau}{h} \right), \tag{7}$$

where $\rho_\theta(e) = e(\theta - \mathbf{1}_{\{e<0\}})$ is the asymmetric loss function used in the quantile regression method by [4]; $\mathbf{1}_{\{\cdot\}}$ is an indicator function, taking the value of 1 if the condition in $\{\cdot\}$ is satisfied, the value of 0 otherwise; $K(\cdot)$ is the kernel function, with bandwidth h, whereas $\widehat{F}_{t|t-1}(x_{i,t})$ is the empirical conditional quantile of $x_{i,t}$. [5] used a similar approach to estimate the relations in quantiles between oil prices and stock returns.

In contrast to [5], we estimate $\widehat{F}_{t|t-1}(x_{i,t})$ dynamically using a rolling window procedure. For each window, we estimate a large set of $x_{i,t}$ quantiles in the support $\tau \in (0,1)$ from the quantile regression model (1), using the method proposed by [2] to ensure the monotonicity of the multiple quantiles for $\tau \in (0,1)$. Then, we linearly interpolate the set of quantiles to obtain the conditional distribution of $x_{i,t}$ at time t, denoted as $\widehat{F}(x_{i,t}|m_{t-1})$. Finally, we recover $\widehat{F}_{t|t-1}(x_{i,t})$, as the probability level, extrapolated from $\widehat{F}(x_{i,t}|m_{t-1})$, corresponding to the realization $x_{i,t}$.

From the method described above, we then compute the QL-CoVaR at the τ-th level as follows:

$$\mathit{QL\text{-}CoVaR}^{(i)}_{t,\theta,\tau} = \widehat{\delta}^{(i)}_{\theta,\tau} + \widehat{\lambda}^{(i)}_{\theta,\tau} \widehat{Q}_\tau(x_{i,t}) + \widehat{\gamma}^{(i)}_{\theta,\tau} m_{t-1}, \tag{8}$$

where $\widehat{Q}_\tau(x_{i,t}) = \widehat{\alpha}^{(i)}_\tau + \widehat{\beta}^{(i)}_\tau m_{t-1}$.

Then, given $\theta = \tau$, and evaluating the model also for $\tau = 1/2$, we define the ΔQL-CoVaR as:

$$\begin{aligned} \Delta\mathit{QL\text{-}CoVaR}^{(i)}_{t,\tau} &= \mathit{QL\text{-}CoVaR}^{(i)}_{t,\theta,\tau} - \mathit{QL\text{-}CoVaR}^{(i)}_{t,\theta,1/2} = \widehat{\delta}^{(i)}_{\theta,\tau} - \widehat{\delta}^{(i)}_{\theta,1/2} \\ &+ \widehat{\lambda}^{(i)}_{\theta,\tau} \left[\widehat{Q}_\tau(x_{i,t}) - \widehat{Q}_{1/2}(x_{i,t}) \right] + (\widehat{\lambda}^{(i)}_{\theta,\tau} - \widehat{\lambda}^{(i)}_{\theta,1/2}) \widehat{Q}_{1/2}(x_{i,t}) \\ &+ (\widehat{\gamma}^{(i)}_{\theta,\tau} - \widehat{\gamma}^{(i)}_{\theta,1/2}) m_{t-1}. \end{aligned} \tag{9}$$

It is important to highlight that $\Delta\mathit{QL\text{-}CoVaR}^{(i)}_{t,\tau}$ includes more components than $\Delta\mathit{CoVaR}^{(i)}_{t,\tau}$ in (5), as the coefficients in (9) also depend on the state of the i-th company. We then have further information about the relationships between the financial system and the individual companies when we focus on the left tails of their distributions. We compute the standard errors of the ΔCoVaR and the ΔQL-CoVaR coefficients using the bootstrap approach (see, e.g., [3]).

2 Empirical results

We implement the methods discussed in Section 1 on the daily returns of 1,155 U.S. financial institutions (952 banks and 203 insurance companies) in the period between October 10, 2000 and July 31, 2015, for a total of 3,864 days.[1] We note that some of the companies enter the dataset after October 10, 2000, whereas others exit before July 31, 2015. We estimate the models described in Section 1 for each of the financial companies for which we have at least 200 observations, resulting in 1,030 companies. We also build an index providing the return of the financial system (y_t) from the returns of the 1,155 financial institutions included in our dataset, weighted by their market values. As for m_t, we use the first principal component of variables that are related to bond, equity and real estate markets: i) the CBOE Volatility Index (VIX); ii) the liquidity spread (LS), computed as the difference between the three-month collateral repo rate and the three-month bill rate; iii) the change in the three-month Treasury bill rate (TB); iv) the change in the slope of the yield curve (YC), computed as the spread between the ten-year Treasury rate and the three-month bill rate; v) the change in the credit spread between BAA-rated bonds and the Treasury rate (CS), both with the ten years maturity; vi) the daily equity market return (EM); vii) the excess return of the real estate sector over the market return (RE).[2] In particular, we checked that the first principal component (m_t) of the variables listed above explains 96.50% of the variability in the data.

Table 1 Estimation of $Q_\theta^{(i)}(y_t) = \delta_\theta^{(i)} + \lambda_\theta^{(i)} x_{i,t} + \gamma_\theta^{(i)} m_{t-1}$

	$\theta = 0.01$					$\theta = 0.05$				
COEF	5P	MED	95P	IQR	PS	5P	MED	95P	IQR	PS
δ_θ	-0.042	-0.031	-0.021	0.012	99.90	-0.027	-0.019	-0.014	0.007	99.61
λ_θ	-0.033	0.117	0.536	0.236	45.15	-0.006	0.112	0.561	0.276	57.09
$100 \times \gamma_\theta$	-0.295	-0.197	-0.074	0.107	88.45	-0.202	-0.128	-0.079	0.064	95.24

The table reports the summary statistics of the CoVaR's parameters estimated for the N financial companies included in our dataset. The estimates are obtained using two quantile levels—θ. In each panel, from left to right, we report the following descriptive statistics of the coefficients: the 5-th percentile (5P), the median (MED), the 95-th percentile (3Q), the interquartile range (IQR) and the percentage of times, out of N, in which they are statistically significant at the 5% confidence level (PS).

We estimate the CoVaR and the QL-CoVaR using two quantile levels—$\theta = \tau = 0.01$ and $\theta = \tau = 0.05$. As for the estimation of the QL-CoVaR parameters, we use the Gaussian kernel as $F(\cdot)$, with $h = 0.15$.[3] On the basis of the empirical set-up

[1] The data are recovered from Thomson Reuters Datastream.

[2] The control variables listed in i)—v) are taken from Thomson Reuters Datastream, whereas EM and RE are recovered from the industry portfolios built by Kenneth R. French, available at *http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.*

[3] We used other values of h for a robustness check, that is, $h = \{0.10, 0.20\}$. We checked that the results obtained with different h values are qualitative similar and are available upon request.

Table 2 Estimation of $Q^{(i)}_{\theta,\tau}(y_t) = \delta^{(i)}_{\theta,\tau} + \lambda^{(i)}_{\theta,\tau} x_{i,t} + \gamma^{(i)}_{\theta,\tau} m_{t-1}$ and $Q^{(i)}_{\theta,1/2}(y_t) = \delta^{(i)}_{\theta,1/2} + \lambda^{(i)}_{\theta,1/2} x_{i,t} + \gamma^{(i)}_{\theta,1/2} m_{t-1}$

COEF	5P	MED	95P	IQR	PS	5P	MED	95P	IQR	PS
	$\theta = \tau = 0.01$					$\theta = \tau = 0.05$				
$\delta_{\theta,\tau}$	-0.049	-0.029	-0.016	0.014	95.05	-0.030	-0.018	-0.010	0.007	95.44
$\delta_{\theta,1/2}$	-0.041	-0.028	-0.019	0.011	99.22	-0.026	-0.018	-0.012	0.007	99.61
$\lambda_{\theta,\tau}$	-0.199	0.248	1.025	0.559	30.87	-0.081	0.212	0.902	0.471	40.49
$\lambda_{\theta,1/2}$	-0.163	0.214	0.788	0.387	41.65	-0.047	0.231	0.731	0.366	55.34
$100 \times \gamma_{\theta,\tau}$	-0.375	-0.194	-0.013	0.143	77.57	-0.242	-0.143	-0.061	0.076	88.93
$100 \times \gamma_{\theta,1/2}$	-0.300	-0.180	-0.045	0.099	86.21	-0.186	-0.115	-0.066	0.059	92.14

The table reports the summary statistics of the QL-CoVaR parameters estimated for the N financial companies included in our dataset. We estimated the conditional quantiles for two quantile levels of θ and $h = 0.15$. From left to right, we report the following descriptive statistics of the coefficients: the 5-th percentile (5P), the median (MED), the 95-th percentile (95P), the interquartile range (IQR) and the percentage of times, out of N, in which they are statistically significant at the 5% confidence level (PS).

described above, we estimate the QL-CoVaR parameters. The results are reported in Table 2.

As expected, we can see from Table 1 that positive returns of the individual companies have a positive impact on the VaR of the financial system, as $\lambda^{(i)}_{\theta}$ takes, on average, positive values. In contrast, Table 2 reports the statistics of the QL-CoVaR coefficients, where we condition the estimates to the distress and to the median state of a single financial company. As before, the average impact exerted by the companies to both *QL-CoVaR*$^{(i)}_{\tau}$ and *QL-CoVaR*$^{(i)}_{1/2}$ is positive, but greater with respect to the standard CoVaR (the medians of both $\widehat{\lambda}^{(i)}_{\theta,\tau}$ and $\widehat{\lambda}^{(i)}_{\theta,0.5}$ are greater than the median of $\widehat{\lambda}^{(i)}_{\theta}$). Therefore, the relationships between the system and the companies become stronger by focusing on particular regions of the $x_{i,t}$ support, i.e. when $x_{i,t}$ is in a neighbourhood of a distress state.

On average, we observe larger values for $\widehat{\lambda}^{(i)}_{\theta,\tau}$ at $\theta = 0.01$ than at $\theta = 0.05$, whereas the opposite holds for $\widehat{\lambda}^{(i)}_{\theta,0.5}$. $\widehat{\lambda}^{(i)}_{\theta,\tau}$ measures the relation between $x_{i,t}$ and y_t, when the companies and the system simultaneously lie in the left tail of their distributions. The fact that $\widehat{\lambda}^{(i)}_{\theta,\tau}$ increases as θ and τ simultaneously decrease means that the co-movements between the financial system and the companies become stronger when moving leftwards along the left tails of their distributions. Consequently, the risk of contagion increases by accentuating the distress degree in the connections between the financial system and the companies. For both CoVaR and QL-CoVaR, the percentage of times in which the coefficient measuring the impact of the individual companies (λ) is statistically significant at the 5% level is greater at $\theta = \tau = 0.05$ than at $\theta = \tau = 0.01$.

References

1. Adrian, T., Brunnermeier, K. (2016). CoVaR. *American Economic Review*, 106, 1705–1741.
2. Bondell, H.D., Reich, B.J., Wang, H. (2010). Non-crossing quantile regression curve estimation. *Biometrika*, 97, 825–838.
3. Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1–26.
4. Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.
5. Sim, N., Zhou, H. (2015). Oil prices, US stock return, and the dependence between their quantiles. *Journal of Banking & Finance*, 55, 1–8.

Forensic Statistics

Cause of effects: an important evaluation in Forensic Science

Cause degli Effetti: una rilevante valutazione nelle Scienze Forensi

Fabio Corradi and Monica Musio

Sommario *Causes of Effects* (COE) queries concern the assessment of causal relationship in individual cases by evaluating the *probability of causation.* However is not always clear how and whether, to usefully employ scientific data for this pourpose. Given even a randomized sample we can typically only provide bounds for the *probability of causation* if some fundamental conditions, namely exogeneity, comparability and sufficiency [4] , are satisfied. In this work we make the fundamental conditions operative by means of a Bayesian model selection procedure.
Sommario *Le problematiche di Cause di Effetti (COE) riguardano la valutazione di relazioni causali individuali attraverso il calcolo della probabilità di causazione. In questo contesto non è sempre chiaro come utilizzare dati provienti da studi scientifici. Infatti anche in presenza di esperimenti randomizzati si possono solo calcolare degli intervalli per la probabilità di causazione, qualora siano soddisfatte le così dette condizioni fondamentali di esogeneità, comparabilità e sufficienza [4]. L'obiettivo di questo lavoro è di rendere operative queste condizioni tramite una procedura di selezione di modelli in ambito Bayesiano.*

Key words: Causes of Effects, Fundamental conditions, Bayesian Model Selection

1 Introduction

In causal inference it is important to distinguish two types of causal query which, although not entirely unconnected, are nevertheless very different, both in form and in the type of answer they require. The following example traces back to Holland,

Fabio Corradi
Dipartimento di Statistica, Informatica e Applicazioni, Università di Firenze, e-mail: corradi@disia.unifi.it

Monica Musio
Dipartimento di Matematica ed Informatica, Università di Cagliari e-mail: mmusio@unica.it

1986 [6] and is archetypical in causal literature.
Effects of Causes (EoC): " Ann has a headache. She is wondering whether to take aspirin. Will that cause her headache to disappear?"
Causes of Effects (CoE): "Ann had a headache and took aspirin. Her headache went away. Was that caused by the aspirin?"
EoC type questions have the form "Will A cause B ?" CoE questions have the form "Did A cause B ?", and are also known as problems of individual causation.

How to use scientific data on individual cases is a problem that is central to the courts of justice and, in forensic literature, is referred as the G2i ("Group-to-individual") problem Dawid *et al.*, (2014) [3]).
In the CoE case, Ann actually chose to take the aspirin ($E = 1$) and assumed the drug ($T = 1$), this produces her headache to disappear ($R = 1$).

In this work we concentrate on the general CoE setting and more explicitly on how to support the conditions required to evaluate the probability of causation.

2 Basic of COE

Let's consider again the CoE example about the aspirin: Ann had a headache and decided ($E = 1$) to took aspirin ($T = 1$). Her headache went away ($R = 1$). Was that caused by the aspirin? One possible answer, "Probability of Causation", PC, relies on potential responses (R_0, R_1) (where R_x denotes the value R takes when $T = x$, *i.e.* T is set to x). We know that, for Ann, $T = 1$ so that $R_1 = 1$: she took the aspirin and her headache disappeared. Now if in "counterfact" $R_0 = 1$, then Ann's headache would have disappeared even if she had not taken the aspirin, so we must conclude that it was not the aspirin that cured her. Conversely, if $R_0 = 0$ then we can indeed attribute her cure to having taken the aspirin. Everything must be also evaluated considering her choice to take the drug, $E = 1$, which could be informative of her health status or, in general, on her preference to assume the treatment. In this way our CoE causal question is formulated in terms of the contrast between the factual outcome R_1 and the counterfactual outcome R_0. Formally we can define the PC as the *conditional probability* (see, *e.g.*, Dawid *et al.*, (2016) [4])

$$\mathrm{PC}_A = \Pr(R_0 = 0 \mid H^A, E = 1, R_1 = 1) = \frac{\Pr(R_0 = 0, R_1 = 1 \mid H^A, E = 1)}{\Pr(R_1 = 1 \mid H^A, E = 1)}, \qquad (1)$$

where PC_A denotes the judge's probability distribution over attributes of Ann and H^A the background information, some relevant known information about Ann. We denote by H a set of variables $H = \{H_1, \ldots, H_K\}$ and with $H^A = \{H_1^A, \ldots, H_K^A\}$ their corresponding values for Ann. This formal approach does however leave open the questions of how to evaluate PC and what evidence can be used. The numerator of (1) is not estimable, since we can never observe both R_0 and R_1 for the same individual, hence we can never assess this dependence without making any further assumptions. CoE questions may simply not have well-determined answers, but we can sometimes set bounds so long as some *fundamental conditions* are satisfied (Dawid *et al.*, (2016) [4]). In such case it is possible to show that

$$\Pr(R_0 = 0 \mid H^A, E = 1, R_1 = 1) \geq \max\{0, 1 - \frac{1}{RR_A}\} \tag{2}$$

with

$$RR_A := \frac{\Pr(R_1 = 1 \mid H^A, E = 1)}{\Pr(R_0 = 1 \mid H^A, E = 1)}.$$

Whenever $RR_A > 2$ the Probability of Causation PC_A will exceed 50%. In a civil court this is often taken as the criterion to assign legal responsibility ”on the balance of probabilities”. How much information H^A to take into account is still an unsolved problem. The aim of this work is to evaluate how different specifications of H support the fundamental conditions. We pose the issue as a model selection problem within the class of models induced by the fundamental conditions, where each model specifies a particular choice for H. Models will be evaluated by considering their marginal likelihood and a prior on the model space.

3 Fundamental conditions

3.1 Exogeneity

The potential outcomes (R_0, R_1) have the same joint distribution among both treated and untreated study subjects having the same background information H^A as Ann.
This condition essentially assumes no confounding $(R_0, R_1) \perp\!\!\!\perp T \mid H^A$ for every specification of H. This assumption cannot be tested empirically since in the treated patients we only observe R_1 and in the untreated patients we only observe R_0. So any argument we make for exogeneity has to be justified because of the randomization and ignorance of the influence of the H characteristics on the response.

3.2 Comparability

Conditional on knowledge of the pre-treatment characteristics of Ann and the trial subjects we can regard Ann's potential responses as comparable with those of the treated subjects having characteristics H^A.
Comparability essentially says that we are not able to distinguish between Ann and the group of treated individuals as it concerns the uncertainty of their response to the treatment. This is nothing but exchangeability, a basic form of dependence introduced by de Finetti. If exchangeability holds, the joint distribution of the random variables considered for a problem are invariant to permutation, i.e. there is no way

to make a distinction among them. According to what characteristics are included in H, different individuals in the randomized sample will be compared with Ann. Then using the representation theorem we can evaluate the probability to observe Ann and the group of responses if exchangeability holds among them.

3.3 Sufficiency

Conditional on H^A, Ann's intention to take (or not) the treatment does not affect the distribution of her potential responses.

This condition requires that, given Ann's own background information H^A, her potential outcomes are not further influenced by her (known) desire to take the aspirin. We need first too address the condition required to evaluate the denominator of RR_A: *Conditional on knowledge of the pre-treatment characteristics of Ann and the trial subjects, we can regard Ann's potential responses as comparable with those of the* untreated *subjects having characteristics H^A.* This potential probability clearly concerns a counterfactual event, since Ann decided and actually took the aspirin; obviously we don't know the response if she had not taken the aspirin but we know her will to assume the drug. We assume that E is observed in the sample: for example, a patient entered into a randomized trial is conscious that he may not receive the treatment but only a placebo and nevertheless agrees to express the wish to take aspirin or not. The ambition is to evaluate $\Pr(R_0 = 1 \mid H^A, E = 1)$ using traditional experimental data for $T = 0$, without ignoring the decision of Ann to take the drug, which is *observational* in nature. $E = 1$ describes the Ann's desire to take the aspirin, information not included in H^A but possibly relevant to determine the probability of the response. If we can obtain reasonable support to the condition

$$R_0 \perp\!\!\!\perp E \mid H^A \qquad \text{i.e. if} \tag{3}$$

$$\Pr(R_0 = 1 \mid H^A, E = 1) = \Pr(R_0 = 1 \mid H^A, E = 0) \tag{4}$$

it would be possible to estimate $\Pr(R_0 = 1 \mid H^A, E = 1)$ by $\Pr(R = 1 \mid H^A, T = 0)$.

4 Model selection

We pose the problem of finding the group most fitting the fundamental conditions as one of model selection solved, as usual, by computing the marginal likelihood, based on the data, for each possible specifications of H. We have 2^K possible different choices for the characteristics that can be selected from H. Let J be one of these choices, that can be identified as a subset of $\{1, \ldots, K\}$. We introduce an equivalence relation $\sim$ on the set of treated individuals that agree to receive the treatment ($T = 1, E = 1$). Namely if I_1 and I_2 are two such individuals then

$$I_1 \sim I_2 \iff H_j^{I_1} = H_j^{I_2}, \quad \forall j \in J.$$

A similar definition can be given for the other subgroups ($T = 1, E = 0$), ($T = 0, E = 1$) and ($T = 0, E = 0$). Thus each equivalence relation determines a partition of one of the above four groups. These four partitions identify a *model* M_J. We denote each element of one quotient set by $X^s_{t,e}$, where $t \in \{0,1\}$, $e \in \{0,1\}$ and $s \in \{0,1,2\ldots,N\}$, $N-1$ being the number of elements of each partition. We reserve the notation $X^0_{t,e}$ to the sets of individuals who share all the characteristics with Ann. Using the same notation as before, we indicate by $\mathbf{r}_{t,e}$ the vector of responses in the main 4 groups, by $\mathbf{r}_{\{\setminus X\}}$ the vector of responses in the complementary of the set X, by $\mathbf{r}^s_{t,e}$ the vector of responses referred to a specific group of individuals and by r_A the Ann's response. The responses of the treated and untreated individuals are not considered exchangeable each other and are modeled separately using the conditional representation indexed by their own θs, which are a priori assumed independent. So the marginal likelihood we want to evaluate factorizes and we evaluate each contribution separately.

$$\begin{aligned}
&\Pr(r_A, \mathbf{r}_{1,1}, \mathbf{r}_{1,0}\mathbf{r}_{0,1}, \mathbf{r}_{0,0}|M_J) = \qquad (5)\\
&= \int_{\Theta_{T=1}} \int_{\Theta_{T=0}} \Pr(r_A, \mathbf{r}_{1,1}, \mathbf{r}_{1,0}\mathbf{r}_{0,1}, \mathbf{r}_{0,0}, \theta_{T=1}, \theta_{T=0}|M_J) d\theta_{T=1} d\theta_{T=0}\\
&= \int_{\Theta_{T=1}} \Pr(r_A, \mathbf{r}_{1,1}, \mathbf{r}_{1,0}, \theta_{T=1}|M_J) d\theta_{T=1} \int_{\Theta_{T=0}} \Pr(\mathbf{r}_{0,1}, \mathbf{r}_{0,0}, \theta_{T=0}|M_J) d\theta_{T=0}.
\end{aligned}$$

4.1 Marginal likelihood for comparability

Because of the sufficiency condition, we assume the same mixing parameter $\theta^s_{T=1}$ in the sets $X^s_{1,1}$ and $X^s_{1,0}$, whose prior is assumed a non-informative $Be(1,1)$, while for the comparability condition Ann is suppose to be exchengeable with the treated individuals sharing with Ann the same characteristics. The marginal likelihood is

$$\begin{aligned}
&\Pr(r_A, \mathbf{r}_{1,1}, \mathbf{r}_{1,0}|M_J) = \Pr\left(r_A, \mathbf{r}^0_{1,1}, \mathbf{r}^0_{1,0}, \mathbf{r}_{\setminus\{X^0_{1,1}, X^0_{1,0}\}}\right) \qquad (6)\\
&= \int_{\Theta^0_{T=1}} \Pr(r_A, \mathbf{r}^0_{1,1}, \mathbf{r}^0_{1,0}, \theta^0_{T=1}) d\theta^0_{T=1} \cdot \prod_{s \neq 0} \int_{\Theta^s_{T=1}} \Pr(\mathbf{r}^s_{1,e}, \theta^s_{T=1}) d\theta^s_{T=1} =\\
&= \frac{x^0_{1,e} + 1}{n^0_{1,e} + 2} \prod_{s \neq 0} \frac{1}{n^s_{1,e} + 1}
\end{aligned}$$

where $x^0_{1,e}$ is the number of success in the group $X^0_{1,1} \cup X^0_{1,0}$, $n^0_{1,e} = |X^0_{1,1}| + |X^0_{1,0}|$ and $n^s_{1,e} = |X^s_{1,1}| + |X^s_{1,0}|$.

4.2 Marginal likelihood for sufficiency

The sufficiency assumption requires E to have not influence on the response for the untreated with the Ann's characteristics. So, if $R_0 \perp\!\!\!\perp E|H^A$ holds, then the two groups of r.v., $\mathbf{r}^0_{0,1}, \mathbf{r}^0_{0,0}$, share a common mixing distribution parameter $\theta^0_{T=0}$, whose prior is assumed a non-informative $Be(1,1)$. We denote by $\theta^s_{T=0,E=e}$ the mixing parameter in the quotient set $X^s_{0,e}$. The marginal likelihood is:

$$\Pr(\mathbf{r}_{0,1},\mathbf{r}_{0,0}|M_J) = \Pr(\mathbf{r}^0_{0,1},\mathbf{r}^0_{0,0},\mathbf{r}_{\setminus\{X^0_{0,1},X^0_{0,0}\}}) = \int_{\Theta^0_{T=0}} \Pr(\mathbf{r}^0_{0,1},\mathbf{r}^0_{0,0},\theta^0_{T=0}) d\theta^0_{T=0} \cdot$$

$$\prod_{s\neq 0}\prod_{e\in\{0,1\}} \int_{\Theta^s_{T=0,E=e}} \Pr(\mathbf{r}^s_{0,e},\theta^s_{T=0,E=e}) d\theta^s_{T=0,E=e}$$

$$= \binom{n^0_{0,0}}{x^0_{0,0}}\binom{n^0_{0,1}}{x^0_{0,1}}\binom{n^0_{0,0}+n^0_{0,1}}{x^0_{0,0}+x^0_{0,1}}^{-1} \cdot \frac{1}{n^0_{0,0}+n^0_{0,1}+1}\prod_{s\neq 0}\prod_{e\in\{0,1\}} \frac{1}{n^s_{0,e}+1} \qquad (7)$$

where we have used a similar notation as before.

Remark 1 (Marginal likelihood and the Irving-Fisher exact test). The marginal likelihood evaluated for a subset of H formally shares the hypergeometric part with the conditional Irving-Fisher exact test statistic used to evaluate differences among the rate of success of an event in two populations. The result is not surprising since we are looking for the set of H making E irrelevant, so supporting the so called H_0 hypothesis of no-difference between the success ratio in the two groups.

4.3 Prior and posterior in the model space

By (5), the required marginal likelihood is the product of (6) and (7). The goal is to evaluate the posterior of M_J given the responses observed on Ann and on the sample. To this end we introduce a prior over the space of models. The simpler choice is to consider an uniform distribution on this space. Another choice is this one proposed by Chen and Chen, (2008) [1]. They give the same prior probability (equal to $\frac{1}{k+1}$) to all models sharing the same number of characteristics k. In this way for the generic model M_J we have:

$$\Pr(M_J) = \frac{1}{k+1}\binom{k}{|M_J|}^{-1} \cdot I(|M_J| \leq k/2) \qquad (8)$$

where the search spans all models including at most $k/2$ characteristics. This choice favours model selection according to the Occam razor principle: the fewer characteristics employed, the more probable is the model. This rationale is reasonably objective. Combining (6) and (7) and (8), we get the required posterior.

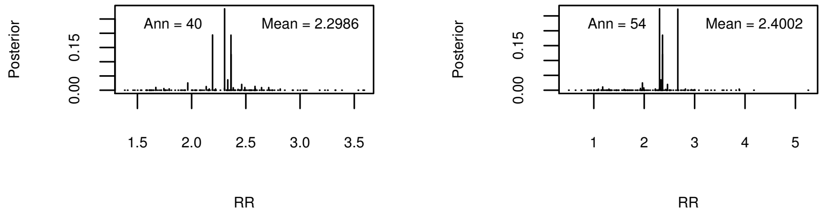


Figura 1 Risk ratios for two individuals who succeded and required the hint

5 An experiment

We carried on an experiment at the University of Florence, School of Engineering, Fall 2017. We asked to 160 students to solve a simple probabilistic question and we provided randomly an hint (the treatment T). Before the test we asked to the students if they wish to be helped or not ($E = 0$ and $E = 1$).
We are interested to investigate if there is a causal relationship between the hint and the ability to solve the question, for the students who wished to take the hint. We had 8 of these cases and for all of them we found a $RR > 2$ (obtained by averaging the results of different models according to their posterior probability, see figure (1)). This implies a lower bound for the probability of causation greater than 0.5 which suffices to indicate a causal relation between the hint and the positive result obtained by the 8 students.

As a result of our experiment we have also that, among the students who asked and had the hint, 24 did not succeed. We can image that one of them claimed that it was the hint which caused the failure. If we look now to the corresponding RR for such students, obtained as before by model averaging, we note that all of them have a values smaller then 1 (see figure 2). This is not conclusive that there is a causal relation between the hint and the failure.

In a civil trial this would not suggest to a judge to provide a compensation.

6 Conclusions

We introduced a typical Cause of Effect problem by means of an archetyppical example considering Ann and the effect of an aspirin on her headache. We have proposed a possible solution to make operational the choice of variables to include, so as to validate the fundamental assumptions underlying the assessment of Ann's probability of causation. We assume to have the possibility to perform a randomi-

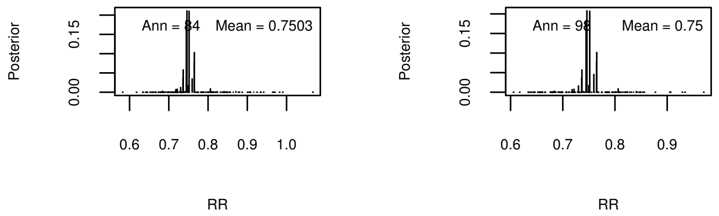


Figura 2 Risk ratios for two individuals who didn't succed and required the hint

zed sample from the Ann's population where, as usual, T is assigned following a randomized protocol and E, this is a novelty, is a question asking to the members of the sample about their preference to be treated or not.

Next step will be to extend the methodology to observational studies, to make possible in a wider range of cases the evaluation of the PC_A for Causes of Effects problems.

Riferimenti bibliografici

1. Chen, J. and Chen, Z., Extended Bayesian Information Criteria for Model Selection with large Model Spaces. *Biometrika*, **95**, 3, 759-771, 2008.
2. Dawid, A. P., The role of scientific and statistical evidence in assessing causality. In *Perspectives on Causation*, (ed. R. Goldberg), 133-147. Hart Publishing, Oxford, 2011.
3. Dawid, A. P., Faigman, D. L., and Fienberg, S. E., Fitting science into legal contexts: Assessing effects of causes or causes of effects? (with Discussion and authors' rejoinder). *Sociological Methods and Research*, **43**, 359–421, 2014.
4. Dawid, A. P., Musio, M. and Fienberg, S. E., From Statistical Evidence to Evidence of Causality. *Bayesian Analysis*, **11**, 725-752, 2016.
5. Dawid, A. P., Musio, M. and Murtas, R., The Probability of Causation, *Law, Probability and Risk*, **16**, 4, 163-179, 2017.
6. Holland, P. W., Statistics and Causal Inference *Journal of the American Statistical Association*, **81**, 396, 945-960, 1986.
7. Rubin, D. B., Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**, 688–701, 1974.

Evaluation and reporting of scientific evidence: the impact of partial probability assignments

La valutazione dell'evidenza scientifica: assegnazioni di probabilità parziali

Silvia Bozza, Alex Biedermann, Franco Taroni

Abstract The assessment of the value of scientific evidence can be performed by the derivation of a likelihood ratio, a rigorous concept that provides a measure of the change produced by an item of information in the odds in favor of a proposition as opposed to another. This represents a demanding task with several sources of uncertainty, due for example to elicitation of probabilities or to computational impasses. While use of such a metric is well established and supported by operational standards, opinions about what should be an appropriate way to deal with such sources of uncertainty while presenting expressions of evidential value at trial differ. Some quarters promote positions according to which practitioners should state a range of values for the probabilities of the evidence given competing propositions, and report a range of values for the likelihood ratio. However, such partial probability assignments may not make good use of available information.

Abstract *La valutazione del valore delle prove scientifiche può essere eseguita attraverso il rapporto di verosimiglianza, un concetto rigoroso che fornisce una misura del cambiamento prodotto da un elemento di prova nelle probabilità a favore di una proposizione rispetto ad un'altra. Questo rappresenta un compito impegnativo con diverse fonti di incertezza, dovute ad esempio alla necessità di assegnare valori di probabilità o a difficoltà computazionali. Mentre l'uso di tale metrica è ben consolidato e supportato da standard operativi, vi sono opinioni discordanti su quale dovrebbe essere il modo più appropriato per affrontare tali fonti di incertezza*

Silvia Bozza
Ca'Foscari University of Venice, Department of Economics, 30121 Venice, Italy
University of Lausanne, School of Criminal Justice, 1015 Lausanne-Dorigny, Switzerland
e-mail: silvia.bozza@unive.it

Alex Biedermann
University of Lausanne, School of Criminal Justice, 1015 Lausanne-Dorigny, Switzerland
e-mail: alex.biedermann@unil.ch

Franco Taroni
University of Lausanne, School of Criminal Justice, 1015 Lausanne-Dorigny, Switzerland
e-mail: franco.taroni@unil.ch

mentre si presentano le espressioni del valore probatorio al processo. Seguendo il dibattito in corso, gli esperti forensi sarebbero invitati ad indicare, per ragioni di trasparenza, un intervallo di valori per le probabilità dell'evidenza date le ipotesi di interesse e riportare un intervallo di valori per il rapporto di verosimiglianza. Tuttavia, tali assegnazioni di probabilità parziali potrebbero non fare un buon uso delle informazioni disponibili.

Key words: Likelihood ratio, partial probability assignments, uncertainty.

1 The likelihood ratio framework

Forensic scientists are typically faced to the evaluation of measurements on characteristics of trace evidence. The use of the likelihood ratio as a metric to assess the probative value of forensic traces is largely supported by operational standards and recommendations in forensic disciplines [3]. However, the progress towards more widespread consensus about foundational principles is still fragile and there are different views on how the strength of evidence conclusions should be reported to the court. The assessment of a likelihood ratio may turn out to be a task subjected to various sources of uncertainty, ranging from the problem of eliciting probabilities, to statistical issues related to the model choice, to sensitivity issues related to the choice of a prior distribution, or to even computational impasses that emerge when the marginal likelihoods are not available in closed form and numerical procedures need to be implemented. There is actually an open debate on the topic of whether the precision of forensic likelihood ratios should be measured and how should be reported to the court. A special edition edited by Geoff Morrison has recently been published in *Science & Justice* ([5] and subsequent papers). From one side, there is a school of thought according to which a forensic expert should report a single value for a likelihood ratio (e.g. [8], [6]). The likelihood ratio being expressed as a ratio of conditional probabilities (or marginal likelihoods whenever the evidence is expressed in terms of continuous measurements) is itself a measure of uncertainty. It represents the best assignment a forensic scientist can provide given data, model and background information. From the other side, it is questioned whether scientists should report interval quantifications as a surrogate for the value of the evidence (e.g. [7]) to acknowledge for uncertainty in likelihood ratio assessment. It is argued that reporting a single value would deprive the legal justice system of essential information needed to assess the reliability of the evidence. This discussion has echoed also in statistical literature, see for example [1].

It must be acknowledged that the discussion took different directions, leading in some cases arguments against the subjectivist interpretation of probabilities or against the Bayesian reasoning scheme, in others starting from different points of view with different interpretations of the same concept of likelihood ratio. It should be emphasized that in reality the fundamental point of this whole discussion is not the defense or not of a subjectivist approach. Nothing prevents, for example, to incorporate ‘reassuring’ relative frequencies to inform subjective probabilites ([9]). What really matters is finding the thread of the whole discussion, understanding whether to bring uncertainty about the expressions of uncertainty can actually lead to a good use of information taking into account that the ultimate goal must be to help justice.

2 On partial probability assignments

The discussion and related disagreements originate (also) from the fact that presenting a numerical value for probabilities or marginal likelihoods in the numerator and in the denominator of the likelihood ratio may give a false impression of exactitude, as such a precision may be in fact rarely realistic.

Consider the case where the evidence, E_1, is expressed in terms of a correspondence of genetic profiles between a person of interest and a recovered stain on a crime scene. What is the probability of observing corresponding evidential findings? Should the expert report his uncertainty, or not reporting it could it be misleading to the court? For this reason a forensic scientist may feel the necessity to present a range of values to minimize their personal involvement in the case. Let us therefore admit a partial probability assignment for both the numerator and the denominator of the likelihood ratio, say $\Pr(E_1 \mid H_p) = (l_p, u_p)$, for some $l_p < u_p$, and $\Pr(E_1 \mid H_d) = (l_d, u_d)$, for some $l_d < u_d$, where H_p and H_d designate propositions put forward by the prosecution and the defence, respectively.

On the other hand, a trier of fact could also be vague about their beliefs as to prior odds on the propositions that the person of interest is the source of the crime stain or another unrelated person is the source of that stain. What is the probability associated to the defendant’s liability? For this reason, let us admit a partial probability assignment for the prior probability of proposition H_p too, say $\Pr(H_p) = (l_h, u_h)$, for some $l_h < u_h$. Suppose now laboratory results are available so that the posterior probabilities of the competing propositions can be computed. For this purpose, it is useful to refer to the example originally sketched out by Frosini in 1989 ([2]) because it is well suited to the forensic issues under discussion. Consider the following partial probability assignments for the probabilites of interest:

$$\Pr(H_p) = (0.1, 0.2) \quad ; \quad \Pr(E_1 \mid H_p) = (0.6, 0.8) \quad ; \quad \Pr(E_1 \mid H_d) = (0.3, 0.5).$$

Assuming that values for the prior probabilities and for the likelihoods of the evidence under the competing propositions are uniformly spread over the assigned intervals, several values are randomly generated from each interval and for each realized triplet the posterior probability of the proposition supported by the prosecution is computed. The posterior probability, expressed by means of intervals, is $\Pr(H_p \mid E_1) = (0.12, 0.3)$. The impact of the evidence is to increase vagueness in the probability assignment for the propositions of interest changing the range of the probabilities from $(0.1, 0.2)$ to $(0.12, 0.3)$.

Suppose now that new findings are available, giving rise to new evidence (e.g. in terms of a correspondence between the recovered and control material from a suspect) denoted by E_2. Following the same line of reasoning, and considering for sake of simplicity the same partial probability assignments for the numerator and the denominator of the likelihood ratio that were assigned to evidence E_1, a new posterior partial probability assignment is obtained for the prosecution proposition, $\Pr(H_p \mid E_1, E_2) = (0.14, 0.45)$. This process can be reiterated many times. One may easily observe that the range of vagueness, at least initially, increases, though this may be felt counterintuitive as the effect of the evidence should be of reducing the initial size of the range of probabilities on the propositions of judicial interest. Posterior probabilities of the prosecution proposition H_p expressed in terms of partial probability assignments are depicted in Figure 2. Note that the range of probabilities assigned for new available findings E_i is kept fixed, $\Pr(E_i \mid H_p) = (0.6, 0.8)$ and $\Pr(E_i \mid H_d) = (0.3, 0.5)$ for $i = 1, 2, \ldots, n$. Though the observation of a correspondence between evidential findings will clearly shift the posterior odds versus the prosecution statement, one may observe that the effect of the evidence is to increase prior vagueness, at least until a large amount of findings is available.

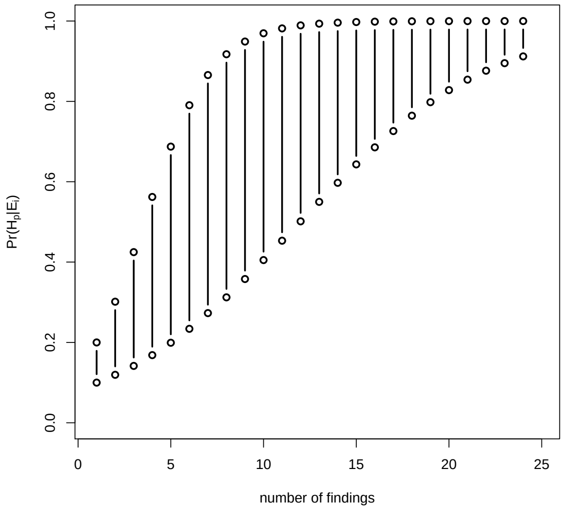


Fig. 1 Posterior probabilities proposition H_p expressed in terms of partial probability assignments: $\Pr(H_p) = (0.1, 0.2)$. Likelihoods are also expressed in terms of partial probability assignments, $\Pr(E_i \mid H_p) = (0.6, 0.8)$ and $\Pr(E_i \mid H_d) = (0.3, 0.5)$, $i = 1, \ldots, 24$.

This is clearly a simulated example. In a real case, the assumption of constant partial probability assignments for the likelihoods in the numerator and denominator in correspondence of new available evidence may be felt too restrictive and difficult to defend. Each assignment will reflect the uncertainty of the expert (or experts, whenever there are different laboratories in charge of the analyses of different recovered stains or marks) about a case-specific result. However, it serves the purpose to show in a simple way that the produced effect of such partial probability assignments is counterintuitive and it does not represent the answer the legal system would expect.

3 Conclusions

There may be different levels of resolution for the value of the likelihood ratio. This prompts scientists to elaborate ways to construct intervals or distributions over probabilities and likelihood ratios. It is argued that by reporting a single value, a forensic scientist deprives the legal justice system of essential information needed to assess the reliability of the evidence and this would amount to be highly misleading. However, nothing will be gained if a particular expression for uncertainty is itself obscured by an additional level of uncertainty ([4]). Not only such intervals provide no guidance to recipient of expert information as to how such pairs of values ought to be used, but also but also ranges of posterior probability may be larger than ranges of the corresponding prior probability. The conclusion of the discussion is that, in a given case at hand, forensic scientists ought to offer to a court of justice a given single value for the likelihood ratio. It is obviously desirable that reported likelihood ratios be accompanied with information to help fact-finders understand how and on what bases forensic scientists have reached their conclusions. Reporting a single value does not prevent a scientist, whenever asked, to respond about the strength of their beliefs.

References

1. Bickel, D.R.: Reporting Bayes factors or probabilities to decision makers of unknown loss functions. Communications in Statistics - Theory and Methods (2018) doi: 10.1080/03610926.2018.1459713
2. Frosini, B.V..: La statistica metodologica nei convegni della SIS In: Atti del convegno 'Statistica e società', pp. 197—227, Pisa, 9-10 ottobre 1989.
3. ENFSI Guideline for evaluative reporting in forensic science. Bruxelles (2015).
4. Lindley, D. V. (2014) Understanding Uncertainty, revised edition. John Wiley & Sons, Hoboken.
5. Morrison, G.S.: Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. Science & Justice **56**, 371–373 (2016).
6. Ommen, D.M., Saunders, C.P., Neumann, C. An argument against presenting interval quantifications as a surrogate for the value of the evidence. Science & Justice **56**, 383–387 (2016).
7. Sjerps, M.J., Alberink, I., Bolck, A., Stoel, R.D., Vergeer, P., van Zanten, J.H.: Uncertainty and LR: to integrate or not to integrate, that's the question. Law, Probability & Risk **15**, 23–29 (2016).

8. Taroni, F., Bozza, S., Biedermann, A., Aitken, A.: Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. Law, Probability & Risk **169**, 321–354 (2016)
9. Taroni, F., Garbolino, P., Biedermann, A., Aitken, A., Bozza, S.: Reconciliation of subjective probabilities and frequencies in forensic science Law, Probability & Risk (in press).

Missing Data Handling in Complex Models

Dependence and sensitivity in regression models for longitudinal responses subject to dropout

Dipendenza e sensibilità nei modelli ad effetti casuali per dati longitudinali in presenza di dropout

Marco Alfò and Maria Francesca Marino

Abstract In longitudinal studies, subjects may be lost to follow-up and present incomplete response sequences. When the mechanism that leads to exit the study is non ignorable, a possible route is to define a model that accounts for potential dependence between the longitudinal and the dropout process. This model should have, at least, two major features: (*i*) it should (simply) reduce to an ignorable missing data model, when some conditions are met; (*ii*) the nested structure should give the way to measure sensitivity of parameter estimates to assumptions on non ignorability. In this work, we discuss random coefficient based dropout models and review measures of local sensitivity.

Abstract Negli studi longitudinali, alcuni soggetti abbandonano lo studio prima del suo completamento, presentando sequenze incomplete. Quando il meccanismo di generazione del dato mancante è non ignorabile, si può considerare un modello che descriva la dipendenza tra processo longitudinale e generazione del dato mancante stesso. Tale modello dovrebbe includere, come caso particolare, il modello per dati mancanti di tipo ignorabile, e permettere un'analisi di sensibilità delle stime rispetto alle ipotesi fatte circa il meccanismo di generazione dei dati mancanti stessi. In questo lavoro, si discutono i modelli a coefficienti casuali per l'analisi di studi longitudinali con dati mancanti non ignorabili e si confrontano misure di sensibilità locale.

Marco Alfò
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, e-mail: marco.alfo@uniroma1.it

Maria Francesca Marino
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, e-mail: mariafrancesca.marino@unifi.it

1 Introduction

Longitudinal studies entail repeated measurements from the same units over time. Often, units leave the study before the planned end, leading to *dropout* (also referred to as *attrition*) According to Rubin's taxonomy [19], if the probability of a missing response, conditional on observed data, does not depend on the responses that should have been observed, the data are said to be missing completely at random (MCAR), or missing at random (MAR). When, even after conditioning on observed data, the mechanism still depends on the unobserved responses, data are referred to as missing not at random (MNAR). In the context of likelihood inference, when either the parameters in the measurement and the missingness process are not distinct or the missing data process is MNAR, missing data are non ignorable (NI). In this case, some form of joint modeling of the longitudinal response and the missing data process is required [12].

Random Coefficient Based Dropout Models (RCBDMs, [11]) may be used as a quite general approach in this context. Here, two separate (conditional) models are built for the longitudinal response and the missingness indicator. Dependence arises due to models sharing common/dependent unit- and (possibly) outcome-specific random parameters. The model specification is completed by adopting an appropriate distribution for these random parameters, which can be either fully parametric [23, 8], or semi-parametric [2, 3]. This latter approaches have been introduced in the literature to avoid the impact that parametric assumptions may have on inference [20], especially in the case of short longitudinal sequences. More elaborated approaches are also available in the literature [5, 6, 4]. Besides the advantages of the semi-parametric approach, it presents the substantial drawback that dependence *within* outcomes can not be separated by dependence *between* outcomes. Starting from this drawback, we define a bi-dimensional finite mixture model for longitudinal data subject to dropout [21]. When the missing data mechanism is ignorable, such MNAR model directly reduces to its MAR counterpart. See also [14] for a dynamic extension of the model. Sensitivity of parameter estimates to assumptions on non ignorability of the dropout process can be explored by adopting either a global or a local perspective. Within the latter, we discuss the so-called *index of sensitivity to non ignorability* (ISNI) proposed by [22] and [13]. We show that, if the proposed model specification is employed, this approach to sensitivity analysis can be seen as a particular version of local influence diagnostics [10, 17, 18]. Obviously, a *global* influence approach could be adopted as well, for example by looking at the *mean score* approach by [25].

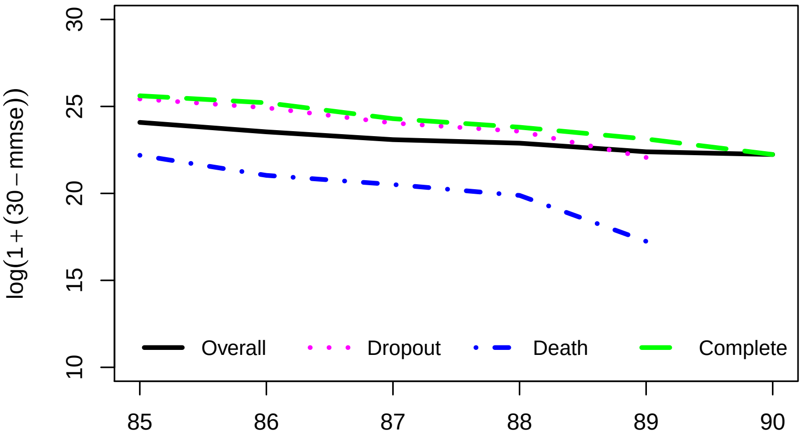
The structure of the paper follows. In section 2 we introduce the motivating application, the Leiden 85+ study, entailing the dynamics of cognitive functioning in the elderly. Section 3 discusses general random coefficient based dropout models, while sensitivity analysis is described in section 4.

2 Motivating example: Leiden 85+ data

To discuss our proposal, we consider data from the Leiden 85+ study, a retrospective study on 705 Leiden (Netherlands) inhabitants, who reached the age of 85 years between September 1997 and September 1999. The study aim was at identifying demographic and genetic determinants of cognitive functioning dynamics in the elderly. The following covariates were collected at the beginning of the study: gender, educational status (primary/higher education), plasma Apolipoprotein E (APOE) genotype (22-23, 24, 33, 34-44). Only 541 subjects present complete covariate information and will be considered in the following. Study participants were visited at their place of residence once a year until the age of 90; orientation, attention, language skills and ability to perform simple actions were assessed through a 30-items questionnaire. The Mini Mental State Examination index (MMSE, [7]), is obtained by summing the binary scores on such 30 items.

We report in Figure 1 the evolution of the mean response over time, stratified by participation.

Fig. 1: Mean *MMSE* value over time stratified by subjects' participation to the study.



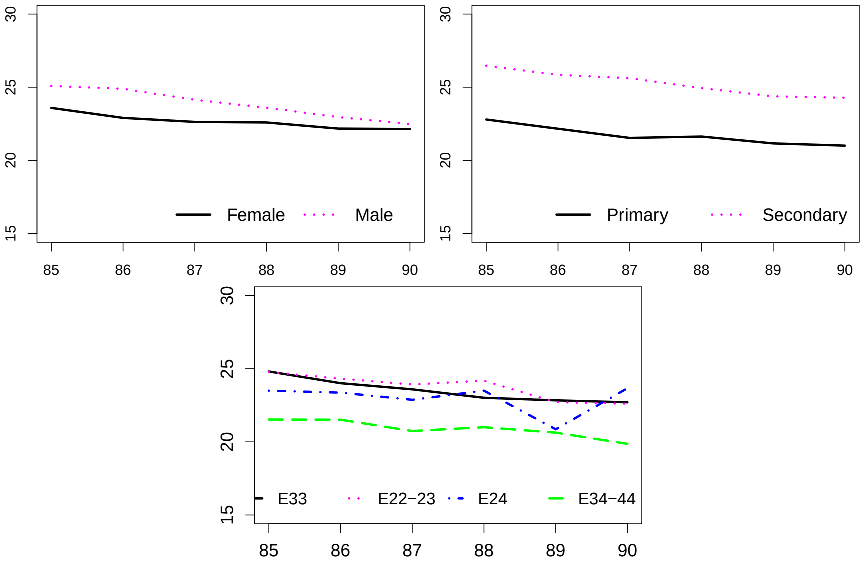
By looking at this figure, we may observe that, while the decline over time in the MMSE mean is (at least approximately) constant across the groups defined by patterns of dropouts, the differential participation in the study leads to a different slope for the overall mean score. Such a finding highlights a potential dependence between the evolution of the response over time and the dropout process, which may bias parameter estimates and corresponding inference. We report in Table 1 the distribution of the observed covariates by pattern of participation. This suggests a differential participation in the study by gender and educational level, while differences can be observed only for $APOE_{34-44}$ group.

Table 1: Leiden 85+ Study: demographic and genetic characteristics of participants

Variable	Total	Completed (%)	Did not complete (%)	(Row) Total
Gender				
Male	180 (33.27)	74 (41.11)	106 (58.89)	100
Female	361 (66.73)	192 (53.19)	169 (46.81)	100
Education				
Primary	351 (64.88)	166 (47.29)	185 (52.71)	100
Secondary	190 (35.12)	100 (52.63)	90 (47.37)	100
APO-E				
22-23	96 (17.74)	54 (56.25)	42 (43.75)	100
24	12 (2.22)	6 (50.00)	6 (50.00)	100
33	319 (58.96)	162 (50.78)	157 (49.22)	100
34-44	114 (21.08)	44 (38.60)	70 (61.40)	100
Total	541 (100)	266 (49.17)	275 (50.83)	100

Figure 2 depicts the dynamics of mean MMSE score over time by available covariates. From this figure, it is evident that cognitive impairment is lower for males than females, even if the difference seems to decrease with age, maybe due to a differential dropout by gender. No further interaction with age can be evinced, as the dynamics seem to be consistent for both levels of education, and for all the 4 levels of APOE genotype, but for the one with a very reduced sample size ($APOE_{24}$).

Fig. 2: Leiden 85+ Study: mean of MMSE score stratified by age and gender, educational level, APOE



3 Random coefficient-based dropout models

Let Y_{it} represent a response recorded on $i = 1,\dots,n$, subjects at time occasions $t = 1,\dots,T$, and let $\mathbf{x}_{it} = (x_{it1},\dots,x_{itp})'$ be a vector of observed covariates. We assume that, conditional on a q-dimensional set of individual-specific random coefficients $\mathbf{b}_i$, the observed responses are independent draws from a distribution in the Exponential Family with canonical parameter defined by

$$\theta_{it} = \eta_{it}^{Y} = \mathbf{x}_{it}'\beta + \mathbf{z}_{it}'\mathbf{b}_i.$$

The terms $\mathbf{b}_i$, $i = 1,\dots,n$, describe unobserved, individual-specific, heterogeneity (which may also be time-varying), while β is a vector of fixed parameters. Usually, $\mathbf{z}_{it} = (z_{it1},\dots,z_{itq})'$ represents a subset of $\mathbf{x}_{it}$. For identifiability purposes, standard assumptions on the random coefficient vector are introduced: $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{b}_i) = \mathbf{D}$ for $i = 1\dots,n$.

Let $\mathbf{R}_i$ denote the vector of missing data indicators, with generic element $R_{it} = 1$ if the i-th unit drops-out at any point in the window $(t-1,t)$, $R_{it} = 0$ else. As we focus on dropouts, we have $R_{it'} = 1, \forall t' > t$, so that $T_i \le T$ measures are available for each study participant. We consider studies in discrete time; however, most of the following arguments may apply, with a limited number of changes, to (continuous time) survival process as well. To describe potential dependence between the longitudinal and the dropout process, we introduce an explicit model for the latter, conditional on a set of covariates, say $\mathbf{w}_i$, and a subset of the random coefficients in the longitudinal model. That is, we assume that, conditional on $\mathbf{b}_i^* = \mathbf{C}\mathbf{b}_i$, $i = 1,\dots,n$, $\mathbf{C} \in \mathscr{M}_{\mathbf{q},\mathbf{q_R}}$, dropout indicators are independent and follow a Bernoulli distribution with probability ϕ_{it} defined by:

$$\mathrm{logit}(\phi_{it}) = \eta_{it}^{R} = \mathbf{w}_{it}'\gamma + \mathbf{v}_{it}'\mathbf{b}_i^*. \tag{1}$$

Previous equations define a so-called shared (random) coefficient model [27, 26]. The assumption is that the longitudinal response and the dropout indicator are independent conditional on the individual-specific random coefficients:

$$f_{Y,R}(\mathbf{y}_i,\mathbf{r}_i \mid \mathbf{X}_i,\mathbf{W}_i) = \int \left[\prod_{t=1}^{T_i} f_Y(y_{it} \mid \mathbf{x}_{it},\mathbf{b}_i) \prod_{t=1}^{\min(T,T_i+1)} f_R(r_{it} \mid \mathbf{w}_{it},\mathbf{b}_i)\right] dG(\mathbf{b}_i), \tag{2}$$

Dependence between the measurement and the missingness, if any, is completely accounted for by the latent effects which are also used to describe unobserved, individual-specific, heterogeneity in each of the two (univariate) profiles. This class of models has been further extended by [28] to *joint* models where a continuous time setting and a survival data model are considered:

$$h_i(t) = h_0(t)\exp(\mathbf{w}_{it}'\gamma + \alpha\eta_{it}^Y). \tag{3}$$

As an alternative, we may consider equation-specific random coefficients [1]. In this respect, let $\mathbf{b}_i = (\mathbf{b}_{i1}, \mathbf{b}_{i2})$ denote an individual- and outcome-specific random coefficient. Introducing a local independence assumption, the joint density for the couple $(\mathbf{Y}_i, \mathbf{R}_i)$ can be written as follows:

$$f_{Y,R}(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{W}_i) = \int \left[\prod_{t=1}^{T_i} f_Y(y_{it}|\mathbf{x}_{it}, \mathbf{b}_{i1}) \prod_{t=1}^{\min(T,T_i+1)} f_R(r_{it}|\mathbf{w}_{it}, \mathbf{b}_{i2})\right] dG(\mathbf{b}_{i1}, \mathbf{b}_{i2}). \tag{4}$$

A further approach is that proposed by [6], where common, partially shared and independent (outcome-specific) random coefficients are considered in the measurement and the dropout process. For example, in the current context, we may write

$$\mathbf{b}_{i1} = \mathbf{b}_i + \varepsilon_{i1}\mathbf{b}_{i2} = \mathbf{b}_i + \varepsilon_{i2}, \qquad \varepsilon_{i1} \perp \varepsilon_{i2}.$$

This can be further extended to consider partially shared effects.

4 Sensitivity analysis: definition of the index

As highlighted by [15], for every MNAR model we may define a MAR counterpart that produces exactly the same fit to the observed data. Two issues are worth to be noticed. First, the MNAR model is fitted to the observed data only, assuming that the distribution of the missing responses is identical to that of the observed ones. Second, the structure describing dependence between the longitudinal responses (observed and missing) and the dropout indicators is just one out of several possible choices. Therefore, we may be interested in evaluating how much maximum likelihood estimates are influenced by hypotheses on the dropout mechanism.

Looking at *local* sensitivity, [22] defined the index of local sensitivity to non ignorability (ISNI) using a first-order Taylor expansion of the log-likelihood function. The aim was at describing the behaviour of parameter estimates in a neighbourhood of the MAR solution. The index was further extended by [9] by considering a second-order Taylor expansion; more general settings and different metrics were also considered [13, 31, 29, 30, 24].

To specify the index of local sensitivity, let $\lambda = (\lambda_{11}, \ldots, \lambda_{K_1K_2})$ denote the vector of non ignorability parameters, with $\lambda = \mathbf{0}$ corresponding to the MAR model. Furthermore, let $\hat{\Phi}(\lambda)$ denote the ML estimates obtained conditional on a given value of λ. The ISNI may be written as

$$ISNI_{\Phi} = \left.\frac{\partial\hat{\Phi}(\lambda)}{\partial\lambda}\right|_{\Phi(0)} \simeq -\left(\left.\frac{\partial^2\ell(\Phi,\Psi,\pi)}{\partial\Phi\Phi'}\right|_{\Phi(0)}\right)^{-1} \left.\frac{\partial^2\ell(\Phi,\Psi,\pi)}{\partial\Phi\lambda}\right|_{\Phi(0)} \tag{5}$$

It measures the displacement of model parameter estimates from their MAR counterpart, in the direction of λ. Following [29], the following equation holds:

$$\hat{\Phi}(\lambda) = \hat{\Phi}(\mathbf{0}) + ISNI_{\Phi}\lambda;$$

The ISNI may be also interpreted as the linear impact that λ has on $\hat{\Phi}$. By using the proposed bi-dimensional model specification, we may show that the sensitivity analysis based on $ISNI_{\Phi}$ can be linked to local influence diagnostics developed for regression models to check for influential observations by perturbing individual-specific weights [10, 17, 18]. Here, we perturb weights associated to groups of subjects, rather than individual observations, See e.g. [16] for a comparison between multiple imputation and perturbation schemes in the more general setting of masking individual microdata.

References

1. Aitkin, M., Alfò, M.: Variance component models for longitudinal count data with baseline information: epilepsy data revisited. Statistics and Computing **16**, 231–238 (2006)
2. Alfò, M., Aitkin, M.: Random coefficient models for binary longitudinal responses with attrition. Statistics and Computing **10**, 279–287 (2000)
3. Alfò, M., Maruotti, A.: A selection model for longitudinal binary responses subject to non-ignorable attrition. Statistics in Medicine **28**, 2435–2450 (2009)
4. Bartolucci, F., Farcomeni, A.: A discrete time event-history approach to informative drop-out in mixed latent markov models with covariates. Biometrics **71**, 80–89 (2015)
5. Beunckens, C., Molenberghs, G., Verbeke, G., Mallinckrodt, C.: A latent-class mixture model for incomplete longitudinal gaussian data. Biometrics **64**, 96–105 (2008)
6. Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., Kenward, M.: Generalized shared-parameter models and missingness at random. Statistical Modelling **11**, 279–310 (2011)
7. Folstein, M., Folstein, S., McHig, P.: Mini-mental state: a pratical method for grading the cognitive state of patients for the clinician. Journal of Psychiatry Research **12**, 189–198 (1975)
8. Gao, S.: A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. Statistics in Medicine **23**, 211–219 (2004)
9. Gao, W., Hedeker, D., Mermelstein, R., Xie, H.: A scalable approach to measuring the impact of nonignorable nonresponse with an ema application. Statistics in Medicine **35**, 5579–5602 (2016)
10. Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., Van Steen, K.: A local influence approach applied to binary data from a psychiatric study. Biometrics **59**, 410–419 (2003)
11. Little, R.: Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association **90**, 1112–1121 (1995)
12. Little, R., Rubin, D.: Statistical analysis with missing data, 2nd edition. Wiley (2002)
13. Ma, G., Troxel, A., Heitjan, D.: An index of local sensitivity to non-ignorability in longitudinal modeling. Statistics in Medicine **24**, 2129–2150 (2005)
14. Marino, M., Alfò, M.: A non-homogeneous hidden markov model for partially observed longitudinal responses. arXiv p. arXiv:1803.08255 (2018)
15. Molenberghs, G., Beunckens, C., Sotto, C., Kenward, M.: Every missing not at random model has got a missing at random counterpart with equal fit. Journal of the Royal Statistical Society, Series B **70**, 371–388 (2008)
16. Muralidhar, K., Sarathy, R.: A comparison of multiple imputation and data perturbation for masking numerical variables. Journal of Official Statistics **22**, 507–524 (2006)

17. Rakhmawati, T., Molenberghs, G., Verbeke, G., Faes, C.: Local influence diagnostics for hierarchical count data models with overdispersion and excess zeros. Biometrical journal **58**, 1390–1408 (2016)
18. Rakhmawati, T., Molenberghs, G., Verbeke, G., Faes, C.: Local influence diagnostics for generalized linear mixed models with overdispersion. Journal of Applied Statistics **44**, 620–641 (2017)
19. Rubin, D.: Inference and missing data. Biometrika **63**, 581–592 (1975)
20. Scharfstein, D., Rotnitzky, A., Robins, J.: Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association **94**, 1096–1120 (1999)
21. Spagnoli, A., Marino, M., Alfò, M.: A bidimensional finite mixture model for longitudinal data subject to dropout. Statistics in medicine **DOI:10.1002/sim.7698**, to appear (2018)
22. Troxel, A., Ma, G., Heitjan, D.: An index of local sensitivity to non-ignorability. Statistica Sinica **14**, 1221–1237 (2004)
23. Verzilli, C., Carpenter, J.: A monte carlo em algorithm for random-coefficient-based dropout models. Journal of Applied Statistics **29**, 1011–1021 (2002)
24. Viviani, S., Rizopoulos, D., Alfò, M.: Local sensitivity to non-ignorability in joint models. Statistical Modelling **14**, 205–228 (2014)
25. White, I., Carpenter, J., Horton, N.: A mean score method for sensitivity analysis to departures from the missing at random assumption in randomised trials. Statistica Sinica **to appear** (2017)
26. Wu, M., Bailey, K.: Estimation and comparison of changes in the presence of informative right censoring: conditional linear models. Biometrics **45**, 939–955 (1989)
27. Wu, M., Carroll, R.: Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics **44**, 175–188 (1988)
28. Wulfsohn, M., Tsiatis, A.: A joint model for survival and longitudinal data measured with error. Biometrics **53**, 330–339 (1997)
29. Xie, H.: A local sensitivity analysis approach to longitudinal non-gaussian data with non-ignorable dropout. Statistics in Medicine **27**, 3155–3177 (2008)
30. Xie, H.: Analyzing longitudinal clinical trial data with nonignorable missingness and unknown missingness reasons. Computational Statistics and Data Analysis **56**, 1287–1300 (2012)
31. Xie, H., Heitjan, D.: Sensitivity analysis of causal inference in a clinical trial subject to crossover. Clinical Trials **1**, 21–30 (2004)

Multilevel analysis of student ratings with missing level-two covariates: a comparison of imputation techniques

Analisi multilivello dell'opinione degli studenti universitari in presenza di valori mancanti delle variabili di secondo livello: un confronto tra metodi di imputazione

Maria Francesca Marino e Carla Rampichini

Abstract We analyse the relationship between student ratings of university courses and several characteristics of the student, the course and the teacher. In particular, we exploit data from a survey collecting information about teacher beliefs and practices at the University of Padua in a.y. 2012/13. Student ratings are nested into classes, calling for multilevel modelling. However, due to survey non-response, the information about beliefs and practices is missing for about half of the teachers, posing a serious issue of missing data at level 2. To avoid listwise deletion, we make multiple imputation via fully conditional specification, exploiting information at all hierarchical levels. The proposed approach turns out to be effective. We found that some of the teacher beliefs and practices are significantly related to student ratings.

Abstract *Il presente lavoro analizza la relazione tra le valutazioni dei corsi universitari fornite dagli studenti e le caratteristiche di docenti e studenti. I dati provengono da un'indagine ad hoc svolta dall'Università di Padova nell'a.a. 2012/13 su convinzioni e pratiche dei docenti universitari. La struttura gerarchica dei dati, con valutazioni raggruppate per insegnamento, richiede l'uso di modelli multilivello. L'indagine presenta un alto tasso di non-risposta, ponendo un serio problema di dati mancanti a livello 2. Si considera un'imputazione multipla basata sull'approccio* fully conditional*, sfruttando anche l'informazione proveniente dalle unità di livello 1. I risultati ottenuti evidenziano l'efficacia del metodo utilizzato. Inoltre, risulta che alcune delle convinzioni e delle pratiche adottate dai docenti sono significativamente correlate alle valutazioni degli studenti.*

Key words: fully conditional specification, multiple imputation, multivariate mixed model

Maria Francesca Marino
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence - Italy, e-mail: mariafrancesca.marino@unifi.it

Carla Rampichini
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence - Italy, e-mail: carla.rampichini@unifi.it

1 Introduction

We analyse student satisfaction, as measured by student evaluation of teaching (SET). The peculiarity of the study lies in the availability of many variables about teacher characteristics and beliefs, and teaching practices. This work exploits data from the University of Padua for academic year 2012/13 about bachelor degree courses. The data set is obtained by merging three different sources: (*i*) the traditional SET survey with 18 items, measured on a ten-point scale (1: low, 10: high); (*ii*) administrative data on students, teachers and didactic activities; (*iii*) an online survey carried out by the PRODID project on teacher beliefs and practices.

Data have a two-level hierarchical structure, with $56,775$ student ratings at level 1 and $1,016$ classes at level 2. The average class size is 79 (min 5, max 442). We are interested in student satisfaction about two key aspects of teaching, i.e. teacher ability to involve students (item D06 of the SET questionnaire) and teacher clarity (item D07).

The analysis is based on the following bivariate two-level linear mixed model for item *m* (*m*: 1 for D06, 2 for D07) recorded on student *i* in class *j*:

$$Y_{mij} = \alpha_m + \boldsymbol{\beta}'_m \mathbf{x}_{ij} + \boldsymbol{\gamma}'_m \mathbf{w}_j + u_{mj} + e_{mij} \qquad (1)$$

where $\mathbf{x}_{ij}$ is the vector of student covariates (level 1), and $\mathbf{w}_j$ is the vector of teacher and class covariates (level 2). Level 1 errors e_{mij} are assumed to be independent across students; level 2 errors (random effects) u_{mj} are assumed to be independent across classes and independent from level 1 errors. We make standard assumptions for the distributions of the model errors, including homoscedasticity (within each outcome) and normality. Therefore, the response vector $\mathbf{Y}_{ij} = (Y_{1ij}, Y_{2ij})'$ has residual variance-covariance matrix $Var(\mathbf{Y}_{ij}) = \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e$, where $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ are the covariance matrices of the errors at level 2 and level 1, respectively.

The survey on teacher beliefs and practices has about fifty percent of missing questionnaires, posing a serious issue of missing data at level 2. An analysis based on listwise deletion would discard the entire set of student ratings for non responding teachers, causing two main problems: (*i*) a dramatic reduction of the sample size, and thus of the statistical power, and (*ii*) possibly biased estimates if the missing mechanism is not MCAR.

2 Multiple imputation of level 2 covariates

In multilevel models, the treatment of missing data requires special techniques. In fact, the data have a hierarchical structure and, thus, missing values can be at any level of the hierarchy. Moreover, missing values can alter the variance components and the correlations. Multiple imputation (MI) is the most flexible approach to missing data. MI has been extended to the multilevel setting to deal with these special issues, following two main approaches (Mistler and Enders, 2017; Grund *et* *al.*, 2018): fully conditional specification, also known as multivariate imputation by chained equations, and joint modelling.

In our case study, the substantive model (1) is multilevel, however missing data are only at level 2. This feature makes the imputation simpler than in the general multilevel setting. Indeed, we can apply standard MI techniques to level 2 data and, then, merge level 1 and level 2 data to obtain complete datasets.

Besides this simplification, the MI step for the data under investigation remains a challenging matter, since we have to deal with a large number of categorical variables having a high percentage of missingness. As stated before, about 50% of the teachers did not respond to the whole questionnaire, thus producing missing values on 10 binary items (teacher practices) and 20 ordinal items (teacher beliefs on a seven-point scale). In particular, the imputation model at level 2 should include all the level 2 covariates and information on level 1 variables, especially on the response variables.

Several strategies to summarize information from level 1 variables can be adopted (Erler *et al.*, 2016; Grund *et al.* 2017). Here, we choose to consider the cluster means as they are rather effective and easy to be implemented in cases where level 1 variables (including the response) are completely observed, as in our case.

We perform multivariate imputation by chained equations using the `mi` command of Stata (StataCorp, 2017). In our case, we use a binary logit as imputation model for the 10 binary items (teacher practices), and a cumulative logit for the 20 ordinal items (teacher beliefs). For all imputation models, we consider the following covariates: the fully observed class and teacher characteristics, the cluster means of level 1 variables (covariates and outcomes), and the cluster size.

3 Results

The bivariate two-level model (1) is fitted by maximum likelihood on 10 imputed data sets, and the results are combined with the standard MI rules. The analysis is conducted using the `gsem` and `mi` commands of Stata (StataCorp, 2017).

We first fit the model without covariates in order to explore the correlation structure of the two outcomes, i.e. teacher ability to involve students and teacher clarity. We find out that the ICC is about 30% for both outcomes. The two outcomes are highly correlated (0.83), especially at level 2 (0.933).

Then, we add the available covariates in the model. These include 6 student characteristics, centered on class averages (gender, age, high school grade, year of enrollment, regular student, number of exams passed in 2012), 8 fully observed covariates at level 2, including teacher and course characteristics (gender, age, role, and involment of the teacher in the course; credits, class size, school and compulsoriness of the course) and 30 further level 2 covariates summarizing teacher practices and beliefs. To ensure parsimony, we select a subset of the 30 practices and beliefs using *leaps and bounds* method proposed by Furnival and Wilson (1974) and implemented in the `gvselect` command of Stata (Lindsey and Sheather, 2010). Specifically, we apply the procedure to each equation of model (1) separately and retain in the final model only those covariates significant for at least one of the two equations according to the AIC criterion. Such a procedure leads to the selection of the following binary indicators of teacher practices: Q01 – *practicals*, Q02 – *exploiting contribution from experts*, Q07 – *using multimedia resources*. As regards teacher beliefs, the following ordinal indicators are retained into the final model: Q12 – *passion for teaching*, Q14 – *usefulness of practicals*, Q17 – *usefulness of student opinions*, Q23 – *usefulness of student-oriented teaching*, Q25 – *need for teaching support*, Q27 – *usefulness of sharing teaching experiences with colleagues*.

To quantify the influence of missing data on the sampling variance of the parameter estimates, we can consider the fraction of missing information (FMI), i.e. the ratio between the sampling variance among imputations and the total sampling variance. For the imputed covariates, the FMI ranges from 0.15 to 0.68, with a median value equal to 0.44. The fraction of missing data is about 0.5; therefore, for the majority of imputed covariates, the trade-off between the sampling error inflation due to MI and its reduction due to sample size increase is favourable.

The main substantive finding is that teacher practices and beliefs from the PRODID survey are significantly related with the SET ratings. In particular, the item *practicals* is negatively related to the outcomes, while *contribution from experts* is positively related. As for teacher beliefs, *passion for teaching* and *usefulness of student opinions* are positively associated with the ratings, while *usefulness of practicals* and *need for teaching support* are negatively associated with the ratings.

An alternative imputation method is joint modelling (e.g. Grund *et al.*, 2017), which allows to impute binary and ordinal data through latent normal variables. We implement this approach using the R package `jomo` (Quartagno and Carpenter, 2017). However, in our case with 10 binary and 20 ordinal variables this approach is not feasible due to the computational burden, unless we treat the ordinal variables as continuous or we previously select a subset of the ordinal variables. The second option is preferable to preserve the original scale of measurement. A thorough comparison between chained equations and joint modelling is the object of our future research.

Acknowledgements This work has been supported by the research project ”Advances in Multilevel and Longitudinal Modelling” (BIRD162088/2016) of the University of Padua.

References

1. Erler, N.S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V.W.V., Franco, O.H., Lesaffre, E.M.E.H: Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian. Statistics in Medicine. **35**, 2955 – 2974 (2016)
2. Furnival, G. M., Wilson, R. W.: Regression by leaps and bounds. Technometrics. **16**, 499511 (1974)
3. Grund, S., Lüdtke, O., Robitzsch, A.: Multiple Imputation of Missing Data for Multilevel Models Simulations and Recommendations. Organizational Research Methods. **21**, 111 – 149

(2018)

4. Grund S., Lüdtke O., Robitzsch A.: Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs, Journal of Educational and Behavioral Statistics. (2017)
5. Lindsey, C., Sheather, S.: Variable selection in linear regression. The Stata Journal. **10**, 650–669 (2010)
6. Mistler, S. A., Enders, C. K.: A comparison of joint model and fully conditional specification imputation for multilevel missing data. Journal of Educational and Behavioral Statistics. **42**, 432 – 466 (2017)
7. Quartagno, M., Carpenter, J.: jomo: A package for Multilevel Joint Modelling Multiple Imputation. https://CRAN.R-project.org/package=jomo (2017)
8. StataCorp.: Stata, Release 15. Statistical Software. StataCorp LLC, College Station, TX (2017)

Multilevel Multiple Imputation in presence of interactions, non-linearities and random slopes

Imputazione Multipla Multilivello in presenza di interazioni, non-linearità e pendenze casuali

Matteo Quartagno and James R. Carpenter

Abstract Multiple Imputation is a flexible tool to handle missing data that has been increasingly used in recent years. One of the conditions for its validity is that the two models used for (i) imputing and (ii) analysing the data need to be compatible. For example, when the partially observed data have a multilevel structure, both models need to reflect this. Choosing an appropriate imputation model is more complicated when data are missing in a variable included in the substantive multilevel analysis model as a covariate with a random slope, an interaction or a non-linear term. We propose an imputation method based on joint modelling of the partially observed variables. We factor this joint model in two parts: a joint multilevel distribution for the covariates, and a conditional multilevel distribution for the outcome given the covariates. We guarantee compatibility by using as the second term the substantive analysis model. We fit this model with a Gibbs sampler, and we use a Metropolis-Hastings step to accept/reject the proposed draws for the missing values, to guarantee that they are actual random draws from the desired distribution. Our proposed imputation approach is theoretically consistent with the substantive model, and we demonstrate the marked improvements this brings by simulation.

Abstract *L'imputazione multipla é uno strumento flessibile per gestire dati mancanti la cui popolarità è aumentata considerevolmente negli ultimi anni. Una delle condizioni necessarie per la sua validità è che i due modelli utilizzati per (i) imputare e (ii) analizzare i dati siano compatibili. Per esempio, quando il dataset parzialmente osservato ha una struttura multilivello, entrambi i modelli devono tenerne conto. Scegliere un modello di imputazione adeguato è più complicato quando i dati sono mancanti in una variabile che è inclusa nel modello d'analisi multilivello di interesse come una covariata con una pendenza casuale, un'interazione o un*

Matteo Quartagno
London School of Hygiene and Tropical Medicine, Keppel Street, e-mail: matteo.quartagno@lshtm.ac.uk

James R. Carpenter
London School of Hygiene and Tropical Medicine, Keppel Street e-mail: james.carpenter@lshtm.ac.uk

termine non-lineare. Proponiamo qui un metodo di imputazione basato sulla modellizzazione congiunta delle variabili parzialmente osservate. Fattorizziamo questo modello congiunto in due parti: una distribuzione congiunta per le covariate ed una distribuzione condizionata per la variabile risposta date le covariate. Garantiamo la compatibilità usando per questo secondo termine la stessa formulazione del modello d'analisi di interesse. Fittiamo questo modello con un campionatore di Gibbs, ed utilizziamo un passo Metropolis-Hastings per accettare/rifiutare i valori proposti per i dati mancanti, per garantire che siano effettive estrazioni casuali dalla distribuzione desiderata. Mostriamo con simulazioni che questo metodo performa in modo appropriato e supera una strategia di imputazione alternativa.

Key words: Missing Data, Multiple Imputation, Multilevel Model, Random Slope, Interactions, Non-linearities

1 Introduction

Multiple imputation (MI) is a missing data handling method that has become very popular in recent years, particularly in the world of medical and social research. Key reasons for its growing popularity include its flexibility, the possibility to use for the analysis step the same model of substantive scientific interest that we would have used on a fully observed dataset and the chance to make use of auxiliary variables to retrieve some information [8, 3].

A key role in MI is played by the *imputation model*, i.e. the model that is used to impute the missing data. In order for MI to lead to valid inference, this needs to be consistent with the substantive analysis model [2]. For example, if the partially observed dataset has a multilevel structure, this needs to be reflected in the imputation model as well as in the analysis model [6].

Different methods have been proposed recently for multilevel MI [1]. The most flexible of these is Joint Modelling Multiple Imputation (JM-MI), which consists in assuming a joint multivariate normal model for the partially observed data, and in fitting this model with a Bayesian (e.g. Gibbs) sampler to impute the missing data. A multilevel version of JM-MI was first introduced in [9], and later extended to allow for binary and categorical data [5] and for cluster-specific covariance matrices [10]. However, in some circumstances it is not possible to find a simple joint imputation model that is fully compatible with the analysis model; some examples include the imputation of variables that are included in the substantive analysis model as covariates with a random slope, an interaction or a non-linear (e.g. quadratic) term. Using a heteroscedastic imputation model can be useful to deal with random slopes and interactions, as it allows for cluster-specific associations between variables [7]. However, full compatibility is still not guaranteed.

Goldstein et al. (2014) proposed a fully bayesian approach that broadly consists in factoring the joint distribution in two terms: a joint model for the covariates of the analysis model and a conditional model for the outcome given the covariates, that usually corresponds with the substantive analysis model. Although it was proposed as a fully bayesian method, it can be used as a multiple imputation approach compatible with the substantive model. The advantage of this is that it allows auxiliary variables to be included.

The aim of this paper is to introduce substantive model compatible JM-MI, and to compare it with standard JM-MI, when the substantive analysis model includes a random slope, an interaction or a quadratic term. We illustrate the advantage of the newly proposed method by simulations.

2 Methods

Assume we have a partially observed dataset with individuals i nested in clusters j. We intended to collect data on three continuous variables Y, X_1 and X_2, but we end up with some missing data in each of the three variables. The substantive analysis model of scientific interest is a linear mixed model:

$$y_{i,j} = (\beta_0 + u_{0,j}) + (\beta_1 + u_{1,j})x_{1,i,j} + \beta_2 x_{2,i,j} + \varepsilon_{i,j}$$
$$\begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix} \sim N(\mathbf{0}, \Sigma_u) \qquad \varepsilon_{i,j} \sim N(0, \sigma_e^2) \tag{1}$$

In order to deal with missing data, we can use JM-MI. But what imputation model should we use?

2.1 *JM-Hom: Homoscedastic Joint Modelling Imputation*

One possibility is to assume a 3-variate normal joint model for the three variables:

$$\begin{cases} y_{i,j} = \alpha_0 + v_{0,j} + e_{0,i,j} \\ x_{1,i,j} = \alpha_1 + v_{1,j} + e_{1,i,j} \\ x_{2,i,j} = \alpha_2 + v_{2,j} + e_{2,i,j} \end{cases}$$
$$\begin{pmatrix} v_{0,j} \\ v_{1,j} \\ v_{2,j} \end{pmatrix} \sim N(\mathbf{0}, \Omega_u) \qquad \begin{pmatrix} e_{0,i,j} \\ e_{1,i,j} \\ e_{2,i,j} \end{pmatrix} \sim N(\mathbf{0}, \Omega_e) \tag{2}$$

This model can be easily fitted with a standard Gibbs sampler, creating K different imputed datasets. These are then analysed with (1) to obtain K copies of the parameter estimates that are finally combined with Rubin's rules.

This approach naturally extends to include binary or categorical variables. This is achieved by means of a latent normal variables approach, as outlined in [5].

2.2 JM-Het: Heteroscedastic Joint Modelling Imputation

Because of the presence of a random slope, Model (1) is not compatible with Model (2), i.e. the conditional distribution of Y given X_1 and X_2 derived from (2) is not (1). To overcome this issue, one possibility is to assume instead an heteroscedastic imputation model, similar to (2) but with random cluster-specific covariance matrices following an inverse Wishart distribution:

$$\Omega_{e,j} \sim IW(a,A) \tag{3}$$

This model makes an attempt at modelling cluster-specific associations between variables, by assuming cluster-specific covariance matrices at level 1. However, it is still not a fully compatible approach.

It can be fitted with a similar Gibbs sampler to the one used for model (2).

2.3 JM-SMC: Substantive Model Compatible Joint Modelling Imputation

In order to define an imputation model fully compatible with (1), following along the lines of [4], we can factorise the joint distribution of the three variables in two terms: (i) a joint model for the two covariates and (ii) a conditional model for the outcome given the covariates. This way, we can make sure that the conditional model for the outcome corresponds to (1):

$$\begin{gathered}
\begin{cases} x_{1,i,j} = \alpha_1 + v_{1,j} + e_{1,i,j} \\ x_{2,i,j} = \alpha_2 + v_{2,j} + e_{2,i,j} \end{cases} \\
\begin{pmatrix} v_{1,j} \\ v_{2,j} \end{pmatrix} \sim N(\mathbf{0}, \Omega_u) \qquad \begin{pmatrix} e_{1,i,j} \\ e_{2,i,j} \end{pmatrix} \sim N(\mathbf{0}, \Omega_e) \\
y_{i,j} = (\beta_0 + u_{0,j}) + (\beta_1 + u_{1,j})x_{1,i,j} + \beta_2 x_{2,i,j} + \varepsilon_{i,j} \\
\begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix} \sim N(\mathbf{0}, \Sigma_u) \qquad \varepsilon_{i,j} \sim N(0, \sigma_e^2)
\end{gathered} \tag{4}$$

In order to impute from this model, an additional Metropolis-Hastings step within the Gibbs sampler is needed to impute the missing X_1 and X_2 values: imputations are drawn from a proposal distribution and accepted or rejected depending on the value of the Metropolis ratio. If using a symmetrical proposal distribution, this is simply the ratio of the likelihood of the model with the new proposed imputed value over the likelihood of the model with the previous imputed value.

This method naturally extends to allow for interactions or non-linearities in the conditional model for the outcome given the covariates in (4). Hence, it is possible to impute compatibly with the analysis model at the only cost of having to know the functional form of the substantive model in advance.

2.4 Software

We fit and impute from all three models (2), (3) and (4) using functions jomo and jomo.smc from our R package jomo, freely available on CRAN. The substantive model (1) is fitted with the R package lme4, and the results are combined with Rubin's rules as implemented in the mitml package.

3 Simulations

To illustrate the improvements that random coefficient compatible multiple imputation brings, in a base-case scenario we generate 1000 multilevel datasets, each constituted of 6000 observations, equally divided in 60 clusters, on three variables Y, X_1 and X_2. X_1 and a latent normal Z are generated from a bivariate normal distribution. X_2 is then created as a binary variable that takes the value 1 when $Z > 0$. The data-generating mechanism for the outcome is the following:

$$y_{i,j} = (0.5 + u_{0,j}) + (1 + u_{1,j})x_{1,i,j} - 0.3x_{2,i,j} + \varepsilon_{i,j}$$
$$\begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \qquad \varepsilon_{i,j} \sim N(0,1)$$

We assume that the desired analysis model is (1). We fit this model on the fully observed data (FD) and store all the parameter estimates. Then, for the fixed effect parameters β_0, β_1 and β_2 we calculate the mean, the empirical and model based standard errors and the coverage level across 1000 simulations. We additionally report the mean of the three variance components: the random intercept variance σ^2_{u0}, the random slope variance σ^2_{u1} and the residual variance σ^2_e.

We then make around 35% of the data on X_1 and X_2 missing at random conditional on the outcome Y, and we re-analyse the data using the complete records (CR). Finally, we handle missing data with the three different MI strategies presented in the previous section.

We investigate four additional scenarios:

- A scenario where X_2 is 3-level categorical;
- One with an additional continuous variable X_3, highly correlated with X_1, but not included either in the data generating process for Y or in the substantive analysis model (i.e. an auxiliary variable);
- One where Y is generated from a model with a quadratic effect on X_1;
- One where there is an interaction between X_1 and X_2.

3.1 Results

Table 1 shows the base-case simulation results. While Complete Records (CR) estimates are strongly biased, because of the dependence of the missingness mechanism from the outcome Y, all imputation methods are preferable; however, JM-Hom also leads to biased estimates and marked undercoverage for most parameters. Inference on the fixed effect parameters after imputation with JM-Het is affected by smaller biases, and leads to good coverage levels. However, bias is larger in the estimation of the variance components. Finally, JM-SMC leads to unbiased parameter estimates and good coverage levels.

Table 1 Base-case scenario: mean and coverage level of fixed effect parameter estimates and mean of variance component estimates over 1000 simulations. We compare Full Data (FD), Complete records (CR), JM imputation with a homoscedastic (JM-Hom) or heteroscedastic (JM-Het) imputation model and substantive model compatible JM-MI (JM-SMC).

	β_0		β_1		β_2		σ^2_{u0}	σ^2_{u1}	σ^2_e
Method	Mean	Cov	Mean	Cov	Mean	Cov	Mean	Mean	Mean
True value	0.50	0.95	1.00	0.95	-0.30	0.95	1.00	1.00	1.00
FD	0.49	0.94	1.00	0.93	-0.30	0.96	0.99	1.01	1.00
CR	1.08	0.00	0.89	0.82	-0.25	0.72	0.57	0.79	0.92
JM-Hom	0.40	0.88	1.07	0.85	-0.28	0.91	1.12	0.54	1.25
JM-Het	0.44	0.92	1.04	0.92	-0.29	0.94	1.09	0.93	1.05
JM-SMC	0.49	0.94	1.00	0.93	-0.30	0.95	0.99	1.01	1.00

Figure 1 pools the results across all the five simulation scenarios into a single panel. This is in terms of relative bias and coverage level for all the fixed effect parameter estimates. While JM-SMC always leads to negligible bias and coverage very close to the nominal level, JM-Hom and JM-Het are prone to bias in the estimation of some parameters in all scenarios. This is particularly serious for the two scenarios with an interaction and a quadratic effect. CR estimates are again always the most seriously biased, because of the missing data mechanism.

Finally, Figure 2 compares the level-2 variance component estimates from the three MI methods. Once again, JM-SMC is the only method leading to unbiased parameter estimates, while JM-Hom is the worst imputation method. JM-Het consistently overestimates the random intercept variance and gives biased estimates of the random slope variance.

4 Conclusions

We have investigated the behaviour of a new substantive model compatible MI strategy to deal with missing data in a multilevel dataset, and compared it with two existing multilevel imputation strategies. In particular, we have showed that when the analysis model of scientific interest includes a random slope, an interaction or a non-linearity, our proposed new method is the only one able to take this into account during the imputation, leading to correct inference.

The only additional price to pay for using this method, is that the precise functional form of the substantive model needs to be known in advance of the imputation process. Further research will investigate what is the best approach to take when model selection has to be performed along the imputation. Future work will also explore ways to impute level-2, i.e. cluster-level, variables within this framework.

All the imputation models presented in this paper can be fitted with functions jomo and jomo.smc in the R package jomo. This allows for binary and survival outcomes as well.

In conclusion, while standard JM-MI remains a valuable, and more flexible, method for the imputation of simple multilevel dataset, substantive model compatible JM-MI is preferable in presence of partially observed covariates with a random slope, an interaction or a non-linear term.

Acknowledgements This work was supported by the Medical Research Council, grant numbers MC UU 12023/21 and MC UU 12023/29.

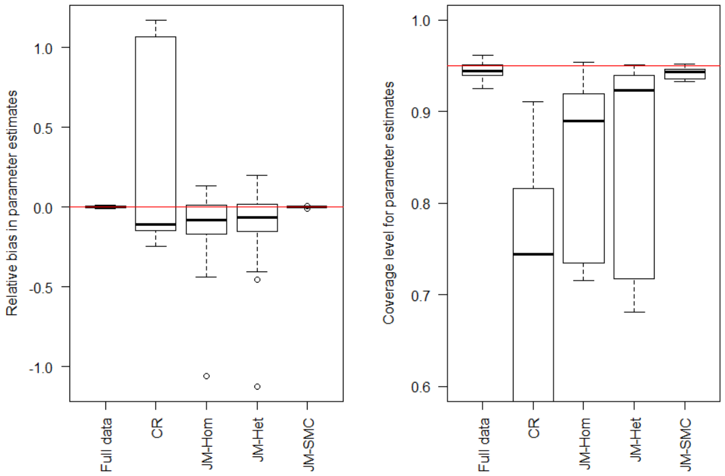


Fig. 1 Boxplots summarising results of five simulation scenarios. We compare relative bias (left panel) and coverage level (right panel) of fixed effect parameter estimates. The red lines indicate 0% relative bias and 95% coverage level. We compare Full Data (FD), Complete records (CR) and the three MI strategies.

References

1. Audigier V., White I.R., Debray T., Jolani S., Quartagno M., Carpenter J.R., van Buuren S., Resche-Rigon M. Stat Science. In press (2018)
2. Bartlett J. W., Seaman S.R., White I.R., Carpenter J.R., for the Alzheimers Disease Neuroimaging Initiative*. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Kenward MG, ed. Statistical Methods in Medical Research. 24(4):462-487 (2015)
3. Carpenter, J. R., Kenward M. G.: Multiple Imputation and its Application. Wiley, Chichester (2013)
4. Goldstein H., Carpenter J.R., Browne W.J., Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. Journal of RSS Series A. 177(2), 553-564 (2014)
5. Goldstein H., Carpenter J.R., Kenward M.G., and Levin K.A. Multilevel models with multivariate mixed response types. Statistical Modelling. 9(3), 173 – 197 (2009).
6. Lüdtke, O., Robitzsch, A. and Grund, S., Multiple imputation of missing data in multilevel designs: a comparison of different strategies, Psychological Methods. 22(1): 141–165 (2017)
7. Quartagno M., Carpenter J.R., Multiple Imputation of IPD Meta-analysis: allowing for heterogeneity and studies with missing covariates. Statistics in Medicine. 35(17), 2938–54 (2016)
8. Rubin, D.: Multiple Imputation for non-response in Surveys - a phenomenological bayesian approach to nonresponse. Wiley, New York (1987)
9. Schafer, J. L., and R. M. Yucel. Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values. Journal of Computational and Graphical Statistics. 11(2) , 437–57 (2002).
10. Yucel, R. M. Random-Covariances and Mixed-Effects Models for Imputing Multivariate Multilevel Continuous Data. Statistical modelling. 11(4), 351-370 (2011)

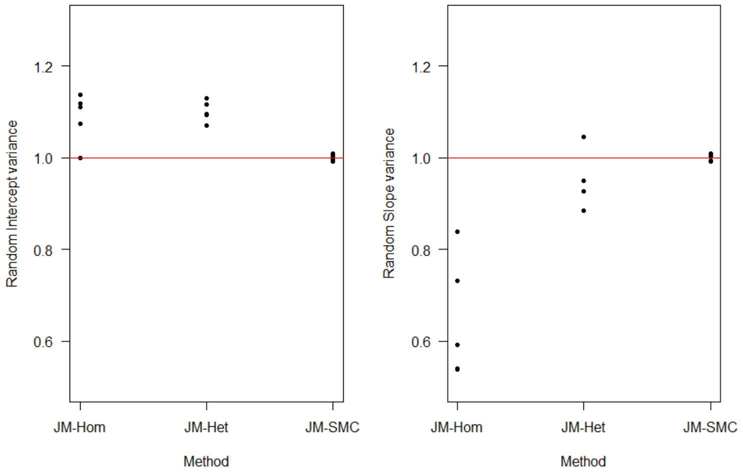


Fig. 2 Comparison of random intercept (left panel) and slope (right panel) variance estimates with the three MI strategies. Each point represents a different scenario. The red line indicates the correct value of 1.

Monitoring Education Systems. Insights from Large Scale Assessment Surveys

Educational Achievement of Immigrant Students. A Cross-National Comparison Over-Time Using OECD-PISA Data

I risultati scolastici degli studenti immigrati. Un confronto fra nazioni attraverso i dati PISA

Mariano Porcu

Abstract According to the Organization for Economic Cooperation and Development (OECD) a substantial performance differential between students with immigrant background and natives is observed in most countries. On average, immigrants tend to underperform their native peers even after their socio-economic conditions are controlled. In this work we study, in a time span perspective, the gap in school performances between native and immigrant students in five different European countries. Two of them are considered as new destination countries, namely Italy and Spain; the others three are traditional immigration countries (albeit with different migration history): France, United Kingdom, and Germany. We analyze data collected for the OECD Program for International Student Assessment (PISA) surveys of 2009, 2012, and 2015 by fitting in a multilevel setting multiple regressions to simultaneously model students' performances in reading and mathematics. We control for gender, socio-economic background, and immigration status (1st generation or 2nd generation immigrants). Results display that the performance gap between immigrant and native students is narrower in mathematics and that it is far from being bridged over-time. No substantial differences in trend are observed differentiating destination countries as new or traditional.
Abstract *Secondo l'Organizzazione per la cooperazione e lo sviluppo economico (OCSE), nella maggior parte dei paesi si osserva una sostanziale differenza di rendimento scolastico tra studenti con background migratorio e nativi. In media, gli immigrati tendono ad avere performance inferiori a quelle dei loro pari nativi, anche dopo aver controllato rispetto alle condizioni socio-economiche. In questo lavoro, in una prospettiva temporale viene studiato il divario nelle prestazioni scolastiche tra studenti nativi e immigrati in cinque diversi paesi europei. Due di questi sono considerati come paesi di nuova immigrazione, ovvero Italia e Spagna; gli altri tre sono paesi di immigrazione tradizionale (anche se con una storia migratoria diversa): Francia, Regno Unito e Germania. Verranno analizzati i dati raccolti*

Mariano Porcu
Università degli Studi di Cagliari, Dipartimento di Scienze Sociali e delle Istituzioni, e-mail: mrporcu@unica.it

nell'indagine PISA del 2009, 2012 e 2015 adattando alle performance in lettura e matematica dei modelli di regressione multilivello controllando rispetto al genere, al contesto socio-economico e allo status migratorio (immigrati di prima generazione o di seconda generazione). I risultati mostrano che il divario di prestazioni tra studenti immigrati e nativi è meno pronunciato in matematica, non si riduce nel tempo e i trend non si differenziano tra i paesi di nuova immigrazione e gli altri.

Key words: Immigration, Educational inequalities, PISA, Multilevel models

1 Introduction

Every year, the hope for a better life or the escape from wars or conditions of economic hardship push millions of people to cross the boundaries between the nations. This has happened since there were boundaries to cross and, perhaps, we can say that the drive to emigrate is itself a characteristic of the human kind.

However, in the last thirty years, the phenomenon has taken on impressive dimensions and is probably the most present issue at stake in the global political agendas. Modern means of transport, the globalization of economies, and the aging of Western populations will make this issue even more pressing in the coming years. So, the integration of immigrants in the hosting countries is crucial both for the economic systems and for a long-term growth of the social welfare.

Indeed, the best way to evaluate how well the immigrants are integrated into a society is certainly not that to compare their performance in the labor market (or in the economy in general) with those of the natives. Reasons for justifying a performance gap between native and immigrants workers are clear: difficulty in using the language of the host country, qualifications or work experience that are not recognized or exploitable are the most important reasons to explain the observed gaps. These motivations, however, should not apply to school performance of their children and the success of immigrant integration policies will be increasingly mirrored by the school performances of no-native students.

According to the Organisation for Economic Cooperation and Development (OECD), on average, in the last two decades, the percentage of 15-year-old students with a migratory background has increased by more than 2 points starting from 2000 OECD [2015, 2012a]. This is a very important evidence for education policies because at the same time a substantial performance differential between students with immigrant background and natives is also observed. On average, immigrants tend to underperform their native peers even after the socio-economic conditions are controlled.

Assessing what are the causes of the observed gaps is very difficult because the social groups of immigrant students are very heterogeneous, educational systems differ among countries, and there are different ways in which resources are distributed and educational policies are defined. Therefore, the contexts in which the immigrant students learn are different both from the historical-political point of view and looking at the governance of the school-system.

In this paper, we will try to assess in a time span perspective the gap in school performances between native and immigrant students in five different European countries. Two of which are considered as new destination countries, namely Italy and Spain. Both have been, for most of the twentieth century, emigration nations and only after the fall of the Iron Curtain they became destinations for migrants from South America, Eastern Europe, Middle East, and Africa. The others three countries have a consolidated albeit different migration history having long been a destination for important migratory flows. On one side, France and United Kingdom experimented incoming flows of migrants that reflected the colonial history of the countries. On the other, Germany rampant industrialization process, has long acted as a magnet for foreign workers since the beginning of the second half of twentieth century. This aim has been pursued by adopting two multilevel regression models which consider students' test scores as Level-1 units and schools as Level-2 units. Differences in student test scores have been analised taking into account a wide range of students and schools socio-economic and cultural characteristics and introducing interaction terms between immigrant status (native, 1st generation or 2nd generation immigrant), country (France, United Kingdom, Germany, Italy and Spain) and waves (2009, 2012, 2015). The approach allowed to capture the effect of having an immigrant background in the five countries across waves.

2 Data

Since the year 2000, the Organisation for Economic Cooperation and Development (OECD) carries on its Program for International Student Assessment (PISA). It is administered every three years to provide comparisons of students' achievement among the participating countries. In this analysis data collected in three rounds of the PISA survey have been considered, say 2015, 2012, and 2009. PISA surveys could be considered as the most comprehensive and accurate international assessment of students's skills in reading, mathematics, and sciences. In addition PISA assesses not only students' competences, but also collects information on their socio-demographic background and on the school context in which their are enrolled. In each round PISA carries on a detailed assessment of each of the three subjects and the 2009 survey marks the return to a focus on reading so that our analysis considers three different subject focuses (nonetheless, in each round the three subjects are, however, considered).

The PISA target population is that of students aged between 15 and 16 years at the time of the survey and who have completed a minimum of 6 years of formal education regardless of the type of institution where they are enrolled. The age of 15-16 represents, for many countries, the transition time from a basic education to a more advanced one. Detailed information on PISA sampling design and procedures are available in a collection of thematic and technical reports at PISA-OECD website [OECD, 2012b, 2014, 2017].

We consider as dependent variables the student's performance in reading and mathematics tests. OECD defines reading literacy as the ability in "[...] understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society ..." [OECD, 2012b]. Math literacy is "[...] the extent to which students can use their mathematical knowledge and skills to solve various kinds of numerical and spatial challenges and problems [...]" [OECD, 2017].

In PISA surveys, in order to minimise the assessment burden on each student and to avoid that the scaling of skills would be influenced by the "booklet effect" each student is asked to handle only a part of the whole test in the three domains assessed (reading, maths, science) following a systematic booklet assembly and rotation procedure. For that reason rather than one single measure of achievement, the PISA databases provides 5 plausible values (PV) of student's score in each topic. The use of PV allows to to take into account the uncertainty associated with the estimate of a measure of achievement for each student by reproducing the likely distribution of students' competencies in each topic [Monseur and Adams, 2009, OECD, 2017].

At student level we considered (for each wave) the following information:

- Country of residence (`COUNTRY`): France (FRA), Germany (DEU), Italy (ITA), Spain (ESP), United Kingdom (GBR).
- Immigrant status (`IMMSTAT`): according to OECD-PISA classification we differentiate between immigrant and non-immigrant students based on the information on the country of birth of both their parents; if both parents were born in a country different than the country where the student take on the test, then the student is classified as immigrant. Non-immigrant or natives (`IMMSTAT`$=0$) are the remainder. Among immigrants we distinguish between second-generation (`IMMSTAT`$=1$) and first-generation students (`IMMSTAT`$=2$). Second-generation are immigrant students born in the country of PISA assessment; first-generation students are foreign-born alike their parents.
- Language spoken at home (`LHOMEDIFF`): an additional relevant difference among immigrant students is the language they speak at home. We distinguish between immigrants who speak at home a foreign language (`LHOMEDIF`$=1$) (i.e. different from the PISA assessment language; dialects or regional languages are considered as test language).
- Gender (`GENDER`$=0$: female).
- Parental educational level (`PARED`): highest parental education in years of schooling.
- Parental occupational status (`HISEI`): highest parental occupational status. In PISA surveys, occupational data for both parents are obtained from responses to open-ended questions. Responses are then coded to four-digit ISCO codes (International Standard Classification of Occupations) and mapped to the international socio-economic index of occupational status (ISEI) Ganzeboom and Treiman [2003]. Higher `HISEI` scores indicate higher occupational status.

- Family possession of culture related items (CULTPOSS): the PISA index of family cultural possession is derived from what the students report on the availability of specific household items at home such as classic literature, books of poetry, works of art, musical instruments, etc. Highest values indicate an higher family endowment of culture related items.
- Family possession of educational resources (HEDRES): the PISA index of home educational resources (a desk to study at, a computer, educational software, a dictionary, etc.). Highest values indicate higher availability of educational resources at home.
- The percentage of non native students in the school (SCHNONATIVE).

In PISA databases, the HISEI index is calculated using Principal Component Analysis (PCA). Except for the parental education variable (PARED), the remaining variables as well as the PV are calculated by using a model-based scaling procedures belonging to the family of Item Response Theory (IRT) applied to dichotomous or Likert-type responses to questionnaire items OECD [2017].

3 First findings

In this framework we discuss only the main findings we have observed by applying two multilevel regression models for assessing trends in the divergences in mathematics and reading across the five countries in the three years between students with different immigrant background. Results observed for mathematics and reading are listed in Tables 1 and 2. In the following we rapidly focus the attention only on the estimated effect of the country × year × immigrant status combination. Looking at the effect jointly exerted overtime by the combination of country and immigrant status we considered, both in reading and mathematics, the achievement of a German native in 2009 as the baseline. Figure 1 displays the caterpillar plot of the estimated 44 parameters, each one with its associated confidence interval limits. In total 44 parameters are displayed for reading and 44 for mathematics: $44 = [(\text{countries} \times \text{immigrant status} \times \text{wave}) - 1] = [(5 \times 3 \times 3) - 1]$. Parameters are displayed in ascending order of magnitude. Looking at reading competencies, we can spot that the highest ranks are hold by native or second generation immigrants of Germany, France, and Great Britain (although the only significant difference for these outperforming students are those of natives from Germany in 2015 and from France in 2012). Whenever we consider trends of performances according to immigrant status within each countries, situations are noteworthy differentiated. Figures 2 groups the estimated parameters by immigration status for reading. The caterpillar plots clearly show divergences across countries according to the immigrant backgrounds. In Germany we can observe that performances in reading of natives increase from 2009 to 2015; the same occurs for second generation immigrants while performances of first generation steadily underperform.

Table 1 Reading. Model parameter estimates

Variable	Beta	se	z-score	pvalue	95% CI	
					lw	up
Intercept	494.898	3.574	138.472	0.000	487.893	501.903
PARED	0.370	0.062	5.981	0.000	0.249	0.492
HISEI	0.506	0.010	50.887	0.000	0.486	0.525
CULTPOSS	10.436	0.191	54.639	0.000	10.062	10.811
HEDRES	5.258	0.191	27.550	0.000	4.884	5.632
LHOMEDIF = Yes	-13.706	0.929	-14.751	0.000	15.527	-11.885
SEX = M	-22.740	0.330	-68.895	0.000	23.387	-22.093
SCHNONATIVE	-72.754	4.123	-17.647	0.000	-80.835	-64.674
DEU.NAT.09 (*bas.*)	—	—	—	—	—	—
DEU.NAT.12	8.829	4.819	1.832	0.067	-0.615	18.274
DEU.NAT.15	16.241	4.665	3.481	0.000	7.097	25.385
⋮*	⋮	⋮	⋮	⋮	⋮	⋮
Random-effect Parameters			Estimate	se	95% CI	
Between Schools Std.Dev.			47.631	0.438	46.782	48.495
Residual Std.Dev.			63.912	0.110	63.696	64.129

*The estimates for the other countries and the related 95% CI have been plotted in the Figure 1.

Table 2 Mathematics. Model parameter estimates

Variable	Beta	se	z-score	pvalue	95% CI	
					lw	up
Intercept	486.682	3.487	139.557	0.000	479.847	493.517
PARED	0.301	0.061	4.908	0.000	0.181	0.422
HISEI	0.542	0.010	54.967	0.000	0.523	0.561
CULTPOSS	8.799	0.189	46.452	0.000	8.428	9.170
HEDRES	6.292	0.189	33.241	0.000	5.921	6.663
LHOMEDIF = Yes	-7.676	0.921	-8.330	0.000	9.482	-5.870
SEX = M	20.874	0.327	63.776	0.000	20.233	21.516
SCHNONATIVE	-76.029	4.024	-18.896	0.000	83.915	-68.143
DEU.NAT.09 (*bas.*)	—	—	—	—	—	—
DEU.NAT.12	0.125	4.698	0.027	0.979	-9.082	9.332
DEU.NAT.15	-3.374	4.547	-0.742	0.458	-12.287	5.539
⋮*	⋮	⋮	⋮	⋮	⋮	⋮
Random-effect Parameters			Estimate	se	95% CI	
Between Schools Std.Dev.			46.312	0.425	45.488	47.152
Residual Std.Dev.			63.393	0.109	63.178	63.608

*The estimates for the other countries and the related 95% CI have been plotted in Figure 1.

4 Conclusion

On the light of the main findings we can state that in general in the traditional immigration countries we see that first generation immigrants steadily lag behind. Nonetheless, we observe similar performances for natives and second generation immigrants: the upper limits of the confidence intervals for second generation overlap with the bottom limits of the confidence intervals estimated for natives (albeit in France differences are sensibly wider).

In Italy and Spain differences in performances according to immigrant status are wider; immigrants underperform and no trends of improvement turn up overtime. Nonetheless, we shall note that in Spain reading performances of second generation immigrants are similar to those of natives in 2015 (likely due to the fact that a large part of the immigrant population in Spain comes from Latin America). In mathematics performances of natives and immigrants are in somewhat less noticeable.

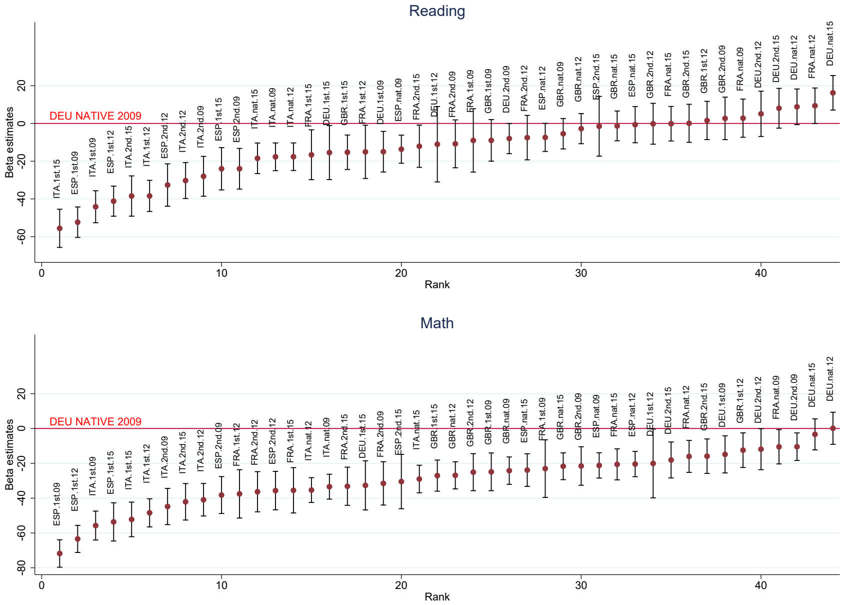


Fig. 1 Caterpillar plot of estimated parameters for immigrant status-country-year effect. Baseline: German-Native 2009

References

OECD. Can the performance gap between immigrant and non-immigrant students be closed? PISA in Focus 53, OECD, Paris, 07 2015.

OECD. *Untapped Skills: Realising the Potential of Immigrant Students*. OECD, Paris, 2012a.

OECD. Pisa 2009 Technical Report. Technical report, OECD, Paris, 2012b.

OECD. Pisa 2012 Technical Report. Technical report, OECD, Paris, 2014.

OECD. Pisa 2015 Technical Report. Technical report, OECD, Paris, 2017.

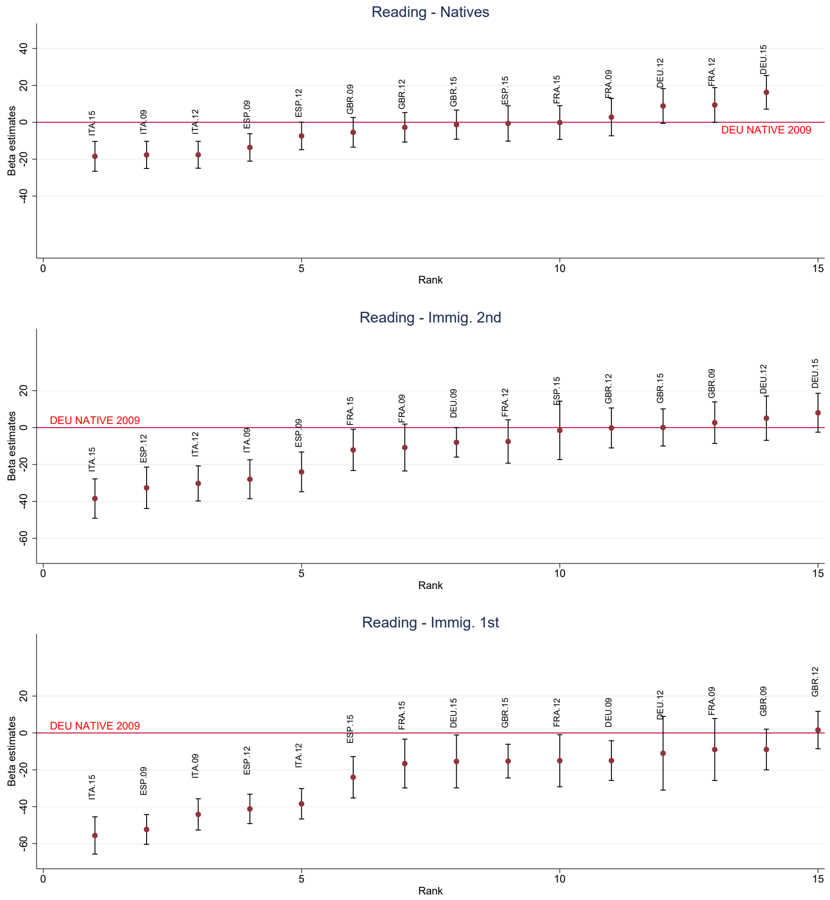


Fig. 2 Caterpillar plot of estimated parameters for country-year effect by immigrant status. Baseline: German-Native 2009 – Reading

Christian Monseur and Raymond Adams. Plausible values: how to deal with their limitations. *Journal of Applied Measurement*, 10(3):320–334, 2009.

Harry B.G. Ganzeboom and Donald J. Treiman. Three Internationally Standardised Measures for Comparative Research on Occupational Status. In *Advances in Cross-National Comparison*. Springer, Boston, MA, 2003.

New Perspectives in Time Series Analysis

Generalized periodic autoregressive models for trend and seasonality varying time series

Francesco Battaglia and Domenico Cucina and Manuel Rizzo

Abstract Many nonstationary time series exhibit changes in the trend and seasonality structure, that may be modeled by splitting the time axis into different regimes. We propose multi-regime models where, inside each regime, the trend is linear and seasonality is explained by a Periodic Autoregressive model. In addition, for achieving parsimony, we allow season grouping, i.e. seasons may consist of one, two, or more consecutive observations. Identification is obtained by means of a Genetic Algorithm that minimizes an identification criterion.

Key words: Genetic algorithms, Structural break, Regime change

1 Introduction

Many seasonal time series exhibit an autocorrelation structure which depends not only on the time between observations but also on the season of the year. Moreover, the time series of observations within a given season is usually second order stationary (Hipel and McLeod, 1994). In order to model appropriately these and similar types of time series, Periodic AutoRegressive models (PAR) can be employed. When fitting a PAR model to periodic time series a separate AR model for each season of the year is estimated. These models are appropriate for describing time series drawn from different areas such as economics, hydrology, climatology and

Francesco Battaglia
Department of Statistical Sciences, University La Sapienza, Rome, Italy e-mail: francesco.battaglia@uniroma1.it

Domenico Cucina
Department of Economics and Statistics, University of Salerno, Italy e-mail: dcucina@unisa.it

Manuel Rizzo
Department of Statistical Sciences, University La Sapienza, Rome, Italy e-mail: manuel.rizzo@uniroma1.it

signal processing (e. g. Franses and Paap, 2004; Hipel and McLeod, 1994; Ursu and Turkman, 2012).

In this study we consider a generalization of PAR models with linear trend in two directions. First, the model may follow different regimes in time, and regime changes may occur at any time. The regime changes may affect the linear trend, the seasonal means and the autoregressive parameters. We also allow a discontinuous trend which can identify changes in level. Second, inside each regime the model structure may be different for each seasonal position (e.g. months) or vary more slowly, changing only according to grouped seasons like quarters or semesters.

The number of regimes and change times (or break points) are assumed to be unknown. The problem of their identification can be treated as a statistical model selection problem according to a specified identification criterion (Aue and Horväth, 2013). This approach has been used for identification of structural breaks e.g. in Davis et al (2008) and Lu et al (2010). In these works Genetic Algorithms (GAs) are proposed to solve the selection problem.

To the best of our knowledge, there are no articles that handle the changing parameters and changing trend problem in PAR models simultaneously. We propose a class of GAs to detect the number of regimes of piecewise PAR models and their locations. Our procedure evaluates several regime patterns where the locations that are possibly change times are simultaneously considered. In this way, GAs deal efficiently with the detection of multiple change times. We also allow subset AR models to be selected. Each piecewise subset PAR configuration is evaluated by an AIC identification criterion.

In our paper, since the seasonal effect on means, variances and correlations may show different speed and pattern, we propose to join appropriately season parameters into groups for each of these three features.

The piecewise linear nature of our model makes forecasting very simple.

2 The Model

Suppose that a time series $\{X_t\}$ of N observations is available. The seasonal period of the series is s and is assumed to be known. Assume that there are $m+1$ different regimes, separated by m change times τ_j so that the first regime contains observations from time 1 to $\tau_1 - 1$, the second regime contains data from time τ_1 to $\tau_2 - 1$, the $(j+1)$-th regime contains data from τ_j to $\tau_{j+1} - 1$, and the last regime data from τ_m to N. To ensure reasonable estimates we assume that the minimum regime length is a fixed constant mrl, thus any regime assignment is defined by the set $\{\tau_j, j = 1, \ldots, m\}$ subject to $mrl < \tau_1 < \tau_2 < \ldots < \tau_m < N - mrl$, $\tau_j \geq \tau_{j-1} + mrl, j = 2, \ldots, m$.

The parameters of the model for regime j will be denoted by a superscript (j).

The seasonal effect on means, variances and correlations may show different speed and pattern, thus it seems advisable, for each of these features, to use a different splitting of the year, determined by a different length of the season inside which that feature remains constant. For example, if the seasonal period is s, we may have exactly s different models, one for each seasonal position; or rather only s/c different structures, when c consecutive observations are supposed to belong to the same season. E. g. if $s = 12$ (monthly data) and $c = 3$, the same model works for each quarter, therefore there are only $s/c = 4$ different seasons. This may be useful when the seasonal variation is slow and more detailed models would be redundant.

We allow a different season grouping for means, correlation and variance. We denote by c_M the number of consecutive observations for which the mean remains constant (c_M divides s), and by $ss = s/c_M$ the number of seasons. In an analogue fashion, we denote by c_{AR} the number of consecutive observations for which the AR parameters remain constant, and $sv = s/c_{AR}$ the related number of seasons. E. g. for $s = 12$, if $c_{AR} = 1$, each month has a different set of AR parameters, then the variances of the 12 months may a priori be different. If on the contrary $c_{AR} > 1$, the variances of c_{AR} contiguous observation all are proportional, through the same coefficient, to the residual variances, thus a variance instability is equivalent to a residual variance instability. Therefore, we allow the possibility that, inside each single season for the AR model (containing c_{AR} consecutive observations) the residual variances may change. Thus we must consider sub-seasons composed by c_V observations, where c_V divides c_{AR}, and allow the residual variance to change every c_V observations, in a total number of seasons (concerning the residual variance) equal to $svar = s/c_V$.

A linear trend and a different mean for each season is assumed. The residuals are treated as zero mean and described by an autoregressive model with maximum order p, and parameters varying with seasons. Let k_t denote the season (for the mean) of the t-th observation ($1 \leq k_t \leq ss$) and k_t^* the season for AR structure of the t-th observation, denote by $a^{(j)} + b^{(j)}t$ the linear trend in regime j, by $\mu^{(j)}(k)$ the mean of season k in regime j, and by $\phi_k^{(j)}(i)$ the lag-i autoregressive parameter for the model in season k and regime j. Then for $\tau_{j-1} \leq t < \tau_j$:

$$X_t = a^{(j)} + b^{(j)}t + \mu^{(j)}(k_t) + W_t \ , \ W_t = \sum_{i=1}^{p} \phi_{k_t^*}^{(j)}(i) W_{t-i} + \varepsilon_t$$

where $\tau_0 = 1$ and $\tau_{m+1} = N + 1$.

The innovations ε are supposed independent and zero-mean, with variances $\sigma^2(j,k)$ possibly depending on the regime and season.

As far as subset selection is concerned, we introduce also $m+1$ binary vectors $\delta^1, \ldots, \delta^{m+1}$, which specify presence or absence of autoregressive parameters in each regime as follows: if $\delta^j[p(k_t^* - 1) + i] = 1$ then $\phi_{k_t^*}^{(j)}(i)$ is constrained to zero. In summary, a model is identified by the following:
External parameters (fixed and equal for all models) N, s, maximum order p, maximum number of regimes, and minimum number of observations per regime *mrl*

Structural parameters (determining the model structure)

m	number of change times
$\tau_1, \tau_2, \ldots, \tau_m$	change times or thresholds
$\delta^1, \ldots, \delta^{m+1}$	denote which ϕ's are zero in each regime and season
c_M, c_{AR}, c_V	season grouping parameters subject to constraints: c_M divides s; c_{AR} divides s; c_V divides c_{AR}

Regression parameters to be estimated by Least Squares (LS) or Maximum Likelihood (ML)

$a_1, a_2, \ldots, a_{m+1}$	intercepts
$b_1, b_2, \ldots, b_{m+1}$	slopes
$\mu^{(j)}(k)$	seasonal means, $k = 1, \ldots, ss$; $j = 1, \ldots, m+1$
$\phi_k^{(j)}(i)$	AR parameters, $k = 1, \ldots, sv$; $j = 1, \ldots, m+1$; $i = 1, \ldots, p$ (some of them may be constrained to zero)
$\sigma^2(j,k)$	innovation variances, regime j and season $k = 1, \ldots, svar$.

For estimating trend and seasonal means by LS, note that the intercept and the means are linearly dependent, therefore we assume that the seasonal means sum to zero on one cycle: $\mu^{(j)}(1) + \mu^{(j)}(2) + \ldots + \mu^{(j)}(ss) = 0, \forall j$. Therefore the following equations are estimated:

$$X_t = b^{(j)}t + c(j, k_t) \ , \ \ \tau_j \le t < \tau_{j+1} \tag{1}$$

and then the parameter vector is $\beta' = \{b^{(1)}, b^{(2)}, \ldots, b^{(m+1)}, c(1,1), c(1,2), \ldots, c(1,ss), c(2,1), \ldots, c(2,ss), \ldots, c(m+1,1), \ldots, c(m+1,ss)\}$ with dimension $(m+1) \times (ss+1)$ and the estimates are obtained by least squares. From the $\{\hat{c}(j,k)\}$, the intercepts $\hat{a}^{(j)}$ and seasonal means $\hat{\mu}^{(j)}(k)$ are recovered basing on the above assumption. It follows

$$\hat{a}^{(j)} = \frac{1}{ss} \sum_{k=1}^{ss} \hat{c}(j,k) \ \ , \ \ \hat{\mu}^{(j)}(k) = \hat{c}(j,k) - \hat{a}^{(j)}.$$

Moreover it is possible to prescribe trend continuity by imposing that, if the number of regimes is larger than one, the trend values of two consecutive regimes coincide on the first observation of the second regime. A possible level change at $t = \tau_{j+1}$ is estimated if the trend continuity is not imposed.

Conditioning on thresholds, seasonal arrangement and estimated trend and means, the residual series is computed as $\hat{W}_t = X_t - \hat{a}^{(j)} - \hat{b}^{(j)}t - \hat{\mu}^{(j)}(k_t)$.

For each regime and season a separate autoregressive process is considered:

$$\hat{W}_t = \sum_{i=1}^{p} \phi_{k_t^*}^{(j)}(i) \hat{W}_{t-i} + \varepsilon_t.$$

We denote by $I(j,k)$ the set of times belonging to regime j and season k. The corresponding observations $z_{j,k}$ are selected and the LS estimates of the parameters $\{\phi_k^{(j)}(i), i = 1, \ldots, p\}$ are obtained. As far as subset selection is concerned, the final estimates are obtained via LS constrained optimization, with constraints given by linear system $H\phi = 0$:

$$\hat{\phi} = \phi_{LS} - (Z'Z)^{-1}H'[H(Z'Z)^{-1}H']^{-1}H\phi_{LS},$$

where $\phi_{LS} = (Z'Z)^{-1}Z'z$ are the unconstrained least squares estimates, Z is the $n_{j,k} \times p$ design matrix including lagged observations and H is the constraints matrix that specifies subset models. The residuals $e = z - Z\hat{\phi}$ give the estimate of the innovations for regime j and season k $\{\varepsilon_t, t \in I(j,k)\}$, which allow to obtain $\hat{\sigma}^2(j,k)$. The structural parameters take discrete values and their combinations amount to a very large number. GAs are naturally suitable for the choice of optimal structural parameters.

3 Genetic algorithms

GAs, initially developed by Holland (1975), imitate the evolution process of biological systems, to optimize a given function. A GA uses a set of candidate solutions, called *population*, instead of one single current solution. In GA terminology, any candidate solution is encoded via a numerical vector called *chromosome*. The GA proceeds iteratively by updating the population in rounds, called generations. In each generation, some of the active chromosomes are selected (parents-chromosomes) to form the chromosomes of the next generation (children-chromosomes). The selection process is based on an evaluation measure called *fitness function*, linked to the objective function, that assigns to each chromosome a positive number. Children are formed by recombining (*crossover*) the genetic material of their two parents-chromosomes and perhaps after a random alteration of some of the genes (single digits of the chromosome), which is called *mutation* (see Holland, 1975; Goldberg, 1989, for a detailed description).

A successful implementation of GAs is certainly crucial to obtain satisfactory results. Before a GA can be applied to a problem some important decisions have to be made. The GA methods require a suitable encoding for the problem and an appropriate definition of objective function. In addition operators of selection, crossover and mutation have to be chosen.

Encoding. An appropriate encoding scheme is a key issue for GAs. It must guarantee an efficient coding producing no illegal chromosome and no redundancy. Details of the adopted method may be found in Battaglia et al (2018).

Fitness function. The most natural objective in building statistical models is to minimize an identification criterion such as AIC, BIC, ICOMP, MDL. They all are based on the estimated residual variance $\hat{\sigma}^2(j,k)$ and the total number of estimated parameters: there are $m+1$ parameters for trend, $(m+1) \times ss$ seasonal means, and in regime j there are $p \times sv - |\delta^j|$ autoregressive parameters (where $|x|^2 = \sum_i x_i^2$). So, the total number of estimated parameters is $P = (m+1)(ss+1) + (m+1)p \times sv - |\delta^1| - |\delta^2| - \ldots - |\delta^{m+1}|$.

If continuity constraints on trend are added, the number of parameters decreases by m.

The most obvious generalization of AIC is the NAIC criterion introduced by Tong (1990, p. 379) for threshold models:

$$NAIC = [\sum_j \sum_k AIC_{j,k}]/N = \left[\sum_j \sum_k n_{j,k} \log \hat{\sigma}^2(j,k) + \pi \times P\right]/N,$$

where $AIC_{j,k}$ is identification criterion for series of regime j and season k, $\sigma^2(j,k)$ is the related residual variance, P is the total number of parameters, π is the penalization term (equal to 2 in the original Akaike's proposal). Other alternatives are possible (see Battaglia et al, 2018).

Since the identification criteria are to be minimized, the fitness function is a monotonically decreasing function of the identification criterion. We adopted a negative exponential transformation.
GA operators. For selection we used the "roulette wheel" rule where the probability of a chromosome being selected as a parent is proportional to its fitness. Each selected couple of parents will produce two "children" by methods of crossover and mutation. We implemented uniform crossover — each child receives each gene from one parent or the other randomly with probability $1/2$.

The entire population of chromosomes is replaced by the offsprings created by the crossover and mutation processes at each generation except for the best chromosome, which survives to the next generation. This *elitist* strategy ensures that the fitness will never decrease through generations (Rudolph, 1994).

Our search strategy is in two steps: in the first one the GA tries to determine the best splitting in regimes for complete models, as the chromosome includes only $m, \tau_1, ..., \tau_m$; in the second step we exhaustively enumerate all possible seasons grouping (specified by c_M, c_{AR}, c_V) and subset models (examining $\delta^1, ..., \delta^{m+1}$). This strategy is hybrid, as it combines an exact method with an approximate method, and it is feasible if order p and seasonal period s are not too large.

4 Applications

We briefly summarize the application of our method to the CET series of monthly mean surface temperatures for a location in the Midlands region (see Proietti and Hillebrand, 2017, and references therein). It is a very long and frequently investigated series (we use 2904 observations for years 1772–2013). Many researchers suggest an upward trend since the beginning of the 20th century, due to global warming, and an evolution of seasonal pattern, identified as a precession of Earth's axis of rotation. Four PAR models were fitted to the time series: a *complete* model, with a different AR(2) model for each month and no constraint on the autoregressive parameters; a *subset* model similar to the previous one, but with some AR parameters constrained to zero in order to maximize fitness; a *grouped* subset model where the season are grouped; and finally a *constant* seasonality model, subset as well, where the autoregressive parameters remain equal in each regime. The series plot appears in Figure 1 (left panel).

Two regimes were identified, with change time at July, 1899 (the trend is drawn as a dotted line in the figure). This confirms with clear evidence the suggested trend change at the beginning of the last century. The grouped season model identified the optimal setting $c_M = c_{AR} = 1$, meaning that any grouping of seasons would not increase the fitness, thus the grouped model coincides with the subset model (with other more parsimonious criteria like *BIC* the best grouping results $c_M = 1, c_{AR} = 4, c_V = 2$). The results appear in Table 1: it may be concluded that many autoregressive parameters may be constrained to zero without a sensible loss of fit. Moreover, the smaller fitness of the constant seasons model indicates an evolution in the seasonal pattern; Figure 1 (right panel) reports the monthly means for the two regimes. More applications may be found in Battaglia et al (2018).

Table 1 Models fitted to the CET series

Model	Complete	Subset	Constant seasons
Residual variance	1.696	1.701	1.752
Number of parameters	74	50	30
Fitness	0.594	0.602	0.595

5 Conclusions

In this paper we have proposed models that are able to explain, on one side, regime changes and structural breaks, and on the other side a seasonal behavior that evolves in time.

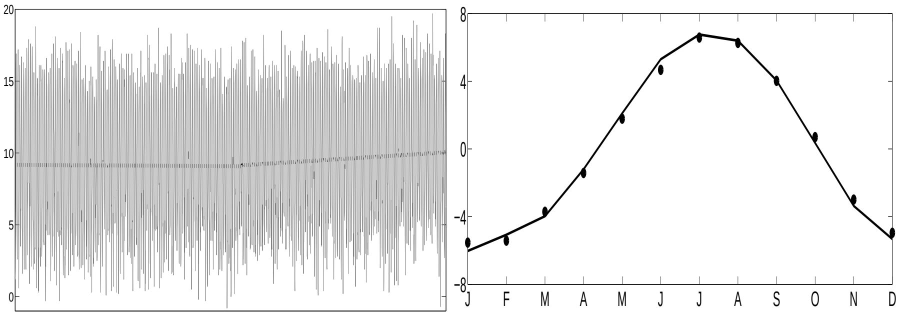


Fig. 1 Left panel: CET series and trend (dotted line). Right panel: monthly means, continuous line: first regime; dotted line: second regime

The complex problem of identifying and estimating such models is solved by GAs. The best model is selected according to a fitness function that is a monotonically decreasing transformation of widely used identification criteria. Experience on real and simulated data suggests that the choice of the fitness function is crucial because a too parsimonious criterion may lead to models that overlook important structure changes.

The results seem to support the usefulness of the proposed methods in detecting relevant changes in the structure of the trend and also possible evolution in the seasonal behavior concerning levels, variance and correlation. The generalized periodic autoregressive models allow a closer analysis of the seasonal behavior, suggesting also the most convenient grouping of seasons in terms of fitness.

References

Aue A, Horväth L (2013) Structural breaks in time series. Journal of Time Series Analysis 34(1):1–16

Battaglia F, Cucina D, Rizzo M (2018) A generalization of periodic autoregressive models for seasonal time series. Tech. Rep. 2, Dept. of Statistical Sciences, University La Sapienza, Rome, Italy, ISBN 2279-798X

Davis R, Lee T, Rodriguez-Yam G (2008) Break detection for a class of nonlinear time series models. Journal of Time Series Analysis 29(5):834–867

Franses PH, Paap R (2004) Periodic Time Series Models. Oxford University Press, New York

Goldberg D (1989) Genetic algorithms in search optimization and machine learning. Addison-Wesley, Reading, MA

Hipel KW, McLeod AI (1994) Time Series Modelling of Water Resources and Environmental Systems. Elsevier, Amsterdam

Holland J (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and AI. The University of Michigan, Ann Arbor, MI

Lu Q, Lund R, Lee T (2010) An MDL approach to the climate segmentation problem. The Annals of Applied Statistics 4(1):299–319

Proietti T, Hillebrand E (2017) Seasonal changes in central England temperatures. Journal of the Royal Statistical Society A 180:769–791

Rudolph G (1994) Convergence analysis of canonical genetic algorithms. IEEE Transactions on Neural Networks 5:96–101

Tong H (1990) Non Linear Time Series: A Dynamical System Approach. Oxford University Press, Oxford

Ursu E, Turkman KF (2012) Periodic autoregressive model identification using genetic algorithms. Journal of Time Series Analysis 33:398–405

Recent Advances in Model-based Clustering

Flexible clustering methods for high-dimensional data sets.

Metodi di cluster analysis flessibili per data set di grandi dimensioni

Cristina Tortora and Paul D. McNicholas

Abstract Finite mixture models assume that a population is a convex combination of densities; therefore, they are well suited for clustering applications. Each cluster is modeled using a density function. One of the most flexible distributions is the generalized hyperbolic distribution (GHD). It can handle skewness and heavy tails, and has many well-known distributions as special or limiting cases. The multiple scaled GHD (MSGHD) and the mixture of coalesced GHDs (CGHD) are even more flexible methods that can detect non-elliptical, and even non-convex, clusters. The drawback of high flexibility is a high parametrization — especially so for high-dimensional data because the number of parameters is depends on the number of variables. Therefore, the aforementioned methods are not well suited for high-dimensional data clustering. However, the eigen-decomposition of the component scale matrix can naturally be used for dimension reduction obtaining a transformation of the MSGHD and MCGHD that is better suited for high-dimensional data clustering.

Key words: Mixture models, generalized hyperbolic distribution, cluster analysis, high dimensional data

1 Background: Model-based clustering

Model-based clustering assumes that a population is a convex combination of a finite number of densities. A random vector $\mathbf{X}$ follows a (parametric) finite mixture distribution if, for all $\mathbf{x} \subset \mathbf{X}$, its density can be written

Cristina Tortora
San Jose State University, One Washington square, San Jose CA USA, e-mail: cristina.tortora@sjsu.edu

Paul D. McNicholas
McMaster University, 1280 Main St W, Hamilton, ON Canada e-mail: paul@math.mcmaster.ca

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g),$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, is the gth mixing proportion, $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the gth component density, and $\boldsymbol{\vartheta} = (\pi, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters, with $\pi = (\pi_1, \ldots, \pi_G)$. The component densities $f_1(\mathbf{x} \mid \boldsymbol{\theta}_1), \ldots, f_G(\mathbf{x} \mid \boldsymbol{\theta}_G)$ are usually taken to be of the same type. Over the past few years, non-Gaussian model-based clustering techniques have gained popularity. [3] proposed the use of the generalized hyperbolic distribution (GHD), which has the advantage of being extremely flexible because it is characterized by five parameters–the mean, the scale matrix, the skewness, the concentration and the index parameters. Many other distributions, e.g. the Gaussian or the the skew-t distribution, can be obtained as a special or limiting cases. The density of a random variable $\mathbf{X}$ from a generalized hyperbolic distribution is

$$f_{\mathrm{H}}(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\chi + \boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma})}{\psi + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right]^{\frac{\lambda - \frac{p}{2}}{2}} \frac{(\psi/\chi)^{\frac{\lambda}{2}} K_{\lambda - \frac{p}{2}}\left(\sqrt{[\psi + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\chi + \boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma})]}\right)}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} K_\lambda(\sqrt{\chi\psi}) \exp\{(\boldsymbol{\mu} - \mathbf{x})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}, \tag{1}$$

where $\boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$, K_λ is the modified Bessel function of the third kind with index λ, and $\boldsymbol{\vartheta}$ denotes the parameters. The parameters have the following interpretation: λ is an index parameter, χ and ψ are concentration parameters, $\boldsymbol{\alpha}$ is a skewness parameter, $\boldsymbol{\mu}$ is the mean, and $\boldsymbol{\Sigma}$ is the scale matrix.

Let $Y \sim \mathrm{GIG}(\psi, \chi, \lambda)$, where GIG indicates the generalized inverse Gaussian distribution [1], and the density is given by

$$h(y \mid \boldsymbol{\theta}_g) = \frac{(y/\eta)^{\lambda - 1}}{2\eta K_\lambda(\omega)} \exp\left\{-\frac{\omega}{2}\left(\frac{y}{\eta} + \frac{\eta}{y}\right)\right\}. \tag{2}$$

Consider Y and a random variable $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Then, a generalized hyperbolic random variable $\mathbf{X}$, see (1), can be generated via

$$\mathbf{X} = \boldsymbol{\mu} + Y\boldsymbol{\alpha} + \sqrt{Y}\mathbf{V}, \tag{3}$$

and it follows that $\mathbf{X} \mid Y \sim \mathcal{N}(\boldsymbol{\mu} + y\boldsymbol{\alpha}, y\boldsymbol{\Sigma})$.

Note that the parameterization used in (1) requires the constraint $|\boldsymbol{\Sigma}| = 1$ to ensure identifiability, but this constraint is not practical for clustering applications. Therefore, an alternative parameterization, setting $\omega = \sqrt{\psi\chi}$ and $\eta = \sqrt{\chi/\psi}$, is used with $\eta = 1$ (see [3]). Under this parametrization the density of the generalized hyperbolic distribution is

$$f_{\mathrm{H}}(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\omega + \boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})}{\omega + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right]^{\frac{\lambda - \frac{p}{2}}{2}} \frac{K_{\lambda - \frac{p}{2}}\left(\sqrt{[\omega + \boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}][\omega + \boldsymbol{\Sigma}(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\Sigma})]}\right)}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} K_{\lambda}(\omega) \exp\{-(\boldsymbol{\mu} - \mathbf{x})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}. \tag{4}$$

Details of this alternative parameterization, as well as maximum likelihood parameter estimates are given by [3]. Parameter estimation for the mixture of generalized hyperbolic distributions model can be carried out via the expectation-maximization (EM) algorithm [4].

1.1 Multiple scaled generalized hyperbolic distribution

The index and concentration parameters, λ and ω are unidimensional, i.e. they are the same for every dimension. Basing on the idea of [5], [8] proposed the multiple scaled GHD, where $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ are p-dimensional vectors, i.e., they can vary in each dimension. To introduce the multiple scaled distribution we need to define the normal variance-mean mixture. The distribution of a p-dimensional random variable $\mathbf{X}$ is said to be a normal variance-mean mixture if its density can be written in the form

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \boldsymbol{\theta}) = \int_0^\infty \phi_p\left(\mathbf{x} \mid \boldsymbol{\mu} + w\alpha, f(w)\boldsymbol{\Sigma}\right) h\left(w \mid \boldsymbol{\theta}\right) dw, \tag{5}$$

where $\phi_p\left(\mathbf{x} \mid \boldsymbol{\mu} + w\alpha, w\boldsymbol{\Sigma}\right)$ is the density of a p-dimensional Gaussian distribution with mean $\boldsymbol{\mu} + w\alpha$ and covariance matrix $f(w)\boldsymbol{\Sigma}$, and $h\left(w \mid \boldsymbol{\theta}\right)$ is the density of a univariate random variable $W > 0$ that has the role of a weight function [2, 6]. This weight function can take on many forms, when the density of W follow a generilized inverse Gaussian distribution, $f_t(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ follows the density of the GHD. [5] show that a multi-dimensional weight variable

$$\boldsymbol{\Sigma}_{\mathbf{W}} = \operatorname{diag}\left(w_1^{-1}, \ldots, w_p^{-1}\right)$$

can be incorporated into (5) via an eigen-decomposition of the symmetric positive-definite matrix $\boldsymbol{\Sigma}$, setting $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'$. Following [7] the formulation of the GHD in (4) can be written as a normal variance-mean mixture where the univariate density is GIG, i.e.,

$$\mathbf{X} = \boldsymbol{\mu} + W\alpha + \sqrt{W}\mathbf{V}, \tag{6}$$

where $\mathbf{V} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and W has density

$$h(w \mid \omega, 1, \lambda) = \frac{w^{\lambda - 1}}{2K_\lambda(\omega)} \exp\left\{-\frac{\omega}{2}\left(w + \frac{1}{w}\right)\right\}, \tag{7}$$

for $w > 0$, where ω and λ are as previously defined. From (6) and (7), it follows that the generalized hyperbolic density can be written

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \omega, \lambda) = \int_0^\infty \phi_p \left(\mathbf{x} \mid \boldsymbol{\mu} + w\alpha, w\boldsymbol{\Sigma}\right) h(w \mid \omega, 1, \lambda) dw. \tag{8}$$

The density of a multiple scaled generalized hyperbolic distribution (MSGHD) is

$$f_{\text{MSGHD}}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \alpha, \omega, \lambda) = \\ \int_0^\infty \cdots \int_0^\infty \phi_p \left(\boldsymbol{\Gamma}'\mathbf{x} - \boldsymbol{\mu} - \Delta_{\mathbf{w}}\alpha \mid \mathbf{0}, \Delta_{\mathbf{w}}\boldsymbol{\Phi}\right) h_{\mathbf{w}}(w_1, \ldots, w_p \mid \boldsymbol{\omega}, \mathbf{1}, \boldsymbol{\lambda}) dw_1 \ldots dw_p, \tag{9}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)'$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$, $\mathbf{1}$ is a p-vector of 1s, and

$$h_{\mathbf{W}}(w_1, \ldots, w_p \mid \boldsymbol{\omega}, \mathbf{1}, \boldsymbol{\lambda}) = h(w_1 \mid \omega_1, 1, \lambda_1) \times \cdots \times h(w_p \mid \omega_p, 1, \lambda_p).$$

Then, a mixture of MSGHDs (MMSGHDs) has density

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_{\text{MSGHD}} \left(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g, \boldsymbol{\Phi}_g, \alpha_g, \omega_g, \lambda_g\right). \tag{10}$$

Details of maximum likelihood parameter estimates and EM-algorithm are given by [7].

1.2 Mixture of Coalesced Generalized Hyperbolic Distributions

The generalized hyperbolic distribution is not a special or limiting case of the MSGHD under any parameterization with $p > 1$. [7] proposed a coalesced generalized hyperbolic distribution (CGHD) that contains both the generalized hyperbolic distribution and MSGHD as limiting cases. The CGHD arises through the introduction of a random vector

$$\mathbf{R} = U\mathbf{X} + (1 - U)\mathbf{S}, \tag{11}$$

where $\mathbf{X} = \boldsymbol{\Gamma}\mathbf{Y}$, $\mathbf{Y} \backsim \text{GHD}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \omega_0, \lambda_0)$, $\mathbf{S} \backsim \text{MSGHD}(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \alpha, \omega, \lambda)$, with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'$, and U is an indicator variable such that

$$U = \begin{cases} 1 & \text{if } R \text{ follows a generalized hyperbolic distribution, and} \\ 0 & \text{if } R \text{ follows a MSGHD.} \end{cases}$$

It follows that $\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\mu} + W\boldsymbol{\Gamma}\alpha + \sqrt{W}\boldsymbol{\Gamma}\mathbf{V}$, where $\boldsymbol{\Gamma}\mathbf{V} \backsim \text{N}_p\left(\mathbf{0}, \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\right)$, $\mathbf{S} = \boldsymbol{\Gamma}\boldsymbol{\mu} + \boldsymbol{\Gamma}\alpha\Delta_{\mathbf{w}} + \boldsymbol{\Gamma}\mathbf{A}$, where $\boldsymbol{\Gamma}\mathbf{A} \backsim \text{N}_p\left(\mathbf{0}, \boldsymbol{\Gamma}\Delta_{\mathbf{w}}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\right)$, and the density of $\mathbf{R}$ can be written

$$f_{\text{CGHD}}(\mathbf{r} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \alpha, \omega, \lambda, \omega_0, \lambda_0, \varpi) \\ = \varpi f_{\text{GHD}}\left(\mathbf{r} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}', \alpha, \omega_0, \lambda_0\right) + (1 - \varpi) f_{\text{MSGHD}}\left(\mathbf{r} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \alpha, \omega, \lambda\right), \tag{12}$$

where $f_{\text{GHD}}(\cdot)$ is the density of a generalized hyperbolic random variable, $f_{\text{MSGHD}}(\cdot)$ is the density of a MSGHD random variable, and $\varpi \in (0,1)$ is a mixing proportion. Note that the random vector $\mathbf{R}$ would be distributed generalized hyperbolic if $\varpi = 1$ and would be distributed MSGHD if $\varpi = 0$.

Parameter estimation can be carried out via a generalized expectation-maximization (GEM) algorithm [4].

2 Dimension reduction

The mixture of GHDs, MSGHDs, and CGHDs are extremely flexible and give good clustering performance; however, the flexibility is obtained increasing the number of parameters. This makes the methods unsuitable for high-dimensional data sets. The problem can be solved considering that the singular value decomposition of the scale matrix $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\Psi\boldsymbol{\Gamma}'$ naturally leads to dimension reduction using $p \times q$ $\boldsymbol{\Gamma}$ and $q \times q$ diagonal $\boldsymbol{\Phi}$ with $q < p$. The random variable $\mathbf{Y}$ is defined as

$$\mathbf{Y} = \boldsymbol{\Gamma}^{*}\mathbf{X} + \varepsilon, \tag{13}$$

with $\mathbf{X} \sim \text{GHD}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \omega, \lambda)$, and $\varepsilon \sim N(0, \Psi)$ where Ψ is a q dimensional diagonal matrix. Using (11) and (13) It follows that

$$\mathbf{Y} = \boldsymbol{\Gamma}^{*}\boldsymbol{\mu} + \boldsymbol{\Gamma}^{*}w\alpha + \boldsymbol{\Gamma}^{*}\sqrt{w}\mathbf{V} + \varepsilon, \tag{14}$$

and

$$\mathbf{Y} \sim GHD(\boldsymbol{\Gamma}^{*}\boldsymbol{\mu}, \boldsymbol{\Gamma}^{*}\alpha, (\boldsymbol{\Gamma}^{*}\boldsymbol{\Phi}(\boldsymbol{\Gamma}^{*})' + \Psi), \lambda, \omega). \tag{15}$$

Similarly if $\mathbf{X} \sim \text{MSGHD}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \omega, \lambda)$,

$$\mathbf{Z} \sim MSGHD(\boldsymbol{\Gamma}^{*}\boldsymbol{\mu}, \boldsymbol{\Gamma}^{*}\alpha, \omega, \lambda, \boldsymbol{\Phi} + \Psi). \tag{16}$$

Define $\tilde{\boldsymbol{\Phi}} := \Psi + \boldsymbol{\Phi}$, a $q \times q$ diagonal matrix. The mixture models obtained using the new proposed density function will be defined as low-dimension mixture of GHDs (LMGHDs) and lo- dimension mixture of MSGHDs (LMMSGHDs) respectively. Using the same procedure used in Section 1.2 we can obtain the lowdimension mixture of CGHDs (LMCGHDs). The parameters that maximize the likelihood for each model can be estimated using the EM-algorithm.

References

1. Barndorff-Nielsen, O., Halgreen, C.: Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. Z. Wahrscheinlichkeitstheorie Verw. Gebiete **38**, 309–311 (1977)
2. Barndorff-Nielsen, O., Kent, J., Sørensen, M.: Normal variance-mean mixtures and z distributions. International Statistical Review / Revue Internationale de Statistique **50**(2), 145–159

(1982)

3. Browne, R.P., McNicholas, P.D.: A mixture of generalized hyperbolic distributions. Canadian Journal of Statistics **43**(2), 176–198 (2015)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B **39**(1), 1–38 (1977)
5. Forbes, F., Wraith, D.: A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering. Statistics and Computing **24**(6), 971–984 (2014)
6. Gneiting, T.: Normal scale mixtures and dual probability densities. Journal of Statistical Computation and Simulation **59**(4), 375–384 (1997)
7. Tortora, C., Franczak, B., Browne, R., McNicholas, P.: A mixture of coalesced generalized hyperbolic distributions. Journal of Classification (accepted) (2018)
8. Tortora, C., McNicholas, P.D., Browne, R.P.: A mixture of generalized hyperbolic factor analyzers. Advances in Data Analysis and Classification **10**(4), 423–440 (2016)

A Comparison of Model-Based and Fuzzy Clustering Methods

Un Confronto tra Metodi di Clustering di tipo Model-Based e Fuzzy

Marco Alfó, Maria Brigida Ferraro, Paolo Giordani, Luca Scrucca, and Alessio Serafini

Abstract Model-based and fuzzy clustering methods represent widely used approaches for soft clustering. In the former approach, it is assumed that the data are generated by a mixture of probability distributions where each component represents a different group or cluster. Each observation unit is ex-post assigned to a cluster using the so-called posterior probability of component membership. In the latter case, no probabilistic assumptions are made and each observation unit belongs to a cluster according to the so-called fuzzy membership degree. The aim of this work is to compare the performance of both approaches by means of a simulation study.

Abstract *I metodi basati sugli approcci model-based e fuzzy rappresentano i piú comuni approcci di soft clustering. Nel primo approccio si assume che i dati siano generati da una mistura di distribuzioni di probabilitá nella quale ciascuna componente individua un gruppo. Le osservazioni sono assegnate ai gruppi ex-post con le cosiddette probabilitá a posteriori (di appartenenza alle componenti). Nell'altro approccio, che non prevede alcuna assunzione probabilistica, le osservazioni ven-*

Marco Alfó
Department of Statistical Sciences, Sapienza University of Rome, P.le Aldo Moro, 5, 00185 Rome, e-mail: marco.alfo@uniroma1.it

Maria Brigida Ferraro
Department of Statistical Sciences, Sapienza University of Rome, P.le Aldo Moro, 5, 00185 Rome, e-mail: mariabrigida.ferraro@uniroma1.it

Paolo Giordani
Department of Statistical Sciences, Sapienza University of Rome, P.le Aldo Moro, 5, 00185 Rome, e-mail: paolo.giordani@uniroma1.it

Luca Scrucca
Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli, 20, 06123 Perugia, e-mail: luca.scrucca@unipg.it

Alessio Serafini
Department of Statistical Sciences, Sapienza University of Rome, P.le Aldo Moro, 5, 00185, Rome e-mail: alessio.serafini@uniroma1.it

gono assegnate ai gruppi con i cosidetti gradi di appartenenza fuzzy. L'obiettivo di questo lavoro é confrontare le performance dei due approcci mediante uno studio di simulazione.

Key words: Cluster analysis, Model-based approach, Fuzzy approach

1 Introduction

In the last years, model-based and fuzzy clustering methods have received a great deal of attention. The two classes of methods are very different from a theoretical point of view. In the model-based framework, probabilistic assumptions are made. The data are generated by a mixture of known probability distributions (usually Gaussian). Each component of the mixture describes a cluster and, therefore, each cluster can be mathematically represented by a parametric distribution. In practice, the observation units are allocated to clusters via the so-called posterior probabilities of component membership and, for each observation unit, the cluster assignment is carried out by looking at the maximum posterior probability. In the fuzzy approach to clustering, the clusters are no longer represented in terms of parametric distributions. The observation units belong to the clusters according to the so-called membership degree, taking values in [0,1]. From a practical point of view, it is quite obvious that such two classes of clustering methods share similar features. Both of them produce a soft partition of the observation units and the posterior probability of component membership may play a role similar to the membership degree. Nevertheless, as far as we know, a thorough comparison between model-based and fuzzy clustering methods has never been carried out except for a few limited cases (see, e.g., [6, 10]). The aim of this work is to fill this gap by comparing the performances of four clustering methods, two from each class, in a simulation experiment.

2 Model-based clustering

Model-based clustering is a popular family of unsupervised learning methods for data classification. Such methods assume that the data are generated by a statistical model and try to recover it from the data. Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in \mathbb{R}^p$ be a random sample of *i.i.d.* observations, where p denotes the number of variables. The random vector $\mathbf{x}_i$ is assumed to arise from a finite mixture of probability density functions:

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f(\mathbf{x}_i | \boldsymbol{\theta}_k), \tag{1}$$

where π_k, $k = 1, \ldots, g$, such that $\pi_k > 0$ and $\sum_{k=1}^{g} \pi_k = 1$, are the mixing proportions, g is the number of components, $f(\mathbf{x}|\boldsymbol{\theta}_k)$ is the component density $(k = 1, \ldots, g)$

and $\boldsymbol{\Phi} = (\pi_1, \pi_2, \ldots, \pi_g, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_g)$ is the parameter vector [20]. Each mixture component density belongs to a specific parametric class and represents a group or cluster. Even thought it is not necessary that each mixture component density arises from the same parametric distribution family, we will focus only on the case where the parametric distribution family is the same for each mixture component.

Maximum likelihood estimation of model parameters in $\boldsymbol{\Phi}$ is done by applying the Expectation-Maximization (EM) algorithm [11]. It is an iterative procedure to estimate the parameters of a finite mixture model by maximizing the expected value of the complete data log-likelihood by alternating two different steps. The Expectation step (E-step) computes the expected value of the complete data log-likelihood, and the Maximization step (M-step) maximises the expected value previously computed with respect to $\boldsymbol{\Phi}$. The log-likelihood can be derived as follows:

$$\ell(\boldsymbol{\Phi}) = \log \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k f(\mathbf{x}_i|\boldsymbol{\theta}_k) = \sum_{i=1}^{n} \log \sum_{k=1}^{g} \pi_k f(\mathbf{x}_i|\boldsymbol{\theta}_k). \tag{2}$$

Suppose to have an unobservable process $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$. We refer to this new dataset $(\mathbf{x}, \mathbf{z})$ as the complete data with density $f(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$. The complete data log-likelihood is given by

$$\ell_c(\boldsymbol{\Phi}) = \log \prod_{i=1}^{n} f(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{f(\mathbf{z}_i|\boldsymbol{\theta}) f(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})\}. \tag{3}$$

The E-step computes the expected value of the complete-data log-likelihood in (3) with respect to the missing part:

$$Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^t) = \mathbb{E}_{\boldsymbol{\Phi}^t}[\ell_c(\boldsymbol{\Phi})] = \mathbb{E}_{\boldsymbol{\Phi}^t}\left[\sum_{i=1}^{n} \log f(\mathbf{z}_i|\boldsymbol{\theta}^t) f(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}^t)\right]. \tag{4}$$

The M-step maximises equation in (4), such that:

$$\boldsymbol{\Phi}^{t+1} = \arg\max_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^t). \tag{5}$$

The procedure is iterated until some convergence criterion is satisfied. The EM algorithm guaranteesthat the observed log-likelihood is nondecreasing and, under fairly general conditions, the sequence converges to at least a local maximum [19]. Further details can be found in, e.g., [19, 20].

2.1 Finite mixtures of Gaussian densities

Due to its flexibility and mathematical tractability, the most popular model for clustering postulates that the data follow a Gaussian mixture distribution, i.e. $f(\mathbf{x}_i, |z_{ik} = 1, \boldsymbol{\theta}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ [15]. The finite mixture of Gaussian densities is then given by

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \phi(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{6}$$

where $\boldsymbol{\theta} = \{\pi_1, \pi_2, \ldots, \pi_{k-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k\}$ denotes the parameter set for the finite mixture model and $\phi(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ the underlying component-specific density function with parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, g$.

Thus, model in (6) generates ellipsoidal clusters centred at the mean vector $\boldsymbol{\mu}_k$, with $\boldsymbol{\Sigma}_k$ controlling the other geometrical properties of each cluster. Parsimonious parametrisations of the cluster covariance matrices can be obtained through the eigendecomposition $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^\top$ [4, 9], where λ_k is a scalar controlling the volume of the ellipsoid, $\mathbf{A}_k$ is a diagonal matrix controlling its shape and $\mathbf{D}_k$ is an orthogonal matrix controlling the orientation of the ellipsoid. Such an eigendecomposition generates a class of models with different geometrical properties. For some covariance parametrisations a closed formula for the M-step in the EM algorithm can be obtained [9]. The estimation for each of the 14 different models resulting from the eigendecomposition of the within clusters covariance matrices is implemented in the R package **mclust** [23].

The number of clusters and the parametrisation of the covariance matrices are selected using selection criteria, such as the Bayesian information criterion (*BIC*) [14, 22].

2.2 Finite mixtures of t densities

In [21], a heavy-tailed alternative to the component-specific density (6) is proposed by replacing the Gaussian distribution with a *t* distribution as follows:

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_t(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k), \tag{7}$$

where $\boldsymbol{\theta} = \{\pi_1, \pi_2, \ldots, \pi_{k-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k, \nu_1, \ldots, \nu_k\}$ and $f_t(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k)$ is the multivariate *t* distribution with mean $\boldsymbol{\mu}_k$, covariance matrix $\boldsymbol{\Sigma}_k$ and ν_k degrees of freedom. As in the Gaussian case, the EM algorithm is employed for maximum likelihood estimation of $\boldsymbol{\theta}$ [2, 21]. In the present context, a parsimonious parametrization can be based on the same eigendecomposition of the covariance matrices as in the Gaussian case and constraining the degrees of freedom to be equal or not across groups [1]. This produces a class of finite mixture models with *t*-distributed components called *tEIGEN*. The *tEIGEN* family is implemented in the R package **teigen** [1, 2].

3 Fuzzy clustering

As opposed to the model-based framework, no probabilistic assumptions are made in the fuzzy approach to clustering. The complexity of the clustering process is managed in terms of fuzziness [24]. The observation units are assigned to the clusters according to the so-called fuzzy membership degree, taking values in [0,1]. This is inversely related to the dissimilarity between the observation units and the cluster prototype. A membership degree approaching 1 implies that the observation unit is close to the corresponding prototype and therefore it can be clearly assigned to the cluster. In the literature, there exist several fuzzy clustering methods. Among them, the most common one is the Fuzzy *k*-Means (F*k*M) algorithm [5]. In the following subsections, we will recall the F*k*M algorithm and the closely related Gustafson-Kessel variant [12]. The algorithms are implemented in the R package **fclust** [13].

3.1 Fuzzy k-Means

The Fuzzy *k*-Means (F*k*M) clustering algorithm [5] aims at grouping *n* observation units in *k* clusters by solving the following constrained optimization problem:

$$\begin{array}{lll} \min\limits_{\mathbf{U},\mathbf{H}} J_{FkM} & = \sum_{i=1}^{n} \sum_{k=1}^{g} u_{ik}^{m} d^2 \left(\mathbf{x}_i, \mathbf{h}_k\right), & \\ \text{s.t.} & u_{ik} \geq 0, \quad i=1,\ldots,n, \quad k=1,\ldots,g, & \\ & \sum_{k=1}^{g} u_{ik} = 1, \quad i=1,\ldots,n. & \end{array} \tag{8}$$

In (8), the term u_{ik} denotes the membership degree of observation unit i to cluster k, as a generic element of the matrix $\mathbf{U}$ of order $(n \times g)$. The row-wise sum of $\mathbf{U}$ is equal to 1. Furthermore, $\mathbf{h}_k = \left[h_{k1}, \ldots, h_{kp}\right]$, the k-th row of the prototype matrix $\mathbf{H}$ of order $(g \times p)$, is the prototype for cluster k, $k = 1, \ldots, g$. Finally, $d^2\left(\mathbf{x}_i, \mathbf{h}_k\right)$ is the squared Euclidean distance between observation unit i and prototype k, while $m > 1$ is the fuzziness parameter which tunes the level of fuzziness of the obtained partition. The higher the values of m, the fuzzier the partition with membership degrees tending to $\frac{1}{k}$. When m is close to 1, the F*k*M solution approaches that of the standard (non-fuzzy or hard) *k*-means [18] with membership degrees equal to either 0 or 1. The standard choice is $m = 2$.

The solution of (8) is carried out through an iterative optimization algorithm by updating the elements of $\mathbf{U}$ as follows

$$u_{ik} = \frac{1}{\sum_{k'=1}^{g} \left(\frac{d^2(\mathbf{x}_i, \mathbf{h}_k)}{d^2\left(\mathbf{x}_i, \mathbf{h}_{k'}\right)} \right)^{\frac{1}{m-1}}}, \quad i=1,\ldots,n, \quad k=1,\ldots,g, \tag{9}$$

and the rows of $\mathbf{H}$ as

$$\mathbf{h}_k = \frac{\sum_{i=1}^{n} u_{ik}^{m} \mathbf{x}_i}{\sum_{i=1}^{n} u_{ik}^{m}}, \quad k=1,\ldots,g. \tag{10}$$

In order to select the optimal number of clusters, several cluster validity indexes can be adopted. A popular choice is the Fuzzy Silhouette index [8].

3.2 Gustafson-Kessel variant of FkM

The major limitation of the F*k*M algorithm is that the obtained clusters are defined to be spherical. Therefore, F*k*M may be inadequate whenever the clusters have different geometrical shapes. In such cases, the so-called Gustafson-Kessel variant of the F*k*M algorithm can be applied, hereinafter GK-F*k*M [12]. The main difference between F*k*M and GK-F*k*M is that, in the latter, a Mahalanobis-type dissimilarity is considered, that is, $d^2\left(\mathbf{x}_i, \mathbf{h}_k\right)$ is replaced by

$$d_M^2\left(\mathbf{x}_i, \mathbf{h}_k\right) = \left(\mathbf{x}_i - \mathbf{h}_k\right)^\top \mathbf{M}_k (\mathbf{x}_i - \mathbf{h}_k), \tag{11}$$

with $\mathbf{M}_k$ symmetric and positive definite. The GK-F*k*M can then be formulated as

$$\begin{array}{lll} \min\limits_{\mathbf{U},\mathbf{H},\mathbf{M}_1,\ldots,\mathbf{M}_g} & J_{GK-FkM} & = \sum_{i=1}^n \sum_{k=1}^g u_{ik}^m d_M^2\left(\mathbf{x}_i, \mathbf{h}_k\right), \\ \text{s.t.} & & u_{ik} \geq 0, \quad i = 1, \ldots, n, \quad k = 1, \ldots, g, \\ & & \sum_{k=1}^g u_{ik} = 1, \quad i = 1, \ldots, n, \\ & & |\mathbf{M}_k| = \rho_k > 0 \quad k = 1, \ldots, g. \end{array} \tag{12}$$

As the cost function is linear with respect to the matrices $\mathbf{M}_k$, a trivial solution with $\mathbf{M}_k = \mathbf{0}, k = 1, \ldots, g$ would be obtained. To avoid it, $\mathbf{M}_k$ must be constrained. A way to do it is to consider volume constraints such that the determinant of $\mathbf{M}_k$ is positive. Note that the most common choice is $\rho_k = 1$, $k = 1, \ldots, g$.

An iterative solution of (12) can be found by updating $\mathbf{U}$ and $\mathbf{H}$ according to (9) and (10) provided that d^2 is replaced by d_M^2 and $\mathbf{M}_k$ by

$$\mathbf{M}_k = [\rho_k |\mathbf{V}_k|]^{\frac{1}{n}} \mathbf{V}_k^{-1}, \quad k = 1, \ldots, g, \tag{13}$$

where $\mathbf{V}_k$ is the fuzzy covariance matrix for cluster k given by

$$\mathbf{V}_k = \frac{\sum_{i=1}^n u_{ik}^m (\mathbf{x}_i - \mathbf{h}_k)(\mathbf{x}_i - \mathbf{h}_k)^\top}{\sum_{i=1}^n u_{ik}^m}, \quad k = 1, \ldots, g. \tag{14}$$

To avoid possible numerical problems for updating $\mathbf{M}_k$, a computational improvement has been proposed in [3] where the condition number of $\mathbf{M}_k$ is constrained to be higher than a prespecified threshold. Note that this condition is similar to that imposed to covariance matrices in finite mixture models with either Gaussian or t components [16, 17].

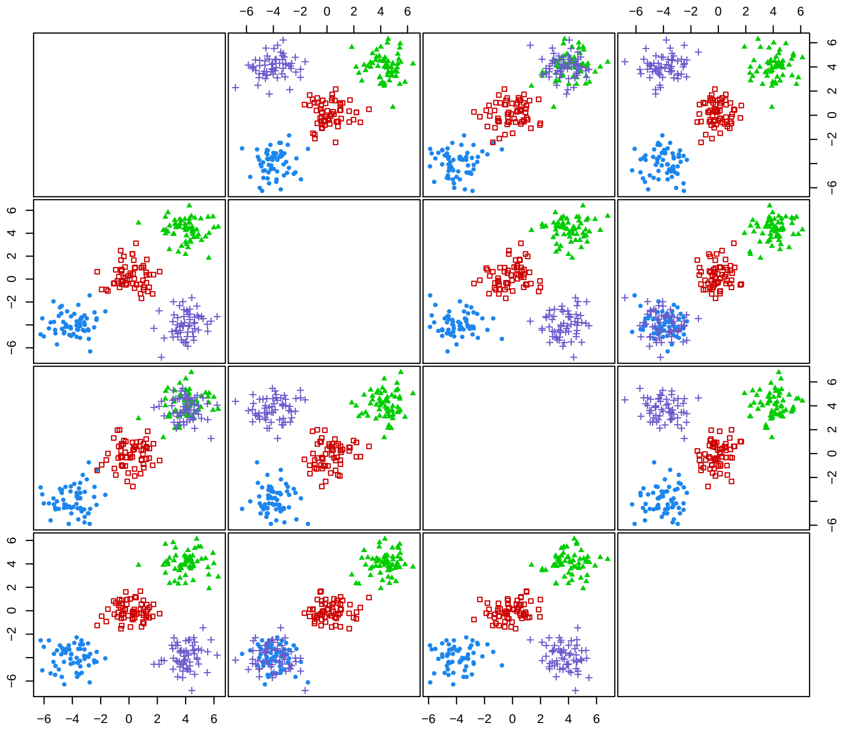


Fig. 1 Example of simulated data

4 Simulation study

A simulation study has been carried out to compare the previously described clustering methods. Simulated data sets have been generated randomly in a full factorial design; an example of generated data is displayed in Figure 1. In the simulation study, the focus lies on assessing the performance of the methods in recovering the cluster structure and checking whether the design variables influence the differential performance of the methods. The results will be presented at the meeting.

References

1. Andrews, J.L., McNicholas, P.D.: Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. Stat. Comput. **22**, 1021–1029 (2012)
2. Andrews, J.L., McNicholas, P.D.: teigen: Model-based clustering and classification with the multivariate t-distribution, R package version 2 (2015)

3. Babuška, R., van der Veen, P.J., Kaymak, U.: Improved covariance estimation for Gustafson-Kessel clustering. In: IEEE International Conference on Fuzzy Systems, pp. 1081–1085 (2002)
4. Banfield, J.D., Raftery, A.E.: Model-based gaussian and non-gaussian clustering. Biometrics **49**, 803–821 (1993)
5. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithm. Plenum Press, New York (1981)
6. Bezdek, J.C., Hathaway, R.J., Huggins, V.J.: Parameter estimation for normal mixtures. Pattern Recognit. Lett. **3**, 79–84 (1985)
7. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 719–725 (2000)
8. Campello, R.J.G.B., Hruschka, E.R.: A fuzzy extension of the silhouette width criterion for cluster analysis, Fuzzy Sets Syst. **157**, 2858–2875 (2006)
9. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**, 781–793 (1995)
10. Davenport, J.W., Bezdek, J.C., Hathaway, R.J.: Parameter estimation for finite mixture distributions. Comput. Math. Applic. **15**, 819–828 (1988)
11. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. Series B **39**, 1–38 (1977)
12. Gustafson, E., Kessel, W.: Fuzzy clustering with a fuzzy covariance matrix. In: IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes, pp. 761–766 (1978)
13. Ferraro, M.B., Giordani, P.: A toolbox for fuzzy clustering using the R programming language. Fuzzy Sets Syst. **279**, 1–16 (2015)
14. Fraley, C., Raftery, A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. **41**, 578–588 (1998)
15. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**, 611–631 (2002)
16. Greselin, F., Ingrassia, S.: Constrained monotone EM algorithms for mixtures of multivariate *t* distributions. Stat. Comput. 20, 9–22 (2010)
17. Ingrassia, S., Rocci, R.: Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. Comput. Stat. Data Anal. 51, 5339–5351 (2007)
18. Mac Queen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp. 281–297 (1967)
19. McLachlan, G., Krishnan, T.: The EM algorithm and extensions. Wiley, New York (2008)
20. McLachlan, G., Peel, D.: Finite mixture models. Wiley, New York (2000)
21. Peel, D., McLachlan, G.: Robust mixture modelling using the t distribution. Stat. Comput. **10**, 339–348 (2000)
22. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6** 461–464 (1978)
23. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. **8** 289–317 (2016)
24. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)

Covariate measurement error in generalized linear models for longitudinal data: a latent Markov approach

Roberto Di Mari*[1], Antonio Punzo[1], and Antonello Maruotti[2,3]

[1]Department of Economics and Business, University of Catania, Italy
[2]Department of Law, Economics, Politics and Modern languages, LUMSA, Italy
[3]Centre for Innovation and Leadership in Health Sciences, University of Southampton, UK

Abstract

One common approach to handle covariate measurement error in Generalized Linear Models (GLM) is classical error modeling. In the past 20 years, classical error modeling has been brought to (Non-Parametric) Maximum Likelihood (NPML) estimation, by means of finite mixture modeling: the supposedly continuous true score is modeled as a discrete (multinomial) static latent variable, and is handled as a part of the model. Nonetheless, the true score is not allowed to vary over time: if the true score has own underlying dynamics, these are either unaccounted for or mistaken for measurement error, or possibly both. The aim of the present paper is to formulate a joint model for the outcome variable, the covariate observed with error (measurement model), and the true score model that accounts for the underlying dynamics in the true score. The true score and its dynamics are modeled non-parametrically as a first-order latent (hidden) Markov chain. Estimation is done extending the NPML approach, in a full maximum likelihood environment with a well-know modification of the EM algorithm (forward-backward algorithm). From an applied researcher perspective, our methodology can safely handle both the case where the latent underlying characteristic is stable over time, as well as providing a suitable framework even when changes across measurement occasions are substantial. Within a GLM framework, it is demonstrated, by means of extensive simulation studies, that this is crucial to get correct estimates of the regression coefficients, as well as good coverages. In the real-data application, the effect of heart rate on the occurrence of cardiovascular diseases in a sample of Chinese elderly patients is measured. Modeling the true (unobserved) heart rate and its dynamics - which, in elderly patients, are likely to be non negligible - will be showed to allow a correct assessment of risk factors of cardiovascular diseases occurrence.

KEY-WORDS: covariate measurement error, errors–in–variables, longitudinal data, generalized linear models, latent Markov models.

*roberto.dimari@unict.it

1 Introduction

In all areas of scientific research, being able to collect high quality measures in order to assess a given phenomenon of interest is crucial. Unaccounted measurement error due to poor measures can severely distort the analysis, leading to unreasonable substantial conclusions.
A relevant type of measurement error is covariate measurement error - or errors in variables. Covariate measurement error issues have long history in epidemiological studies. For instance, Cotton et al. (2005) show how measurement error can affect diagnosis of developmental dyslexia in children with reading difficulties. Kipnis et al. (2003) find that dietary intake assessed with error in a biomarker study produces a severely attenuated estimate of disease relative risk. Guo & Little (2011) report, in pre-menopause women, a negative effect of carotenoids on progesterone, estimated to be zero if measurement error is not accounted for.
The two most common approaches for covariate measurement error modeling are classical error models - known also as error calibration models - and regression calibration models (for an extensive review, see Carroll et al., 2006). In the past 20 years, Aitkin (1996, 1999) and Aitkin & Rocci (2002) among other, have provided a way to bring classical error modeling to maximum likelihood estimation, allowing the user to make no parametric assumption on the true score. That is, nonparametric maximum likelihood (NPML) handles the distribution of the true score as a part of the model non-parametrically, by using finite mixture models (for a recent review, see for instance Alfó & Viviani, 2016. Whereby remaining in a fully ML setup, this approach, in many practical situations, can be more convenient than assuming normality of the true score - as is commonly done in the regression calibration literature. Nonetheless, the true score is not allowed to vary over time. If the true character has own underlying dynamics, these are either unaccounted for or mistaken for measurement error (Alwin, 2007), or possibly both.
Separating unreliability from change in the true score is possible if the true score dynamics are modeled. Quasi Simplex models, or Quasi Autoregressive/Markov Simplex models (Alwin, 2007) are used in survey methods literature to address the issue of changes in the true-score distribution (see, for instance, Uhrig & Watson, 2017). The model is fitted in a confirmatory factor-analytic Structural Equation Modeling (SEM) framework, by assuming a continuous dynamic latent variable with Gaussian error (one-factor model). A similar approach in epidemiology can be found in Sánchez et al. (2009), who study the effects of in-utero lead exposure on child development.
The aim of this work is to formulate a joint model for the outcome variable, the covariate observed with error (measurement model), and the true score model that accounts for the underlying dynamics in the true score. The true score and its dynamics are modeled non-parametrically as a first-order latent (hidden) Markov chain (Bartolucci et al., 2012; Collins & Lanza, 2010; Rabe-Hesketh & Skrondal, 2008; J. K. Vermunt et al., 1999; Wiggins, 1973; Zucchini et al., 2016). Model estimation is done in a fully maximum likelihood environment, with a well-know modification of the EM algorithm (Baum et al., 1970; Welch, 2003).
Our approach is closely related to Aitkin & Rocci (2002)'s, in that we make no distributional assumption on the true score, whereby the key novelty is in modeling the true-score

dynamics. From an applied researcher perspective, our methodology can safely handle both the case where the latent underlying characteristic is stable over time, as well as providing a suitable framework even when changes across measurement occasions are substantial. Within a generalized linear modeling (GLM) framework, we demonstrate that this is crucial to get correct estimates of the regression coefficients, as well as good coverages. Although confirmatory factor-analytic/SEM methodologies allow for dynamics in the true score, estimation relies on identifying restrictions (for instance on the true score variance) and distributional assumptions (normality of each regression errors), which might be restrictive in certain practical situations. In the methodology we propose, we need no identifying restrictions, and we handle non-parametrically a possibly continuous underlying latent variable, modeling it as a dynamic discrete trait (Catania & Di Mari, 2018; Di Mari & Bakk, 2017; J. Vermunt & Magidson, 2004).

We illustrate the proposed methodology by analyzing data from the Chinese Longitudinal Healthy Longevity Survey, where a sample of n Chinese old patients is observed T times, and information on cardiovascular diseases for each person is reported alongside demographics and well-known risk factors, among which heart rate. The aim is to measure the effect of heart rate on the occurrence of a cardiovascular disease, controlling for demographic characteristics and dietary habits.

The structure of the paper is as follows. In Section 2 we will give details on the modeling specification, and describe how the model parameters can be estimated with our latent Markov approach (Section 2. In Section 4 we will summarize the results from the simulation study and the empirical application.

2 Outcome, measurement and error components in common error correction modeling

Let Y_t, W_t and $\mathbf{Z}_t$ be respectively the outcome variable, a continuous covariate corresponding to the true score X_t and a k-vector of error–free covariates, for $t = 0, \dots, T$. In addition, we let $\mathbf{Y}$ be the full vector of outcomes, $\mathbf{Z}$ the full set of available covariates, and $\mathbf{W}$ the full vector of covariate values with corresponding true scores $\mathbf{X}$, observed for the $T+1$ time occasions. We assume that, given the true score, Y_t and W_t are conditionally independent - non–differential measurement error model.

As it is typical in GLM context, we assume Y_t to have distribution belonging to the exponential family, with the following linear predictor

$$\eta_t(\boldsymbol{\theta}) = \alpha + \beta\,\mathrm{X}_t + \boldsymbol{\gamma}'\,\mathbf{Z}_t, \tag{1}$$

where $\eta_t(.)$ is an appropriate link function, $\boldsymbol{\gamma}$ and β are respectively k-vector and scalar regression coefficients, α is an intercept term and $\boldsymbol{\theta} = \{\alpha, \beta, \boldsymbol{\gamma}\}$.

Equation (1) defines the outcome model in terms of its linear predictor.

As for the classical measurement error model, we assume

$$\mathrm{W}_t = \mathrm{X}_t + \xi_t, \tag{2}$$

where $\xi_t \sim N(0, \sigma^2_{\mathrm{W}})$. The classical additive model can be also applied to variables transformed in log-scale, in order to model multiplicative rather than additive error.

The true score and its dynamics can be modeled extending the usual assumption of (conditional) normality of the true score (given the exogenous covariates $\mathbf{Z}_t$; Aitkin & Rocci, 2002), by assuming the score follows an AR(1) process, such that

$$\begin{aligned} \mathrm{X}_0 &= \epsilon_0 + \boldsymbol{\lambda}' \mathbf{Z}_0, \\ \mathrm{X}_t &= \mathrm{X}_{t-1}\, \rho + \boldsymbol{\lambda}' \mathbf{Z}_t + \epsilon_t. \end{aligned} \tag{3}$$

For convenience, we define the transformed true score $\mathrm{X}_0^* = \mathrm{X}_0 - \boldsymbol{\lambda}' \mathbf{Z}_0$, $\mathrm{X}_t^* = \mathrm{X}_t - \boldsymbol{\lambda}' \mathbf{Z}_t$, and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma} + \beta\lambda$. By dropping the stars, the measurement model is now transformed as

$$\mathrm{W}_t = \mathrm{X}_t + \boldsymbol{\lambda}' \mathbf{Z}_t + \xi_t, \tag{4}$$

and the linear predictor of the outcome model becomes

$$\eta_t(\boldsymbol{\theta}) = \alpha + \beta\, \mathrm{X}_t + \boldsymbol{\gamma}' \mathbf{Z}_t\,. \tag{5}$$

We can now express, using Aitkin & Rocci (2002)'s notation, the following joint model for $(\mathrm{Y}_t, \mathrm{W}_t, \mathrm{X}_t \,|\, \mathbf{Z}_t)$

$$P(\mathrm{Y}_t, \mathrm{W}_t, \mathrm{X}_t \,|\, \mathbf{Z}_t) = P(\mathrm{Y}_t \,|\, \mathrm{X}_t, \mathbf{Z}_t) m(W_t |\, \mathrm{X}_t, \mathbf{Z}_t) \pi(X_t), \tag{6}$$

where $P(\mathrm{Y}_t \,|\, \mathrm{X}_t, \mathbf{Z}_t)$ is the density or pmf of the outcome, $m(W_t |\, \mathrm{X}_t, \mathbf{Z}_t)$ is the measurement model density, and $\pi(X_t)$ is the true score density.
We can now define the following joint marginal distribution for (Y_t, W_t)

$$P(\mathrm{Y}_t, \mathrm{W}_t, \mathrm{X}_t \,|\, \mathbf{Z}_t) = \int P(\mathrm{Y}_t \,|\, \mathrm{X}_t) m(W_t |\, \mathrm{X}_t) \pi(X_t) dX_t, \tag{7}$$

Let $\{(\mathrm{Y}_{it}, \mathrm{W}_{it}, \mathbf{Z}_{it})\}_n = \{(\mathrm{Y}_{1t}, \mathrm{W}_{1t}, \mathbf{Z}_{1t}), \ldots, (\mathrm{Y}_{nt}, \mathrm{W}_{nt}, \mathbf{Z}_{nt})\}$ be a sample of n independent observations, observed for $t = 0, \ldots, T$ time points. The sample log-likelihood - corresponding to the model of Equation (7) for $t = 0, \ldots, T$ - can be defined as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \int P(\mathbf{Y}_i \,|\, \mathbf{X}_i, \mathbf{Z}_i) m(\mathbf{W}_i \,|\, \mathbf{X}_i, \mathbf{Z}_i) \pi(\mathbf{X}_i) d\, \mathbf{X}_i \right\}, \tag{8}$$

where

$$P(\mathbf{Y}_i, |\, \mathbf{X}_i, \mathbf{Z}_i) m(\mathbf{W}_i \,|\, \mathbf{X}_i, \mathbf{Z}_i) = \prod_{t=0}^{T} P(\mathrm{Y}_{it} \,|\, \mathrm{X}_{it}, \mathbf{Z}_{it}) m(W_{it} |\, \mathrm{X}_{it}, \mathbf{Z}_{it}), \tag{9}$$

due to local independence of the distributions of outcome and the covariate measured with error across time given the true score. The assumption on the distribution of X_t involves also assumptions on its dynamics. With a relatively simple AR(1) specification, estimating the model parameters by maximixing the (log) likelihood of Equation (8) requires evaluating an integral over a $(T + 1)$-dimensional space. This can be done by using a (nonlinear) filtering algorithm, known in the time series literature (Heiss, 2008), which is based on the sequential application of Gaussian quadrature rules (see also Bartolucci et al., 2014).

3 Handling covariate measurement error with latent Markov modeling

The idea underlying the NPML approach (Laird, 1978; Lindsay, 1983a,b) is that finding the MLE of $\pi(\cdot)$, say $\widehat{\pi}(\cdot)$, involves a standard convex optimization problem and, as long as the model likelihood is bounded, $\widehat{\pi}(\cdot)$ is concentrated over at most as many support points as the number of sample units, and is uniquely identified by locations and related masses. We let S, with $S \leq n$, be the state space of the concentrated distribution at time t, X_{it}^{s} be the realized discretized true score for the i-th observation at time t corresponding to the s-th location x_s, with time–varying mass π_{st}, for $s = 1, \dots, S$, at time t. We propose to model the time–varying masses by using the properties of first–order homogeneous Markov chains. In particular, by letting $\boldsymbol{\delta} = \{\delta_s\}_S$ be the common initial probabilities, where $\delta_s = P(\mathrm{X}_{i0}^{s} = s)$, and $\mathbf{Q}$ the common transition matrix, with elements $\{\mathrm{q}_{rs}\}$, where $\mathrm{q}_{rs} = P(\mathrm{X}_{t}^{s} = \mathrm{x}_s \,|\, \mathrm{X}_{t-1}^{s} = \mathrm{x}_r)$ with $1 < s \leq S$, and $1 < r \leq S$, we can approximate the log likelihood function of Equation (8) as follows

$$\ell(\boldsymbol{\theta}) \approx \sum_{i=1}^{n} \log \left\{ \prod_{t=0}^{T} \sum_{s=1}^{S} P(\mathrm{Y}_{it} \,|\, \mathrm{X}_{it}^{s}, \mathbf{Z}_{it}) m(\mathrm{W}_{it} \,|\, \mathrm{X}_{it}^{s}, \mathbf{Z}_{it}) \pi_{st} \right\}, \tag{10}$$

where, thanks to the properties of Markov chains, $\pi_{s0} = \delta_s$, and π_{st} is the s-th element of the vector $\boldsymbol{\pi}_t = \boldsymbol{\delta}' * \mathbf{Q}^t$.

The elements of the initial state probabilities and the transition probabilities can be parametrized according to logistic parametrizations as follows

$$\log \frac{P(X_0 = s)}{P(X_0 = 1)} = \beta_{s0}, \tag{11}$$

with $1 < s \leq S$, for the initial state probability, and

$$\log \frac{P(\mathrm{X}_t = s | X_{t-1} = r)}{P(\mathrm{X}_t = 1 | X_{t-1} = r)} = \gamma_{0s} + \gamma_{0rs}, \tag{12}$$

with $1 < s \leq S$, and $1 < r \leq S$ for the transitions probabilities. We take the first category as reference - setting to zero the related parameters. For the transition model, this means that parameters related to the elements in the first row and column of the transition matrix are set to zero.

Iterative procedures, like the EM algorithm (Dempster et al., 1977) can be used to maximize Equation (10) in order to estimate the model parameters in one step. However, when using the standard EM, the time and storage required for parameter estimation of latent Markov models increase exponentially with the number of time points (Vermunt, Langeheine, & Böckenholt, 1999). For this reason, the forward–backward algorithm (Baum et al., 1970; Welch, 2003) is typically implemented: this is a special version of the standard EM in which the size of the problem increases only linearly with the number of time occasions (Zucchini et al., 2016).

4 Results and conclusions

We have assessed the proposed latent Markov approach for parameters estimation of generalized linear models with covariate(s) measured with error under a broad set of scenarios, resulting from combinations of different sample size (100, 500 and 1000 for each time point, with $T = 5$), measurement error size ($\sigma^2_{\mathrm{W}} = (1, 1.5, 2)$), and size of the effect of the true score on the outcome $\beta = (1, 1.5)$, for both continuous and dichotomous outcome variables. We have generated the continuous true score from a continuous dynamic model. We found that our latent Markov approach with a number of states between 3 and 5 is enough to approximate the continuous underlying distribution of the true score. The results have showed that the proposed method yields correct parameter estimates in all conditions, except for small sample size (100 observations), as well as good coverages.
In the empirical application on the Chinese Longitudinal Healthy Longevity Survey data, we were able to find, modeling the true (unobserved) heart rate and its dynamics, risk factors for the cardiovascular disease occurrence consistent with the medical literature.

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, *6*(3), 251–262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*(1), 117–128.

Aitkin, M., & Rocci, R. (2002). A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, *12*(2), 163–174.

Alfó, M., & Viviani, S. (2016). Finite mixtures of structured models. In H. C, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of Cluster Analysis* (pp. 217–240). Chapman & Hall: Boca Raton, FL.

Alwin, D. F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement* (Vol. 547). John Wiley & Sons.

Bartolucci, F., Bacci, S., & Pennoni, F. (2014). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(2), 267–288.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2012). *Latent markov models for longitudinal data.* Chapman and Hall / CRC Press.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: a Modern Perspective.* CRC press.

Catania, L., & Di Mari, R. (2018). Hierarchical hidden markov models for multivariate integer-valued time-series with application to crime data. *Under review.*

Collins, L. M., & Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences* (Vol. 718). Wiley.

Cotton, S. M., Crewther, D. P., & Crewther, S. G. (2005). Measurement error: Implications for diagnosis and discrepancy models of developmental dyslexia. *Dyslexia*, *11*(3), 186–202.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Di Mari, R., & Bakk, Z. (2017). Mostly harmless direct effects: a comparison of different latent markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal.*

Guo, Y., & Little, R. J. (2011). Regression analysis with covariates that have heteroscedastic measurement error. *Statistics in Medicine*, *30*(18), 2278–2294.

Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for microeconometric panel data. *Journal of Applied Econometrics*, *23*(3), 373–389.

Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. P., ... Carroll, R. J. (2003). Structure of dietary measurement error: results of the open biomarker study. *American Journal of Epidemiology*, *158*(1), 14–21.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805–811.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, *11*(1), 86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part ii: the exponential family. *The Annals of Statistics*, *11*(3), 783–792.

Rabe-Hesketh, S., & Skrondal, A. (2008). Classical latent variable models for medical research. *Statistical Methods in Medical Research*, *17*(1), 5–32.

Sánchez, B. N., Budtz-Jørgensen, E., & Ryan, L. M. (2009). An estimating equations approach to fitting latent exposure models with longitudinal health outcomes. *The Annals of Applied Statistics*, 830–856.

Uhrig, S. C. N., & Watson, N. (2017). The impact of measurement error on wage decompositions: Evidence from the british household panel survey and the household, income and labour dynamics in australia survey. *Sociological Methods & Research.*

Vermunt, J., & Magidson, J. (2004). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In L. van der Ark, M. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (p. 41-63). Mahwah, NJ: Erlbaum.

Vermunt, J. K., Langeheine, R., & B ockenholt, U. (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*, 179-207.

Welch, L. R. (2003). Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, *53*(4), 10–13.

Wiggins, L. M. (1973). *Panel analysis: latent probability models for attitude and behaviour processes.* Elsevier, Amsterdam.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov Models for Time Series: an Introduction Using R.* Chapman and Hall / CRC Press.

Statistical Modelling

A regularized estimation approach for the three-parameter logistic model

Un approccio regolarizzato per la stima del modello logistico con tre parametri

Michela Battauz and Ruggero Bellio

Abstract The three-parameter logistic model is an item response theory model used with dichotomous items. It is well known that the parameters of the model are weekly identifiable and that the maximization of the likelihood, which is performed using numerical algorithms, is prone to convergence issues. In this paper, we propose the use of a penalized likelihood for the estimation of the item parameters. In particular, the penalty term shrinks the guessing parameters towards a known constant. Cross-validation is used to select such constant and the amount of shrinkage, though estimation via an empirical Bayes approach is also considered. The method is both simple and effective, and it is illustrated by means of a simulation study and a real data example.

Abstract *Il modello logistico con tre parametri è uno dei modelli item response theory usato con item dicotomici. In letteratura è noto che i parametri del modello sono debolmente identificabili e che la massimizzazione della verosimiglianza, che è condotta attraverso algoritmi numerici, è soggetta a problemi di convergenza. In questo articolo, si propone l'uso di una verosimiglianza penalizzata per la stima dei parametri del modello. In particolare, il termine di penalizzazione regolarizza le stime dei parametri di guessing verso una costante nota. L'ammontare di regolarizzazione viene determinato attraverso la validazione incrociata, e in alternativa mediante un approccio Bayesiano empirico. Il metodo è pratico ed efficace, e viene illustrato attraverso uno studio di simulazione e un esempio con dati reali.*

Key words: Cross-validation, Empirical Bayes, Guessing, Item response theory, Penalty

Michela Battauz
University of Udine - Department of Economics and Statistics, via Tomadini 30/A - Udine (Italy),
e-mail: michela.battauz@uniud.it

Ruggero Bellio
University of Udine - Department of Economics and Statistics, via Tomadini 30/A - Udine (Italy)
e-mail: ruggero.bellio@uniud.it

1 Introduction

The Three-Parameter Logistic (3PL) model is an Item Response Theory (IRT) model used with dichotomous responses. This model can be used for multiple-choice items, which are expected to have a non-zero probability of giving a correct response even at very low achievement levels. However, the parameters of this model are weekly identifiable [6] and the algorithms used for the maximization of the likelihood function frequently encounter convergence problems. A possible resolution is provided by regularization by means of penalized likelihood estimation [3, 10]. In particular, this paper studies the inclusion in the likelihood function of a penalty term that shrinks the guessing parameters towards a known constant. Despite a natural choice for this constant would be $1/k$, where k is the number of response options, we follow a data-driven approach for the selection of this value and of the amount of shrinkage. More specifically, the selection is performed by cross-validation [3]. An alternative route based on empirical Bayes estimation is also considered.

The paper is organized as follows. Section 2 introduces the model and the estimation methods, Section 3 shows an application to achievement data and Section 4 presents the results of a simulation study. Finally, Section 5 contains some concluding remarks.

2 Models and methods

In a 3PL model, the probability of a positive response to item j is given by

$$p_{ij} = \Pr(X_{ij} = 1|\theta_i; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\{a_j(\theta_i - b_j)\}}{1 + \exp\{a_j(\theta_i - b_j)\}}, \tag{1}$$

where θ_i is the ability of person i, a_j is the discrimination parameter, b_j is the difficulty parameter, and c_j is the guessing parameter. Here, $i = 1, \dots, n$ and $j = 1, \dots, J$, so that n is the sample size and J the number of items. A convenient parameterization of the model, suitable for estimation, is the following

$$p_{ij} = c_j + (1 - c_j)\frac{\exp(\beta_{1j} + \beta_{2j}\theta_i)}{1 + \exp(\beta_{1j} + \beta_{2j}\theta_i)}, \tag{2}$$

with

$$c_j = F(\beta_{3j}) = \frac{\exp(\beta_{3j})}{1 + \exp(\beta_{3j})}. \tag{3}$$

Let β be the vector containing all item parameters. The marginal maximum likelihood method [1], which is a commonly used estimation method of IRT models, requires the maximization of the following log likelihood function for β

$$\ell(\beta) = \sum_{i=1}^{n} \log \int_{\mathbb{R}} \prod_{j=1}^{J} p_{ij}^{x_{ij}} (1-p_{ij})^{1-x_{ij}} \, \phi(\theta_i) d\theta_i, \tag{4}$$

where $\phi(\cdot)$ denotes the standard normal p.d.f.

The penalized log likelihood function considered here is given by

$$\ell_p(\beta) = \ell(\beta) + J(\beta_3), \tag{5}$$

where β_3 is the vector containing all the guessing parameters, and $J(\beta_3)$ is a quadratic penalty term

$$J(\beta_3) = -\frac{1}{2\sigma^2} \sum_{j=1}^{J} (\beta_{3j} - \mu)^2, \tag{6}$$

proportional to the log p.d.f. of the normal distribution with mean μ and standard deviation σ. Note that to constrain the guessing parameters c_j towards a constant c, it is necessary to choose μ so that $c = F(\mu)$, with F defined in (3). Another option would take a penalty derived from a beta distribution for c_j.

2.1 Parameter tuning by cross-validation

A relevant issue is the choice of the tuning parameters μ and σ. The first proposal is to adopt a typical approach followed for regularized regression [10], namely

i. Estimate the item parameter for fixed values of the tuning parameters μ and σ;
ii. Select the tuning parameters by minimizing some error rate computed by cross-validation [3].

Step i. above is simply performed, for example by recourse to IRT software which allows for the introduction of a penalty term for 3PL models. Step ii. requires the definition of a suitable cross-validation error, here taken as minus the log likelihood of the validation set evaluated at the parameter estimates.

2.2 Empirical Bayes

An alternative method consists in treating μ and σ as parameter estimated by an empirical Bayes approach. In particular, we implement this method by treating β_3 as normal random effects, and then jointly estimate $(\beta_1, \beta_2, \mu, \sigma)$ after integrating them out from joint likelihood function of the data and β_3. Here the Laplace approximation is employed to carry out the latter integration, a strategy greatly simplified by the usage of automatic differentiation software [8].

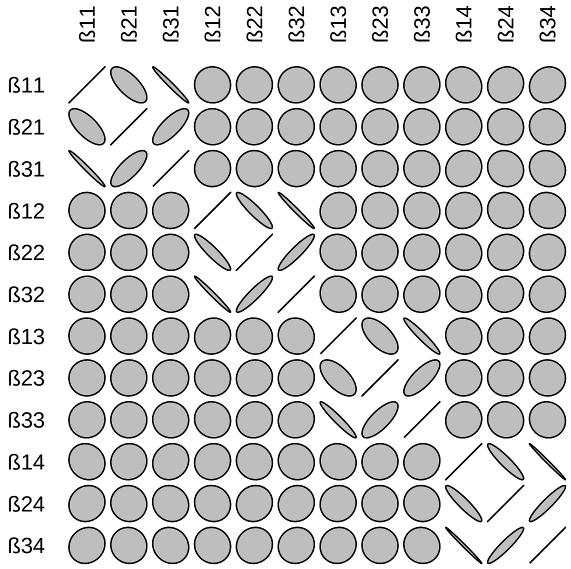
3 A real data example

The proposed methodology was applied to achievement data collected on students attending the third year of vocational high school in Italy. In particular, we used the mathematics test that was administered during the final exam. The sample is composed of 3843 students. Only multiple-choice items were included in the analysis. These were 14 items, all with four response options. All analyses were performed using the `R` software [7].

We started the analysis from ordinary (unpenalized) maximum likelihood estimation. Figure 1 visualizes the estimated correlation matrix among item parameter estimators for the first four items. The estimates were obtained with the `R` package `mirt` [2], and the `R` package `ellipse` [5] was used to obtain the plot.

The correlation matrix is nearly singular, since the estimated guessing parameter of each item $\widehat{\beta}_{3j}$ is negatively correlated with the estimated easiness parameter $\widehat{\beta}_{1j}$, and it is positively correlated with the estimated discrimination parameter $\widehat{\beta}_{2j}$. At times, such correlations are very high (in either direction), so that it is not surprising that parameter estimation may become cumbersome. Some sort of regularization is surely helpful.

Fig. 1 Estimated correlation matrix among item parameter estimators for the first four items.



The function mirt to fit an IRT model in the mirt package has an option parprior to introduce the penalty for the guessing parameters, which turned out to be very handy to apply the methodology endorsed here. The tuning parameters were selected by choosing among a set of candidate values. In particular, $F(\mu)$ was selected in the set $\{0.1, 0.15, 0.2, 0.25\}$, whereas for σ a set formed by 100 values between 0.2 and 5 was considered. The selection was performed by means of 10-fold cross-validation.

Fig. 2 Left: cross-validation error as a function of $1/\sigma$, for given μ. Right: estimates of guessing parameters, as a function of $1/\sigma$, for given μ.

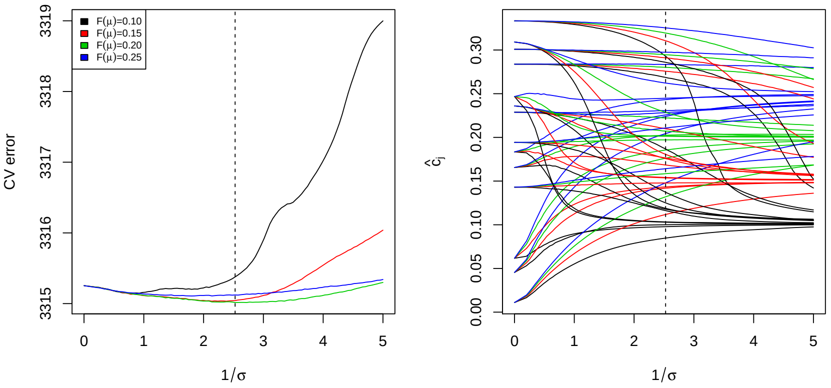


Figure 2 shows the results obtained for this data set. On the left panel of the figure, the cross-validation error is plotted against the reciprocal of σ. Thus, higher values on the x-axis correspond to larger amounts of shrinkage. The different colors refer to different values of $F(\mu)$. The vertical dashed lines indicate the point at which the cross-validation error is smallest. This corresponds to the values $\mu = F^{-1}(0.20)$ and $1/\sigma = 2.5$. The right panel of the figure shows the estimates of the guessing parameters at different levels of shrinkage. For all the values of μ, the smallest value of the cross-validation score is attained for σ values away from 0, pointing to the need of some shrinkage for the guessing parameters.

The estimation based on empirical Bayes has been carried out by means of the Template Model Builder (TMB) R package [9], which allows to define a C++ template used to estimate the random effects model of interest. The use of the package resulted in a rather efficient implementation, the key point being the explicit integration of the ability parameters within the C++ template by means of efficiently-coded Gaussian quadrature.

For the data set of interest, the empirical Bayes method provides estimated tuning parameters equal to $1/\widehat{\sigma} = 4.4$ and $\widehat{\mu} = F^{-1}(0.22)$, implying a higher amount of shrinkage for the guessing parameters with respect to cross validation. The estimates of the remaining item parameters were instead quite similar for the two penalized methods.

The overall message of this example is that, even for a large sample of subjects, the estimation of the guessing parameter is challenging, and penalized maximum likelihood estimation improves the inferential results. The need of some regularization may become more striking for smaller sample sizes, where numerical problems may hamper the estimation routines of IRT software.

4 A simulation study

A small-scale simulation study has been performed to assess the performance of the various methods. In particular, the focus was on a relatively small scale setting, with $n = 500$ subjects and $J = 30$ items. Two different choices for the guessing parameters were considered. In the first setting, we took all the c_j parameters as constant and equal to 0.2, whereas for the second setting we took as guessing parameters the estimates obtained from a large scale educational assessment, with values of c_j ranging between 0.04 and 0.33, with an average value of 0.16. The real data set was employed also to set the values for the other item parameters, for either setting.

Three different methods were considered, given by ordinary maximum likelihood estimation (MLE) as implemented in the `mirt` package, and the two penalized estimation methods with tuning parameters estimated by cross validation (CV) and empirical Bayes (EB), respectively. Table 1 and 2 summarize the result of 100 simulated datasets. In particular, the tables report the average over the three groups of parameters of Root Mean Squared Error (RMSE), the squared root of the average of squared bias (B), and the average of Median Absolute Error (MAE).

Table 1 Summary of simulation results for Setting 1 (equal guessing parameters).

Method	Easiness β_1			Discrimination β_2			Guessing c		
	RMSE	B	MAE	RMSE	B	MAE	RMSE	B	MAE
MLE	0.76	0.19	0.44	0.59	0.18	0.25	0.15	0.02	0.14
CV	0.23	0.03	0.14	0.26	0.03	0.15	0.04	0.01	0.01
EB	0.20	0.02	0.13	0.23	0.02	0.15	0.03	0.01	0.02

The tables points to some interesting results. First, the ordinary unpenalized estimation performs rather poorly in both settings, with unacceptably large variability for all the parameters, thus confirming that this method is essentially useless for datasets of this size. For Setting 1, the two penalized methods perform an excellent adjustment, with negligible bias and greatly reduced variability for all the parameters. This is the setting more relevant to the adopted penalty, so that the good performances are not surprising. For Setting 2, as expected, the two penal-

Table 2 Summary of simulation results for Setting 2 (different guessing parameters).

Method	Easiness β_1			Discrimination β_2			Guessing c		
	RMSE	B	MAE	RMSE	B	MAE	RMSE	B	MAE
MLE	0.70	0.21	0.37	0.53	0.18	0.23	0.14	0.04	0.10
CV	0.32	0.24	0.23	0.29	0.14	0.17	0.07	0.08	0.06
EB	0.30	0.28	0.25	0.24	0.16	0.17	0.08	0.07	0.07

ized methods perform less well, with bias similar to that of the MLE. At any rate, the two penalized estimators have overall better performances, and they represent a clear improvement over the MLE, with shrinkage offering a good level of protection against the large fluctuations affecting ordinary MLE. Finally, the differences between the two penalized methods are generally minor, though the method based on cross-validation seems to be slightly preferable in the most challenging setting.

5 Conclusion and ongoing research

This paper presents a procedure for the estimation of the 3PL model based on a penalized likelihood approach. The application shows that by penalizing the likelihood the error rate of prediction, assessed through cross-validation, is reduced. Even if IRT models are not usually fitted to predict new observations, the procedure can be viewed as a regularization method to obtain a model closer to the data generating process. The simulation study suggests that the penalized estimation represents a notable improvement over ordinary maximum likelihood when the guessing parameters are constant, while the improvement is less substantial when the guessing parameters exhibit large variation. A further finding is that the results obtained via cross validation are generally similar to those obtained by an empirical Bayes approach.

In the Bayesian literature [6], the prior for the guessing parameters is usually a distribution with mean equal to the reciprocal of the number of response options for the items. Despite this seems a sensible choice, the application of the proposed approach typically leads to the selection of a different value of the mean, a fact that seems worth mentioning.

The method introduced here is very practical, since it only requires the introduction of a simple penalty term in the ordinary log likelihood function for MML estimation of a 3PL model. More sophisticated approaches could be considered, such as model-based shrinkage aiming to reduce the bias of guessing parameter estimators [4]. Some investigation in this direction appears worth considering. The introduction of flexible assumptions for the ability parameters, which may be recommendable in some instances [11], appears instead more challenging.

It should be noted that this is a preliminary study. The use of regularization techniques for the estimation of the 3PL model is still under investigation by the authors. Future research will involve more extensive simulation studies to achieve a better understanding of the performance of the procedure, and the consideration of further regularization methods, targeting better inferential properties.

Acknowledgements

This work was supported by PRID 2017, University of Udine.

References

1. Bock, R. D., Aitkin, M.: Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika **46**, 443–459 (1981)
2. Chalmers, R.: mirt: A Multidimensional Item Response Theory Package for the R Environment. J. Stat. Softw. **48**, 1–29 (2012)
3. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning (2nd Edition). Springer, New York (2009)
4. Kosmidis, I., and Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. Electron. J. Statist. **4**, 1097–1112 (2010)
5. Murdoch, D. and Chow, E. D. ellipse: Functions for drawing ellipses and ellipse-like confidence regions. R package version 0.3-8 (2013)
6. Patz, R. J., Junker, B. W.: Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. J. Educ. Behav. Stat. **24**, 342–366 (1999)
7. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2018)
8. Skaug, H. J. (2002). Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. J. Comput. Graph. Statist. **11**, 458–470 (2002)
9. Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. J., Bell, B. TMB: Automatic differentiation and Laplace approximation. J. Stat. Softw. **70**, 1–21 (2016)
10. Tutz, G: Regression for Categorical Data. Cambridge University Press, Cambridge (2011)
11. Woods, C.: Estimating the latent density in unidimensional IRT to permit non-normality. In: Reise, S. P., and Revicki, D. A. (Eds.) Measuring Psychological Constructs: Advances in Model-Based Approaches, pp. 60–84. Routledge, New York (2015)

Statistical modelling and GAMLSS
Modellazione statistica attraverso i GAMLSS

Mikis D. Stasinopoulos and Robert A. Rigby and Fernanda De Bastiani

Abstract This paper reflects on the impact and future development of the generalized additive models for location scale and shape, GAMLSS. GAMLSS application is illustrated with an analysis of a real discrete data set.
Abstract *Questo lavoro discute le potenzialit attuali e future dei modelli GAMLSS (generalized additive models for location scale and shape) attraverso l'analisi di un dataset con risposta discreta e modellata attraverso un modello 'zero inflated beta negative binomial'.*

Key words: Beta negative binomial, Flexible.regression, Zero inflated beta negative binomial

1 Introduction

The generalized additive models for location scale and shape (GAMLSS) was introduced by [17]. It has been applied to a variety of different scientific fields including: actuarial science, [7], biology, [6], economics, [26], environment, [24], genomics, [10], finance, [9], [3], fisheries, food consumption, management science, [1], marine research, medicine, [18], meteorology, and vaccines. GAMLSS have also become standard for centile estimation, e.g. [25], [23], [15].

WHO, [27] and [28], use GAMLSS for centile estimation to produce growth charts for children. Their charts are used by more than 140 countries as the stan-

Mikis D. Stasinopoulos
London Metropolitan University, London, UK, e-mail: dmh.stasinopoulos@gmail.com

Robert A. Rigby
London Metropolitan University, London, UK

Fernanda De Bastiani
Universidade Federal de Pernambuco, Recife, PE, Brazil

dard charts monitoring the growth of children. The Global Lung Function Initiative (GLFI), [http://www.lungfunction.org, [16]] use GAMLSS to provide a unified worldwide approach to monitoring lung function, by obtaining centiles for lung function based on age and height.

Section 2 discussed GAMLSS within the general framework of statical modelling. Section 3 defines GAMLSS and show one of its application. In the conclusions we discuss the future of GAMLSS.

2 What is GAMLSS

GAMLSS was built around the basic principals of statistical modelling which can be summarized as: i) all models are *wrong* but some are useful (attributed to Gorge Box); ii) statistical modelling is an *iterative* process where, after fitting an initial model, assumptions are checked, followed by refitting models until an appropriate model is found; iii) a simple model is preferable to a more complicated one if both explain the data adequately (*Occam's Razor*) and iv) "no matter how beautiful your theory, no matter how clever you are or what your name is, if the model disagrees with the data, it is wrong"[1].

GAMLSS, is a general framework for *univariate* regression where we assume that the response variable depends on many explanatory variables. This dependance can be linear, non-linear or smooth non-parametric. For example, in the classical linear regression model (LM) the mean of the response variable is a linear function of the explanatory variables. In the generalized linear models (GLM), [14], a monotonic function of the mean, called the linear predictor, is a linear function of the explanatory variables. Non linear relationships between the response variable and the explanatory variables, within both LM and GLM, are dealt with by using nonparametric smoothing functions, giving additive models (AM) and generalized additive models (GAM) respectively. The generalized additive models (GAM) introduced by [5] and popularized by [29], have made the smoothing techniques within a regression framework available to a wide range of practitioners.

GAMLSS is an extension of the LM, GLM and GAM and has two main features. Firstly, in GAMLSS the assumed distribution can be any parametric distribution. Secondly, all the parameters (not only the location e.g. the mean) of the distribution can be modelled as linear or smooth functions of the explanatory variables. As a result the shape of the distribution of the response variable is allowed to vary according to the values of the explanatory variables. The GAMLSS models are an example of a "Beyond Mean Regression" model, [11]. and because of their explicit distributional assumption the response variable they also also part of the "distributional regression" modelling approach, [2].

GAMLSS allows a variety of smooth functions of explanatory variables including the ones which employ a quadratic penalty in the likelihood. The basic ideas of

[1] paraphrasing Richard Feynman famous quote

GAMLSS, have been implemented in **R** in a series of packages, [19], where residual based diagnostics facilities are also provided.

3 The GAMLSS framework

The response variable observations $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ are assumed independent with

$$\mathbf{Y} \sim D(\mu, \sigma, \nu, \tau) \tag{1}$$

where D is any (up to four distribution) distribution and where usually μ is the location parameter (e.g. the mean), σ is the scale parameter (e.g. the variance), ν and τ are the shape parameters (e.g. skewness and kurtosis). A GAMLSS model allows the modelling of all the parameters of the distribution as linear i.e. $\mathbf{X}_k \boldsymbol{\beta}_k$ or smooth term functions $s_{kj}(x_{kj})$, for example:

$$\begin{aligned}
g_1(\mu) &= \eta_1 = \mathbf{X}_1\beta_1 + \sum_{j=1}^{J_1} s_{1j}(\mathbf{x}_{1j}) \\
g_2(\sigma) &= \eta_2 = \mathbf{X}_2\beta_2 + \sum_{j=1}^{J_2} s_{2j}(\mathbf{x}_{2j}) \\
g_3(\sigma) &= \eta_3 = \mathbf{X}_3\beta_3 + \sum_{j=1}^{J_3} s_{3j}(\mathbf{x}_{3j}) \\
g_4(\sigma) &= \eta_4 = \mathbf{X}_4\beta_4 + \sum_{j=1}^{J_4} s_{4j}(\mathbf{x}_{4j})
\end{aligned} \tag{2}$$

where $\mathbf{X}_k$ is a known design matrix, $\beta_k = (\beta_{k1}, \ldots, \beta_{kJ'_k})^\top$ is a parameter vector of length J'_k, s_{kj} is a smooth nonparametric function of variable X_{kj} the $\mathbf{x}_{kj}$'s are vectors of length n, and $g_k(.)$ known monotonic link function relating a distribution parameter to a predictor η_k, for $k = 1, 2, 3, 4$ and $j = 1, \ldots, J_k$.

Stasinopoulos et al. [21] provide a variety of examples using GAMLSS. Here we use the example given by [20] pages 12-22. The data consists of 4406 observations, on the following variables: `visits`, number of physician office visits (the response variable), `hospital`, number of hospital stays, `health`, a factor indicating health status, `chronic`, number of chronic conditions, `gender`, a factor, `school`, number of years of education, `insurance`, a factor indicating whether the individual is covered by private insurance. The data are available from the **AER** package under the name NMES1988. There are more than 30 available discrete count distributions in the **gamlss** package. After a stepwise selection procedure the following model using a zero inflated beta negative binomial, *ZIBNB*, distribution was chosen:

$Y \sim \texttt{ZIBNB}(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$,

$$\begin{aligned}
\log(\hat{\mu}) = {} & 0.980 + 0.382\sqrt{\text{hospital}} + 0.332\sqrt{\text{chronic}} \\
& +0.025\text{school} + 0.255(\text{if health=poor}) \\
& -0.313(\text{if health=excellent}) - 0.112(\text{if gender=male}) \\
& +0.123(\text{if insurance=yes}) \\
\log(\hat{\sigma}) = {} & -1.7026 - 0.208\sqrt{\text{chronic}} + 0.394(\text{if health=poor}) \\
& -0.345(\text{if health=excellent}) + 0.197(\text{if gender=male}) \\
\log(\hat{\nu}) = {} & -2.679 + 0.966\sqrt{\text{hospital}} \\
\log[\hat{\tau}/(1-\hat{\tau})] = {} & -1.077 - 0.744\sqrt{\text{chronic}} - 1.546(\text{if insurance=yes}),
\end{aligned} \tag{3}$$

Figure 1 shows the worm plot and the rootogram of the fitted final model both indicating the the fit is adequate except for the extreme right tail.

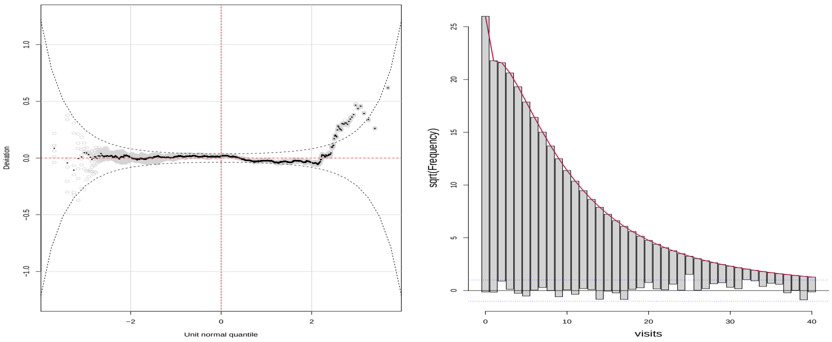


Fig. 1 Worm plot (a) and rootogram of the randomised quantile residuals for the final *ZIBNB* models. .

4 Conclusions

The GAMLSS models are especially useful for continuous response variables if they are negatively skew, highly positively skew, platykurtic, or leptokurtic. For discrete responses, if there exist over-dispersion or under-dispersion, excess or shortage of zero values and long tails. The GAMLSS models allows **any** parametric distribution for the response variable. The current implementation in the **gamlss** package in `R` allows the user to choose from more than 100 distributions with up to four parameters, allowing changes in modelling the location, the scale and shape of the distribution. A boosting and a Bayesian versions of GAMLSS exist in `R` packages , see [8, 13] and [22] respectively. There are alternative approaches to GAMLSS for quantile or centile estimation: for continuous response variable quantile regression, [12], can be used. In quantile regression there are less assumption than GAMLSS and for this reason more difficult to check the model. For mean (and variance) estimation the generalized estimation equation (GEE) [4] also can be used.

At this moment of time the following work is under way for the development and enhancement of GAMLSS:

- A second book on GAMLSS with the title "Distributions for Modelling Location, Scale and Shape: Using GAMLSS in R" is in its final draft.
- Robust methods of GAMLSS model are developed.
- Alternative model selection techniques are explored.
- Time series modelling techniques within GAMLSS are investigated.

The GAMLSS model provide a very general framework for regression type of modelling but it flexibility it is also its burden. More automated procedures would help its spread and its popularity

References

1. Budge, S., Ingolfsson, A., and Zerom, D. Empirical analysis of ambulance travel times: The case of calgary emergency medical services. Management Sci- ence, **56**(4):716-723. (2010).
2. Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. Regression: Models, Methods and Applications. Springer-Verlag, New York. (2013).
3. Giraud, G. and Kockerols, T. Making the european banking union macro- economically resilient: Cost of non-europe report. Report to the European Parliament. (2015).
4. Hardin, J. W. and Hilbe, J. M. Generalized Estimating Equations. Chapman and Hall/CRC. (2003).
5. Hastie, T. J. and Tibshirani, R. J. Generalized additive models. Chapman and Hall, London. (1990).
6. Hawkins, E., Fricker, T. E., Challinor, A. J., Ferro, C. A., Ho, C. K., and Osborne, T. M. Increasing influence of heat stress on french maize yields from the 1960s to the 2030s. Global change biology, **19**(3):937-947. (2013).
7. Heller, G., Stasinopoulos, D., Rigby, R., and De Jong, P. Mean and dis- persion modelling for policy claims costs. Scandinavian Actuarial Journal, 2007(4):281-292. (2007).
8. Hofner, B., Mayr, A., Fenske, N., and Schmid, M. gamboostLSS: Boosting Methods for GAMLSS Models. R package version 2.0-0. (2017).
9. International Monetary Fund Stress Testing, Technical Note. Country Report No. 15/173. (2015).
10. Khondoker, M. R., Glasbey, C., and Worton, B. A comparison of parametric and nonparametric methods for normalising cdna microarray data. Biometrical Journal, **49**(6):815-823. (2007).
11. Kneib, T. Beyond mean regression. Statistical Modelling, **13**:275-303. (2013).
12. Koenker, R. Quantile regression: 40 years on. Annual Review of Economics, **9**:155-176. (2017).
13. Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. Generalized additive models for location, scale and shape for high dimensional data, a flexible approach based on boosting. J. R. Statist. Soc. Series C, **61**:403-427. (2012).
14. Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. Journal of the Royal Statistical Society, Series A, **135**:370-384. (1972).

15. Neuhauser, H. K., Thamm, M., Ellert, U., Hense, H. W., and Rosario, A. S. Blood pressure percentiles by age and height from nonoverweight children and adolescents in germany. Pediatrics, pages peds-2010. (2011).
16. Quanjer, P. H., Stanojevic, S., Cole, T. J., Baur, X., Hall, G. L., Culver, B. H., Enright, P. L., Hankinson, J. L., Ip, M. S. M., Zheng, J., et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung func- tion 2012 equations. European Respiratory Journal, **40**(6):1324-1343. (2012).
17. Rigby, R. A. and Stasinopoulos, D. M. Generalized additive models for location, scale and shape, (with discussion). Applied Statistics, **54**:507-554. (2005).
18. Rodrigues, J., de Castro, M., Cancho, V., and Balakrishnan, N. Com-poisson cure rate survival models and an application to a cutaneous melanoma data. Journal of Statistical Planning and Inference, **139**(10):3605-3611. (2009).
19. Stasinopoulos, D. M. and Rigby, R. A. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, **23**(7):1-46. (2007).
20. Stasinopoulos, D. M., Rigby, R. A., and F., D. B. Gamlss: A distributional regression approach. Statistical Modelling, **18**:1-26. (2018).
21. Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. Flexible Regression and Smoothing: Using GAMLSS in R. Chapman and Hall, Boca Raton. (2017).
22. Umlauf, N., Klein, N., and Zeileis, A. BAMLSS: Bayesian additive models for location, scale and shape (and beyond). Journal of Computational and Graphical Statistics. (2017).
23. Villar, J., Cheikh, I. L., Victora, C. G., Ohuma, E. O., Bertino, E., Altman, D., Lambert, A., Papageorghiou, A. T., Carvalho, M., Jaffer, Y. A., Gravett, M. G., Purwar, M., Frederick, I., Noble, A. J., Pang, R. Barros, F. C., Chumlea, C. Bhutta, Z. A., and Kennedy, S. H. International standards for newborn weight, length, and head circumference by gestational age and sex: The newborn cross-sectional study of the intergrowth-21st project. The Lancet., **384**(9946):857-868. (2014).
24. Villarini, G., Smith, J., Serinaldi, F., Bales, J., Bates, P., and Krajewski, W. Flood frequency analysis for nonstationary annual peak records in an urban drainage basin. Advances in Water Resources, **32**(8):1255-1266. (2009).
25. Visser, G. H., Eilers, P. H. C., Elferink-Stinkens, P. M., Merkus, H. M., and Wit, J. M. New dutch reference curves for birthweight by gestational age. Early human development, **85**(12):737-744. (2009).
26. Voudouris, V., Ayres, R., Serrenho, A. C., and Kiose, D. The economic growth enigma revisited: The EU-15 since the 1970s. Energy Policy. (2015).
27. WHO, M. G. R. S. G. WHO Child Growth Standards: Head circumference- for-age, arm circumference-for-age, triceps circumference-for-age and subscapular skinford-for-age: Methods and development. Geneva: World Health Organization. (2007).
28. WHO, M. G. R. S. G. WHO Child Growth Standards: Growth velocity based on weight, length and head circumference: Methods and development. Geneva: World Health Organization. (2009).
29. Wood, S. N. Generalized additive models. An introduction with R. Chapman and Hall. (2017).

Young Contributions to Statistical Learning

Introducing spatio-temporal dependence in clustering: from a parametric to a nonparametric approach.

Introduzione di dipendenza spazio-temporale nel problema di clustering: da un approccio parametrico a un approccio non-parametrico

Clara Grazian, Gianluca Mastrantonio and Enrico Bibbona

Abstract A huge literature about clustering spatial time data exists. The problem has been studied both in a parametric and in a nonparametric setting. There are several problems in defining a proper clustering procedure, depending on the type of relationship between the clusters. From a parametric point of view, a classic approach is to introduce mixture models and studying the posterior distribution of the mixture weights. We propose a mixture model where the mixing probabilities are time specific and are assumed to follow a Logistic-Normal distribution. We introduce dependence between the vectors of mixing probabilities by means of a Gaussian processes representation. We briefly propose a nonparametric extension of this approach.
Abstract *Esiste un'estesa letteratura riguardante problemi di clustering temporale e spaziale. Il problema è stato studiato sia in ambito parametrico che nonparametrico. I problemi principali delle procedure di clustering spazio-temporali dipendono dal tipo di dipendenza tra i cluster stessi. Da un punto di vista parametrico, l'approccio classico è quello di assumere un modello mistura e studiare la distribuzione a posteriori dei pesi della mistura. In questo articolo proponiamo ancora un modello mistura dove i pesi dipendono dal tempo e dove si assume che seguano un modello logistico-normale. La dipendenza temporale tra i cluster è introdotta attraverso una rappresentazione in base a processi gaussiani. Nell'articolo proponiamo brevemente anche un'estensione nonparametrica dell'approccio.*

Key words: Gaussian processes, logitN, temporal clustering

Clara Grazian
Clara Grazian, University of Oxford, e-mail: `clara.grazian@ndm.ox.ac.uk`

Gianluca Mastrantonio
Politecnico di Torino e-mail: `gianluca.mastrantonio@polito.it`

Enrico Bibbona
Politecnico di Torino e-mail: `enrico.bibbona@polito.it`

1 Introduction

There is more and more interest in spatial data, i.e. data where the response variable is measured at spatial locations or data where the response variable is defined as a set of spatial coordinates. This is due to the increased ability to store and collect this type of data.

A central problem in the analysis of spatial data is spatial clustering, which groups similar spatial objects into classes. Standard applications are the identification of land areas for usage purposes in agricultural sciences or weather patterns in environmental sciences. The goal of spatial clustering may be multiple, focusing on the study of the characteristics of each cluster and on a better understanding and description of the data and, ultimately, influencing policies in public health and environment. Due to its huge usefulness in applied sciences, spatial clustering has been a very active subject, with many contributions from different fields.

In this work, we will focus on the problem of modelling spatial coordinates (or transformation of them) and, in particular, clustering them through time. Direct applications of this may be seen in the modelling of wind directions [16], ocean streams [8], identification of three-dimensional protein structures [12] and animal movements [3].

The problem of clustering in this settings relates to the description of structural changes in the time series. For instance, it is generally assumed that the animal behaviour changes according to a natural cycle in the observational period, for example the resting/feeding cycle, or to out-of-ordinary situations, such as the assault of a predator or changement in human activities.

The joint distribution of the coordinates, or of an appropriate transformation of them, can be seen as a mixture process where the mixture components are the different behaviours or regimes. It is generally assumed that the switching between regimes is temporally structured [17], and sometime also spatially, as in [2], often ruled by a non-observed Markov process leading to the class of hidden Markov models (HMMs) [25].

In this paper, we propose a mixture-type model, as the hidden Markov model, but with a higher level of flexibility given by the assumption that the vector of probabilities is marginally distributed as a Logistic-Normal model (*LogitN*) [1] and the structured temporal dependence is induced via a coregionalization over a multivariate Gaussian process [10]. This is not the first proposal where a structured dependence is introduced over vectors *LogitN* distributed, however, as we will show, our proposal is more general since it focuses on the dependence structure of the probabilities vectors rather than that of the Gaussian process, as in [22] and [4].

The rest of the paper is organized as follows: Section 2 provides the notation and gives some preliminaries, Section 3 describes the proposed method; Section 4 focuses on a real application on data of animal movement. Section 5 concludes the paper.

2 Notation and preliminaries

Suppose the data are indexed by temporal indices $(t_1,\dots,t_T)' \equiv \mathscr{T}$, assuming that t_1 is the starting observational time and t_T the ending observational time, by allowing that $t_i - t_{i-1} = c_i$ which may vary depending on i.

Let, then, $\mathbf{s}_{t_i} = (s_{t_i,1}, s_{t_i,2})' \in \mathbb{R}^2$ be spatial location at time t_i with $\mathbf{s} = (\mathbf{s}_{t_1},\dots,\mathbf{s}_{t_T})'$.

A standard transformation considered when analysing spatial coordinates is looking at the projections of $\mathbf{s}_{t_i}$ on the x- and y- axes of the cartesian coordinates system centred on $\mathbf{s}_{t_{i-1}}$, say $\mathbf{r}_{t_i} = (r_{t_i,1}, r_{t_i,2}) \in \mathbb{R}^2$ and then defining

$$\mathbf{y}_{t_i} = \frac{\mathbf{r}_{t_i}}{d(t_{i+1},t_i)}.$$

The variable $\mathbf{y}_{t_i}$ contains all the sufficient information to recover the trajectory without loosing any significant property. In particular, the sign of $y_{t_i,1}$ provides information about the fact that the movement is on the same direction of the previous one, while the sign of $y_{t_i,2}$ indicates if it turns to the right or to the left. Moreover, $||\mathbf{y}_{t_i}||$ represents the step-length and $\theta_{t_i} = \text{atan2}(y_{t_i,2}, y_{t_i,1})$ the turning angle of the movement, respectively.

A standard clustering methodology is to introduce information indicating the cluster membership as a latent variable $\mathbf{z} = \{z_t\}_{t\in\mathscr{T}}$, with $z_t \in \{1,2,\dots,K\} \equiv \mathscr{K}$, and K the total number of clusters.

The data are assumed to come from a mixture-type model based on bivariate Gaussian densities:

$$f(\mathbf{y}|\mathbf{z}\{\xi_k,\Omega_k\}) = \prod_{t\in\mathscr{T}} f(\mathbf{y}_t|\xi_{z_t},\Omega_{z_t})$$

where

$$\mathbf{y}_t|\xi_{z_t},\Omega_{z_t} \sim N_2(\xi_{z_t},\Omega_{z_t})$$

i.e. given the latent variables $\mathbf{z}_t$, the observations $\mathbf{y}_t$ are independent. The hidden Markov models (HMMs) [5] assume the latent variables follow the Markov rule

$$\Pr(z_{t_i} \mid z_{t_1},\cdots,z_{t_{i-1}}) = \Pr(z_{t_i} \mid z_{t_{i-1}})$$

and, in particular,

$$z_{t_i}|z_{t_{i-1}},\{\pi_{k,k'}\}_{k,k'\in\mathscr{K}} \sim \sum_{k\in\mathscr{K}} \pi_{z_{t_{i-1}},k}\delta_k.$$

Although HMMs are widely used ans easy to implement, the Markov structure may be too restrictive. For instance, the assumption that the probabilities of switching cluster is fixed and independent from time is difficult to accept in all the contexts. In some works, see for example [19], [17], [13] and [11], problems like these are tackled using covariates that model probabilities, but not always these are available and may not be enough for describing the complexity of reality. [18] proposes a mixture model where the latent spatial process is allowed to evolve dynamically over time.

We proposed to consider a more complex model where the mixing probabilities follow a *LogitN* model; this has been proposed by [1] as a distribution for independent compositional data, i.e. vectors of positive proportions with a constant sum. Beyond the obvious fact that it can model data which are positive and less than a constant, the constant-sum constraint induces some more insidious characteristics which will be described in the next Section.

3 The model

We propose to introduce a dependence on time of the vector of probabilities, instead of using a HMM

$$z_t|\pi_t \sim \sum_{k=1}^{K} \pi_{t,k}\delta_k,$$

where $0 \leq \pi_{t,k} \leq 1$, $\sum_{k=1}^{K} \pi_{t,k} = 1$. The probability vector π_t is then defined with the logistic transformation of real valued variables

$$\pi_{t,k} = \frac{e^{\omega_{t,k}}}{1+\sum_{j=1}^{K} e^{\omega_{t,j}}}, k = 1,\dots,K \tag{1}$$

with the last element defined as

$$\pi_{t,K} = \frac{1}{1+\sum_{j=1}^{K-1} e^{\omega_{t,j}}}.$$

It has to be noticed that adding a constant c to each $\omega_{t,k}$ produces the same vector of probabilities, and then an identifiability constraint is needed; without loss of generality, we set to zero the K^{th}.

In this way, it is possible to model the latent variable ω_t. In particular, we propose to define it as

$$\omega_t = (\mathbf{I}_{K-1} \otimes \mathbf{X}_t)\beta + \mathbf{A}\eta_t \tag{2}$$

where $\mathbf{I}_d$ is the identity matrix of dimension $d \times d$, $\mathbf{X}_t$ is a set of p covariates changing over time and β the corresponding coefficients such that each $\omega_{t,k}$ shares the same set of covariates but different coefficient β. $\mathbf{A}$ is the term introducing dependence among the vectors $\omega_{.,k} = \{\omega_{t,k}\}_{t\in\mathscr{T}}$, while η_t is the error term such that it is defined as a Gaussian process, i.e. $\eta_{.k} \sim GP(\mathbf{0}, C_k(\cdot;\psi_k))$ where C_k models the structured dependence. Consequently, the covariance between ω_t and $\omega_{t'}$ is given by $\Sigma_{t,t'} = \mathbf{A}\mathbf{C}_{\eta,|t-t'|}\mathbf{A}'$.

We say that marginally the vector of probabilities π_t follows a logit-normal model [1], i.e. $\pi_t \sim LogitN(\mu_t, \Sigma)$, where $\mu_t = (\mathbf{I}_{K-1} \otimes \mathbf{X}_t)\beta$ and $\Sigma = \mathbf{A}\mathbf{A}'$.

The nature of a compositional vector, in particular its property to sum to a constant term, implies a lack of interpretability since the correlations are not completely free to vary in $(-1,1)$. We propose a different parametrisation to deal with this problem.

Instead of modelling ω_t, we propose to introduce another variable

$$\gamma_t = (\mathbf{I}_K \otimes \mathbf{X}_t)\beta + \mathbf{A}^* \eta_t$$

in similar way than ω_t, with the difference that all the K variables are defined, i.e. $\mathbf{I}_K$ is the identity matrix of dimension $K \times K$, $\mathbf{A}$ is a $K \times K$ matrix introducing dependence among the $\gamma_{.,k} = \{\gamma_{t,k}\}_{t \in \mathscr{T}}$ while η_t is the error term such that $\eta_{.k} \sim GP(\mathbf{0}, C_k^*(\cdot;\psi_k^*))$. Then, γ_t has covariance matrix given by

$$\Sigma^* = \mathbf{A}^*(\mathbf{A}^*)'.$$

Therefore, it is possible to define a different vector of probabilities with elements

$$\pi_{t,k}^* = \frac{\exp(\gamma_{t,k})}{\sum_{j=1}^{K} \exp(\gamma_{t,j})}.$$

It is evident, however, that the model is not identifiable. Notice that π_t^* follows a marginal distribution different from π_t, with a different temporal dependence.

However, if we let

$$\omega_{t,k} \coloneqq \gamma_{t,k} - \gamma_{t,K}$$

we create a link between the two parametrizations that induces

$$\mathbf{A} = [\mathbf{A}^*]_{1:(K-1),1:K} - [\mathbf{A}^*]_{K,1:K},$$

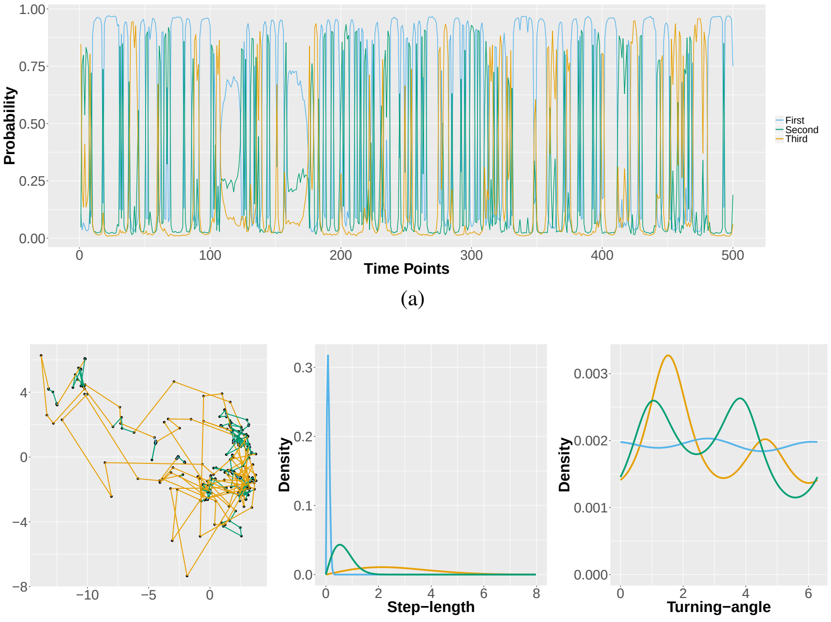
and, consequently,

$$[\Sigma^*]_{1:(K-1),1:(K-1)} + \mathbf{1}_{K-1}[\Sigma^*]_{K,K}\mathbf{1}'_{K-1} - [\Sigma^*]_{1:(K-1),K}\mathbf{1}'_{K-1} - \mathbf{1}_{K-1}[\Sigma^*]'_{1:(K-1),K}.$$

Given this link, it follows that $\pi_{t,k} = \pi_{t,k}^*$ for any $k = 1,\cdots,K$ and $t = 1,\cdots,T$, with the advantage to be able to define a covariance structure on the variables γ_t which induces the desired covariance structure on the corresponding π_t.

There are several interesting issues to highlight at this point. First, our proposed model given in (2) generalizes most of the models available in the literature where transformations as in (1) are used to define probabilities vectors. For instance, the proposal of [15] is obtained by assuming $\eta_t \equiv \mathbf{0}_{K-1}$. The model of [18] is obtained letting η_t be a spatio-temporal process with autoregressive temporal increments and a diagonal $\mathbf{A}$. We can also reduce to the proposals of [24], [14] and [21]. On the other hand, models using the cokriging, such as the ones of [20], can not be expressed with our formulation. However, one may notice that the complexity of our approach is reduced with respect to the cokriging.

Secondly, computational issues often arise for models based on Gaussian processes. We make use of the approach proposed by [7], i.e. a scalable nearest-



(a)

(b) (c) (d)

Fig. 1: Real data: (a) probabilities time series; (b) observed trajectory; (c) posterior estimates of step-length; (d) posterior estimates of turning-angle

neighbour Gaussian process (NNGP) which may be seen as a hierarchical sparse prior and allows for efficient MCMC algorithms to be performed without storing or decomposing large covariance matrices. While [7] empirically shows that 25 neighbours are needed to obtain an approximation close to the complete model, the temporal nature of the data analysed in this paper allow us to use just one neighbour, with a even highest level of saving of computational time.

4 Real Data

Data on 6 free-ranging Maremma sheepdogs positions are recorded by tracking collars every 30 minutes. The behaviour of the dogs is unknown because there is minimal supervision by their owners and the animals are allowed to range freely.

We select a time series of 500 points of one dog and characterize the hidden behaviours using the model proposed above; in Figure 1 (b) we show the observed trajectories, i.e. the observed coordinates.

We estimate model with values of K in $[2,3,\dots,10]$ and, using DIC_5 [6], we select the best model as the one with three components. The three behaviours can be easily characterized looking at the distribution of step-length and turning-angle, Figure 1 (b) and (c). In the first behaviour, the dog has low speed and the distribution of the angle is almost circular uniform, i.e. it moves randomly. In the second one, the speed increases and the direction has two modes with same height, one at $\pi/2$ and one at $3\pi/2$, meaning that it changes direction clockwise and anticlockwise with the same probability. In the last one, the speed is high and it changes direction mainly anticlockwise. From Figure 1 (a) we can see the temporal series of the probabilities and we appreciate that behaviour one is the one that has more temporal stabilities, i.e. it has high values for a long time-period. In Figure 1 (b) we see where the behaviours are spatially localized. The third one occurs in all spatial domain, while the other two behaviours are on the top left (where the house of the owner is localized) and the right part of the map (where there is the livestock).

5 Conclusion

In this work, we propose a model to perform clustering for spatial data, based on a new parametrization of the logit-normal process, with parameters allowing an easier interpretability.

However, the possibility to use this model derives from the knowledge of the exact number of clusters. The nonparametric extension of a mixture model is the Dirichlet process (DP) [9], a stochastic process defined over a measurable space whose random paths are probability measures with probability one. The hierarchical Dirichlet process (HDP), proposed by [23], is an extension of the DP that makes possible to have processes sharing the same set of atoms. In the original construction of the HDP the random vectors $\pi_{j.}$ and $\pi_{k.}$ for any $j \neq k$ are independent. The nonparametric extension of the work proposed in this paper, which will be the subject of further research, focuses on how to introduce dependence in hierarchical Dirichlet processes.

References

1. Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, UK, UK: Chapman & Hall, Ltd.
2. Blackwell, P. G. (2003). Bayesian inference for Markov processes with diffusion and discrete components. *Biometrika*, 90(3): 613–627.
3. Brownlee, J. (1912). The mathematical theory of random migration and epidemic distribution. *Proceedings of the Royal Society of Edinburgh*, 31: 262–289.
4. Brunsdon, T. M. and Smith, T. (1998). The time series analysis of compositional data. *Journal of Official Statistics*, 14(3): 237.
5. Cappé, O., Moulines, E., and Rydén, T. (2009). Inference in Hidden Markov Models. *Proceedings of EUSFLAT Conference.*
6. Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., Futurs, I., and Rhône-alpes, I. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 4: 651–674.

7. Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association*, 111(514): 800–812.
8. Ewing, J. A. (1990). Wind, wave and current data for the design of ships and offshore structures. *Marine Structures*, 3(6): 421–459.
9. Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2): 209–230.
10. Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Chapman and Hall.
11. Gupta, M., Qu, P., and Ibrahim, J. G. (2007). A temporal hidden Markov regression model for the analysis of gene regulatory networks. *Biostatistics*, 8(4): 805–820.
12. Hughes, G. (2007). Multivariate and time series models for circular data with applications to protein conformational angles. *Doctoral dissertation, University of Leeds*.
13. Lagona, F., Maruotti, A., and Picone, M. (2011). A non-homogeneous hidden Markov model for the analysis of multi-pollutant exceedances data. In *Hidden Markov Models, Theory and Applications*. InTech.
14. Martins, A. B. T., Bonat, W. H., and Ribeiro Jr, P. J. (2016). Likelihood analysis for a class of spatial geostatistical compositional models. *Spatial Statistics*, 17: 121 – 130.
15. Maruotti, A., Bulla, J., Lagona, F., Picone, M., and Martella, F. (to appear). Dynamic mixture of factor analyzers to characterize multivariate air pollutant exposures. *Annals of Applied Statistics*, 11(3), 1617–1648.
16. Masseran, N., Razali, A. M., Ibrahim, K., and Latif, M. T. (2013). Fitting a mixture of von Mises distributions in order to model data on wind direction in Peninsular Malaysia. *Energy Conversion and Management*, 72: 94–102.
17. Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., and Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, 85(9): 2436–2445.
18. Paci, L. and Finazzi, F. (2017). Dynamic model-based clustering for spatio-temporal data. *Statistics and Computing*, 1–16.
19. Patterson, T. A., Basson, M., Bravington, M. V., and Gunn, J. S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, 78(6): 1113–1123.
20. Pawlowsky, V. and Burger, H. (1992). Spatial structure analysis of regionalized compositions. *Mathematical Geology*, 24(6): 675–691.
21. Pirzamanbein, B., Lindström, J., Poska, A., and Gaillard, M.-J. (2018). Modelling Spatial Compositional Data: Reconstructions of past land cover and uncertainties. *Spatial Statistics*.
22. Quintana, J. M. and West, M. (1988). Time series analysis of compositional data. *Bayesian Statistics*, 3: 747–756.
23. Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476): 1566–1581.
24. Tjelmeland, H. and Lund, K. V. (2003). Bayesian modelling of spatial compositional data. *Journal of Applied Statistics*, 30(1): 87–100.
25. Zucchini, W. and MacDonald, I. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Bayesian inference for hidden Markov models via duality and approximate filtering distributions

Inferenza bayesiana per modelli di Markov nascosti via dualità e filtraggio approssimato

Guillaume Kon Kam King, Omiros Papaspiliopoulos and Matteo Ruggiero

Abstract Filtering hidden Markov models is analytically tractable only for a handful of models (e.g. Baum-Welch and Kalman filters). Recently, Papaspiliopoulos & Ruggiero (2014) proposed another analytical approach exploiting a duality relation between the hidden process and an auxiliary process, called dual and related to the time reversal of the former. With this approach, the filtering distributions are obtained as a recursive series of finite mixtures. Here, we study the computational effort required to implement this strategy in the case of two hidden Markov models, the Cox-Ingersoll-Ross process and the K-dimensional Wright-Fisher process, and examine several natural and very efficient approximation strategies.

Abstract *Il filtraggio dei modelli di Markov nascosti è un problema trattabile analiticamente solo per un numero limitato di modelli (p.es. i filtri di Baum-Welch e Kalman). Papaspiliopoulos & Ruggiero (2014) hanno proposto un nuovo approccio che sfrutta una relazione di dualità tra il processo nascosto ed un processo ausiliario, detto duale e legato alla reversibilità del segnale. Con questo approccio la soluzione del problema di filtraggio assume la forma di una mistura finita valutata mediante una ricorsione. Qui studieremo il costo computazionale richiesto per implementare tale strategia nel caso di due modelli di Markov nascosti, i cui segnali evolvono come un processo di Cox-Ingersoll-Ross e come una diffusione K-dimensionale di Wright-Fisher; analizzeremo diverse strategie di approssimazione.*

Key words: Optimal filtering, duality, Wright-Fisher, hidden Markov models

Guillaume Kon Kam King
University of Torino and Collegio Carlo Alberto, e-mail: guillaume.konkamking@unito.it

Omiros Papaspiliopoulos
ICREA - Pompeu Fabra University, Barcelona e-mail: omiros.papaspiliopoulos@upf.edu

Matteo Ruggiero,
University of Torino and Collegio Carlo Alberto, e-mail: matteo.ruggiero@unito.it

1 Introduction to optimal filtering using a dual process

Consider a hidden stochastic process and some noisy observations of this process. As new data arrives, obtaining the distribution for the last hidden state given all the values observed previously is called filtering the hidden process. Let the series $\{Y_k, 0 \geq k \geq n\}$ be the sequence of observations, denoted $Y_{0:n}$ for $Y \in \mathscr{Y}$, and let the Markov chain $\{X_k, 0 \geq k \geq n\}$, similarly denoted $X_{0:n}$, be the unobserved stochastic process. We assume $X_{0:n}$ to be the discrete-time sampling of a homogeneous continuous-time Markov process X_t. We also assume that X_t has state-space $\mathscr{X}$, transition kernel $P_t(x, dx')$ and initial distribution $\nu(\mathrm{d}x)$. The observations relate to the hidden signal by means of conditional distributions assumed to be given by the kernel $F(x, \mathrm{d}y)$ and we let $F(x, \mathrm{d}y) = f_x(y)\mu(\mathrm{d}y)$ for some measure $\mu(\mathrm{d}y)$. The filtering distributions, which are the target of inference, are $\mathscr{L}(X_n|Y_{0:n})$, denoted $\nu_n(\mathrm{d}x)$. Define now an update and prediction operator acting on probability measures ν:

$$\text{update:} \qquad \phi_y(\nu)(\mathrm{d}x) = \frac{f_x(y)\nu(\mathrm{d}x)}{p_\nu(y)}, \qquad \text{with } p_\nu(y) = \int_\chi f_x(y)\nu(\mathrm{d}x) \quad (1)$$

$$\text{prediction:} \qquad \psi_t(\nu)(\mathrm{d}x') = \int_\chi \nu(\mathrm{d}x)P_t(x, \mathrm{d}x') \quad (2)$$

Then, the filtering distributions can be obtained by repeated applications of the update and prediction operators, as the recursion: $\nu_0 = \phi_{Y_0}(\nu)$ and $\forall n > 0, \nu_n = \phi_{Y_n}(\psi_{t_n - t_{t-1}}(\nu_{n-1}))$ (see for instance [1]). An explicit solution to the filtering problem is seldom available, except in two notorious cases: unobserved Markov chains with a discrete state-space, and Gaussian unobserved Markov chains with Gaussian conditional distribution. [2] extended the class of models for which an explicit solution is available by exploiting a duality relation between the unobserved Markov chain and a pure death stochastic process. In order to describe this, assume that Θ_t is a deterministic process and that $r : \Theta \to \Theta$ is such that the differential equation: $\mathrm{d}\Theta_t/\mathrm{d}t = r(\Theta_t)$ with $\Theta_0 = \theta_0$ has a unique solution for all θ_0. Let $\lambda : \mathbb{Z}_+ \to \mathbb{R}_+$ be an increasing function, $\rho : \Theta \to \mathbb{R}_+$ be a continuous function, and consider a two-component Markov process (M_t, Θ_t) with state-space $\mathscr{M} \times \Theta$, where Θ_t evolves autonomously according to the previous differential equation, and when at $(M_t, \Theta_t) = (\mathbf{m}, \theta)$, the process jumps down to state $(\mathbf{m} - \mathbf{e}_j, \theta)$ with instantaneous rate $\lambda(|\mathbf{m}|)\rho(\theta)m_j$. We say that (M_t, Θ_t) is *dual* to X_t with respect to a family of functions h, e.g.

$$\mathbb{E}^x\left[h(X_t, \mathbf{m}, \theta)\right] = \mathbb{E}^{\mathbf{m},\theta}\left[h(x, M_t, \Theta_t)\right], \quad \forall x \in \chi, \mathbf{m} \in \mathscr{M}, \theta \in \Theta, t \geq 0.$$

where $\mathbb{E}^x\left[f(X_t)\right] = \mathbb{E}\left[f(X_t)|X_0 = x\right] = \int_\chi f(x')P_t(x, \mathrm{d}x')$ and the duality functions are such that $h : \chi \times \mathscr{M} \times \Theta \to \mathbb{R}_+$, $\Theta \subseteq \mathbb{R}^l$. The dual process (M_t, Θ_t) is separated into a deterministic component Θ_t and a pure death process M_t, whose rates are subordinated to the deterministic process. The transition probabilities of the dual process are denoted $p_{\mathbf{m},\mathbf{n}}(t,\theta) = \mathbb{P}\left[M_t = \mathbf{n} | M_0 = \mathbf{m}, \Theta_0 = \theta\right], \forall\, \mathbf{n}, \mathbf{m} \in \mathscr{M}^2, \mathbf{n} \leq \mathbf{m}$. The duality property is key to the computability of the filters, as it allows to replace the expectation with respect to realisations of the original stochastic process in the prediction operation (Eq. (2)) by an expectation over realisations of the pure death component of the dual process, which involves finite sums.

The transition probabilities can be found by exploiting the duality relation ([2]):

$$p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t,\theta) = \gamma_{|\mathbf{m}|,|\mathbf{i}|} C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t) p(\mathbf{i}; \mathbf{m}, |\mathbf{i}|) \tag{3}$$

with:

$$\gamma_{|\mathbf{m}|,|\mathbf{i}|} = \left(\prod_{h=0}^{|\mathbf{i}|-1} \lambda_{|\mathbf{m}|-h} \right), \text{ and } C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t) = (-1)^{|\mathbf{i}|} \sum_{k=0}^{|\mathbf{i}|} \frac{e^{-\lambda_{|\mathbf{m}|-k} t}}{\prod_{0 \leq h \leq |\mathbf{i}|, h \neq k} (\lambda_{|\mathbf{m}|-k} - \lambda_{|\mathbf{m}|-h})} \tag{4}$$

and $p(\mathbf{i}; \mathbf{m}, |\mathbf{i}|)$ is the hypergeometric probability mass function.

We also define the following notion of *conjugacy*, by assuming that $\mathscr{F}_0 = \{h(x,\mathbf{m},\theta)\pi(\mathrm{d}x), \mathbf{m} \in \mathscr{M}, \theta \in \Theta\}$ is a family of probability measures such that there exist functions $t : \mathscr{Y} \times \mathscr{M} \to \mathscr{M}$ and $T : \mathscr{Y} \times \Theta \to \Theta$ with $\mathbf{m} \to t(y,\mathbf{m})$ increasing and such that $\phi_y(h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)) = h(x, t(y,\mathbf{m}), T(y,\theta))\pi(\mathrm{d}x)$. The filtering algorithm proposed in [2] can be summarised by the two following relations. For the family of finite mixtures $\bar{\mathscr{F}}$ $\{\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x) : \Lambda \subset \mathscr{M}, |\Lambda| < \infty$, $\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} = 1\}$, the update operation acts as:

$$\phi_y \left(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x) \right) = \sum_{\mathbf{n}\in t(y,\Lambda)} \hat{w}_{\mathbf{m}} h\left(x, \mathbf{n}, T(y,\theta)\right) \pi(\mathrm{d}x) \tag{5}$$

with $t(y,\Lambda) = \{\mathbf{n} : \mathbf{n} = t(y,\mathbf{m}), \mathbf{m} \in \Lambda\}$, and $\hat{w}_{\mathbf{m}} \propto w_{\mathbf{m}}$ and for $\mathbf{n} = t(y,\mathbf{m})$, $\sum_{\mathbf{n}\in t(y,\Lambda)} \hat{w}_{\mathbf{n}} = 1$. This updates the signal given the new data by means of the Bayes theorem. The prediction operation acts as:

$$\psi_t \left(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x) \right) = \sum_{\mathbf{n}\in G(\Lambda)} \left(\sum_{\mathbf{m}\in\Lambda, \mathbf{m}\geq\mathbf{n}} w_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t,\theta) \right) h\left(x, \mathbf{n}, \theta_t\right) \pi(\mathrm{d}x) \tag{6}$$

where $G(\Lambda) = \{\mathbf{n} \in \mathscr{M} : \mathbf{n} \leq \mathbf{m}, \mathbf{m} \in \Lambda\}$, propagating the current filtering distribution by means of the signal transition kernel. As such, filtering a hidden Markov model using the duality relation consists in recursive operations on finite mixtures of distributions, where the number of components remains finite and the components remain within the same family of distributions. At each new observation, the mixture distribution is shifted towards the data, then until the next observation, the mixture progressively forgets the past information and drifts back towards the prior distribution.

2 Implementation of the dual filtering algorithm

The filtering algorithm resulting from the method presented above is similar to the Baum-Welch filter and it alternates update and prediction steps. The update step shifts each component and modifies its weight, while the prediction step lets all the components propagate some of their mass towards the components close to the prior. We illustrate this dual filtering algorithm (Eq. (1)) for two stochastic processes: the Cox-Ingersoll-Ross (CIR) process and the Wright-Fisher (WF), presented in full details later. For these two models, the number of mixture components in the filtering distributions evolves as $|\Lambda_n| = \prod_{i=1}^{K}(m_{0,i} + 1 + \sum_{i=1}^{n} Y_i)$, where K is the dimension of the latent space and $\mathbf{m}_0 = m_{0,1:K}$ is the initial state.

The prediction step is much costlier than the update step, as at each iteration it involves computing the transitions from all elements of Λ_i to those reachable by a pure death process in $G(\Lambda_i)$. It is possible to contain the cost of the prediction operation by storing the transition terms $p_{\mathbf{m},\mathbf{n}}$, which will be used multiple times during the successive iterations. However, the rapid growth in the number of those terms (proportional to $|G(\Lambda_n)|^2$) does not permit saving all of them in memory. Yet, the $p_{\mathbf{m},\mathbf{n}}$ are themselves a product of a number of terms which grows only quadratically with the sum of all observations and can be saved (Eq. (4) and the hypergeometric coefficients expressed as a product of binomials coefficients). Another technical difficulty is that the sum with terms of alternated sign in (4) is susceptible to both over and underflow. We compute it using the `Nemo` library for arbitrary precision computation ([3]).

Although considerable efficiency gains are achieved by storing the transition terms, further improvements may be obtained by a natural approximation of the filtering distributions. Indeed, the filtering distributions contain a number of components that grows quickly as new observations arrive, but the complexity of the hidden signal does not necessarily increase accordingly. Hence, if the prior is reasonable and the posteriors appropriately concentrated, there is no reason for the number of components with non negligible weight to explode. Indeed, simulation studies show that the number of components representing 99% of the weight of the mixture saturates as new observations arrive (Eq. (2)). This suggests that some components may be deleted from the mixtures, speeding the computations, without loosing much in terms of precision. We envision three strategies for pruning the mixtures:

- prune all the components who have a weight below a certain threshold, which is an attempt at controlling the approximation error at a given step. This approach will be referred to as the *fixed threshold strategy*.
- retain only a given number of components, hopefully chosen above the saturation number (see Eq. (2)). This is an attempt at controlling the computation budget at each time step. This approach will be referred to as the *fixed number strategy*.
- retain all the largest components needed to reach a certain amount of mass, for instance 99%. This is an adaptive strategy to keep the smallest possible

number of components under a certain error tolerance level. This approach will be referred to as the *fixed fraction strategy*.

In Algorithm 1, the pruning is performed just after the update step. This choice is dictated by two reasons: first, after the update step the mixture should be more concentrated because information from the new observation was just incorporated, leading to a smaller number of components with non negligible weight. Then, as the prediction step is the most computationally expensive, reducing the number of components before predicting entails the maximum computational gain. After pruning, we renormalise all the remaining weights so that they sum to 1. As the pruning operation occurs at each time step, the level of approximation on a given filtering distribution results from several successive approximations.

Algorithm 1: Optimal filtering algorithm using the dual process, with the option of pruning.

```
Data: $Y_{0:n}$, $t_{0:n}$ and $\nu = h(x, \mathbf{m}_0, \theta_0) \in \mathscr{F}$ for some $\mathbf{m}_0 \in \mathscr{M}, \theta_0 \in \Theta$
Result: $\Theta_{0:n}$, $\Lambda_{0:n}$ and $W_{0:n}$ with $W_i = \{w^i_{\mathbf{m}}, \mathbf{m} \in \Lambda_i\}$
Initialise
    Set $\Theta_0 = \theta_0$
    Set $\Lambda_0 = \{t(Y_0, \mathbf{m}_0)\} = \{m^*\}$ and $W_0 = \{1\}$ with $t$ as in (5)
    Let $\Theta_0$ evolve during $t_1 - t_0$ and set $\theta^*$ equal to the new value
    Set $\Lambda^* = G(\Lambda_0)$ and $W^* = \{p_{m^*,\mathbf{n}}(t_1 - t_0, \theta_0), \mathbf{n} \in \Lambda^*\}$ with $G$ as in (6) and $p_{\mathbf{m},\mathbf{n}}$ as in
      (3)
for $i$ from 1 to $n$ do
    Update
        Set $\Theta_i = \theta^*$
        Set $\Lambda_i = \{t(Y_i, \mathbf{m}), \mathbf{m} \in \Lambda^*\}$
        Set $W_i = \{\frac{w^*_{\mathbf{m}} p_{h(x,\mathbf{m},\Theta_i)}}{\sum_{\mathbf{n}\in\Lambda^*} w^*_{\mathbf{n}} p_{h(x,\mathbf{n},\Theta_i)}}, \mathbf{m} \in \Lambda^*\}$ with $p_{h(x,\mathbf{m},\theta)}$ defined as in (1)
    if pruning then
        Prune($\Lambda_i$) and remove the corresponding weights in $W_i$
        Normalise the weights in $W_i$
    Predict
        Let $\Theta_i$ evolve during $t_{i+1} - t_i$ and set $\theta^*$ equal to the new value
        Set $\Lambda^* = G(\Lambda_i)$ and $W^* = \left\{ \sum_{\mathbf{m}\in\Lambda_i, \mathbf{m}\geq\mathbf{n}} w^i_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t_{i+1} - t_i, \Theta_i), \mathbf{n} \in \Lambda^* \right\}$
end
```

3 Filtering two stochastic processes

For illustration we consider two stochastic processes, a 1-dimensional Cox-Ingersoll-Ross process and a 3-dimensional Wright-Fisher process, which we filter using the strategy outlined above. The dimension of the state space of the pure death process is dependent on the dimension of the signal, therefore the number of components in the filtering distributions for the WF process is much greater than for the CIR process, rendering the inference computationally more challenging. The one-dimensional CIR process has the following generator: $\mathscr{A} = (\delta\sigma^2 - 2\gamma x)\frac{d}{dx} + 2\sigma^2 x\frac{d^2}{dx^2}$. with $\delta, \gamma, \sigma > 0$ and stationary distribution $\mathrm{Ga}(\delta/2, \gamma/\sigma^2)$. A conjugate emission density is: $Y_t|X_t \sim \mathrm{Po}(X_t)$. The duality function can be found in [2]. We simulate a CIR process starting from $X = 3$ with $\delta = 3.6$, $\gamma = 2.08$, $\sigma = 2.8$. which corresponds to a stationary distribution $\mathrm{Gamma}(1.8, 0.38)$. Furthermore, we simulate 10 observations at each time, with 200 time steps separated by 0.011 seconds. For the inference, we use as a prior for the stationary distribution a $\mathrm{Gamma}(1.5, 0.15625)$ which corresponds to $\gamma = 2.5$, $\delta = 3.$, $\sigma = 4.$ and $m_0 = 0$.

The Wright-Fisher model is a K-dimensional diffusion, whose generator is $\mathscr{A} = \frac{1}{2}\sum_{i=1}^{K}(\alpha_i - |\alpha|x_j)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{K} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j}$. and its stationary distribution is a Dirichlet(α). A conjugate emission density is: $f_x(Y) = \prod_{i=1}^{J}\left(|\mathbf{n}_i|!\prod_{k=1}^{K}\frac{x_k^{n_{ki}}}{n_{ki}!}\right)$. The duality function can also be found in [2]. We simulate two datasets using a discrete time and finite population Wright-Fisher model of dimension $K = 3$ initialised at random from a Dirichlet(0.3, 0.3, 0.3) with $\alpha = (0.75, 0.75, 0.75)$ and a population size of 50000. 15 observations are collected at each observation time. There are 10 observation times with a time step of 0.1 second for the first dataset and 20 observation times with a time step of 0.004 second for the second dataset. As a prior, we use a uniform distribution Dirichlet(1,1,1). The two different time steps for the WF model are intended to explore two regimes, one for which the time between observations is large, such that information from previous data is almost forgotten (the predictive distribution has almost moved back to the prior) and one for which that time is very short. In these two regimes, the number of components with non negligible weights is expected to be very different. Notably, in the second regime the impact of the successive approximations is expected to be stronger. Fig. 1 shows that in all the studied cases, the filtering distributions are centred around the signal. For the WF model with the short time step, the filtering distributions do not evolve fast enough to follow the signal exactly, but this is to be expected given the rapid rate at which new observations arrive. Considering how the weights are distributed among the components of the filtering distribution, we observe that the mass is mostly concentrated on a number of components several orders of magnitude smaller than the total number of components. This observation suggests that many components may be deleted with a minimal loss of precision.

To quantify this loss of precision by pruning, we compute the Hellinger distance between the exact and the approximate filtering distributions obtained by pruning: $d_H(f_1, f_2) = \frac{1}{2}\int_{\chi}(\sqrt{f_1} - \sqrt{f_2})^2$. As there is one filtering distribution per observation time, to compare two sets of filtering distributions we consider the maximum over time of the distance between the distributions at each time, i.e. $\sup_n(d_H(\nu_{n,\mathrm{exact}}, \nu_{n,\mathrm{approx}}))$. The numerical evaluation of the distances is done using standard quadrature and simplicial cubature rules from the R package `SimplicialCubature`. Parallel to the loss of precision due to the approximation, we consider the gain in efficiency by measuring the computing time needed

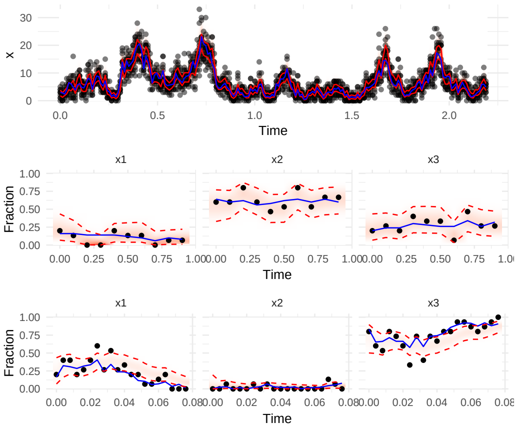


Fig. 1: Hidden signal, data and 95% credible intervals of the filtering distribution for the three datasets. The hidden signal is denoted by the blue line, the data by the black dots and the credible bands by the red dashed lines. Top: CIR, centre: WF, bottom: WF with short time step. For the WF model, each panel corresponds to one marginal, and the data plotted is the proportion of the 15 multinomial observations which are from the corresponding type.

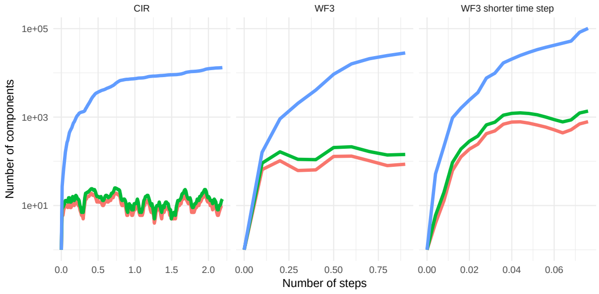


Fig. 2: Number of components (in log scale) in the filtering distributions as a function of the iteration number. Left: CIR, centre: WF, right: WF with short time step. The blue line denotes the total number of components in the filtering distributions, the green line denotes the number of components carrying 99% of the mass and the red 95%.

to filter the whole dataset. Fig. 3 shows that the approximation strategies afford a reduction in computing time by 5 orders of magnitude for the CIR process, or by 2 to 3 orders of magnitude for the three-dimensional WF process. The fixed fraction strategy is noticeably slower in the case of the WF process with the shorter time step because the mass is spread over more components, as was also apparent on Fig. 2. For all strategies and all processes, it seems possible to find a compromise between accuracy and computing time where increasing the computational effort starts yielding diminishing returns. Except in the case of the CIR model where the fixed threshold strategy seems to slightly outperform the others, no strategy seems to offer a fundamentally better precision/cost ratio than the others.

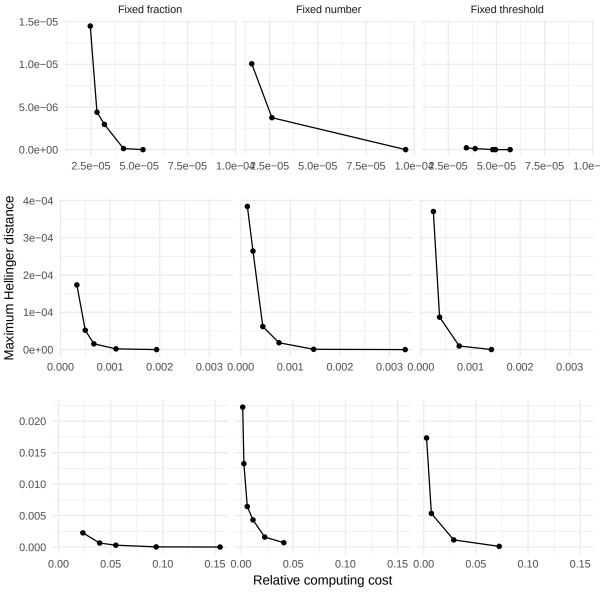


Fig. 3: Approximation error versus computational effort. The computation time is given relative to the time needed for obtaining the exact filtering distributions. The top level represents the CIR process, the middle represents the WF process and the bottom represents the WF process with the shorter time step. Fixed fractions tested are 0.8, 0.9, 0.95, 0.99, 0.999. The fixed numbers tested are 5, 10, 25 for the CIR process, 10, 25, 50, 100, 200, 400 for the WF processes. The fixed thresholds are 0.01, 0.005, 0.001, 0.0005, 0.0001 for the CIR orocess and 0.01, 0.005, 0.001, 0.0001 for the WF processes.

The results presented here are a preliminary study on the computational costs of filtering strategies based on duality. A more thorough investigation of these and other aspects involved in this type of filtering are currently ongoing work.

References

1. Cappé, O., Moulines, E., Ryden, T.: Inference in Hidden Markov Model. Springer, New York, USA (2005)
2. Papaspiliopoulos, O., Ruggiero, M.:Optimal filtering and the dual process. Bernoulli **20**, 1999–2019 (2014)
3. Fieker, C., Hart, W., Hofmann, T., Johansson, F.: Nemo/Hecke: Computer Algebra and Number Theory Packages for the Julia Programming Language. In: Proc. 2017 ACM Int. Symp. Symb. Algebr. Comput., pp. 157–164. ACM, New York, USA (2017)

K-means seeding via MUS algorithm

Inizializzazione del K-means tramite l'algoritmo MUS

Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli

Abstract K-means algorithm is one of the most popular procedures in data clustering. Despite its large use, one major criticism is the impact of the initial seeding on the final solution. We propose a modification of the K-means algorithm, based on a suitable choice of the initial centers. Similarly to clustering ensemble methods, our approach takes advantage of the information contained in a co-association matrix. Such matrix is given as input for the MUS algorithm that allows to define a pivot-based initialization step. Preliminary results concerning the comparison with the classical approach are discussed.

Abstract *L'algoritmo K-medie è una delle procedure di raggruppamento più utilizzate. Tuttavia, una delle maggiori criticità di tale metodo riguarda l'impatto della scelta dei semi iniziali sulla configurazione finale. In questo lavoro viene proposta una variante del K-medie basata su una scelta opportuna dei semi iniziali. In linea con i cosiddetti 'metodi di insieme', l'approccio considerato sfrutta l'informazione contenuta in una matrice di co-associazione. Tale matrice viene utilizzata dall'algoritmo MUS per definire i semi iniziali dei gruppi sulla base di unità pivotali. Vengono discussi alcuni risultati preliminari riguardanti il confronto con l'approccio classico.*

Key words: Clustering, pivotal unit, seeding

Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, 'Bruno de Finetti', Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: leoegidi@hotmail.it, rpappada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it,

1 Introduction

The goal of cluster analysis is to group a given collection of objects in such a way that instances in the same cluster are as similar as possible, according to a suitable "similarity" criterion (see, e.g., [5]). One of the most popular and widely used clustering techniques is K-means algorithm (see [7] and the references therein). For a given dataset of n observations $\mathbf{Y} = (y_1, \dots, y_n)$, with $y_i \in \mathscr{Y} \subset \mathbb{R}^d,\ i = 1, \dots, n$, K-means seeks to find a partition of the data into K clusters $\mathscr{S} = \{S_1, \dots, S_K\}$ so as to minimize

$$\phi(\mathscr{S}) = \sum_{k=1}^{K} \sum_{y_i \in S_k} ||y_i - \mu_k||^2,$$

where μ_k is the k-th cluster center. The algorithm begins with K randomly initialized centers, and assigns each point to the nearest center; then, the clusters' centers are recomputed and the partition updated; such process is iterated until a stable configuration is reached. In many instances, the number of clusters K is specified in advance, and the optimal solution is sought conditional on such value. Also the distance adopted is set in advance, depending on the nature of the data and user subjective preferences. Initial seeding is a more technical issue, usually less discussed, whose impact on the final result is often neglected. Nonetheless, the choice of the initial group centers may strongly affect the clustering solution. As already mentioned, the classical approach in K-means clustering uses a random seeding in the first step of the procedure. In practice, multiple random seeds are considered, and the final K-means solution is chosen as the one that minimizes the objective function. An alternative seeding technique is proposed by [1], among others. Many extensions of the classical approach have been explored in the literature (for a complete review see [7]). As it is discussed in [7], classical approaches have been challenged by using ensembles methods.

Clustering ensembles methods ([9, 10]) explore the idea of the so-called evidence accumulation in order to summarize the information coming from multiple clusterings into a pairwise *co-association* matrix, regarded to as a similarity matrix ([8, 6]). Such matrix is constructed by taking the co-occurrences of pairs of units in the same cluster among the total number of partitions, and then used to deliver the final consensus clustering. This kind of matrix has been estimated in [4] and used in the context of the label switching problem in Bayesian estimation of finite mixture models. In particular, the algorithm of Maxima Units Search (MUS), introduced in [4] and further developed in [3], has proved to be useful in extracting some specific units–one for each mixture component–called pivots, from a large and sparse similarity matrix representing an estimate of the probability that pairs of units belong to the same group.

Here the main idea is to exploit the use of the pivots detected by the MUS algorithm, which are determined as the observations that are "as far away from each other as possible" according to the co-association matrix, as group centers in the initialization step of K-means procedure, where such units.

Section 2 briefly reviews the MUS procedure for the identification of pivotal units from a given set of data, and outlines the proposed approach. In Section 3 the performance of the presented algorithm is preliminarily investigated by means of a small simulation study. Section 4 presents some final remarks.

2 Seeding via MUS algorithm

Consider H distinct partitions of a set of n d-dimensional statistical units into K groups determined by some clustering technique. It is possible to map them into a $n \times n$ co-association matrix C with generic element $c_{i,j} = n_{i,j}/H$, where $n_{i,j}$ is the number of times the pair (y_i, y_j) is assigned to the same cluster with respect to the clustering ensemble. Units which are very distant from each other are likely to have zero co-occurrences; as a consequence, C is a square symmetric matrix expected to contain a non-negligible number of zeros.

The main task of the MUS algorithm is to detect submatrices of small rank from the co-association matrix and extract those units $y_{i_1}, \dots, y_{i_K}$ such that the $K \times K$ submatrix of C with only the $i_1, \dots, i_K$ rows and columns has few, possibly none, nonzero elements off the diagonal (that is, this submatrix is identical or nearly identical). Practically, the resulting units—hereafter pivots—have the desirable property to be representative of the group they belong to. From a computational point of view, the issue is non-trivial and involves a global search row by row; as n, K and the number of zeros within C increase, the procedure becomes computationally demanding. Given that the pivots correspond to well separated units in the data space, they can represent an alternative approach to the random seeding in K-means setting. A similar idea has been discussed in [1], where the initial centers are chosen on the basis of suitable weights assigned to data points.

Although K-means clustering is one of the most popular algorithms due to its simplicity and low computational burden, one major criticism is the impact of the choice of the initial centers on the final solution. However, limited work has been developed for improving the seeding of the centers. A modified version of K-means could benefit from a pivot-based initialization step. In particular, the starting point is performing multiple runs of the classical K-means with K fixed, and build the co-association matrix of data units. Such matrix is given as input for the MUS procedure, yielding the pivots regarded to as cluster centers. Intuitively, such approach represents a careful seeding which may improve the validity of the final configuration. According to the general K-means method, steps 1a—1c of the `MUS`K-means algorithm summarized below collapse in a single step, where the initial centers are chosen uniformly at random from the data space $\mathscr{Y}$. The remaining steps coincide with those of the classical K-means version.

MUSK-means:

1a Perform H classical K-means algorithms, and obtain then H distinct data partitions, with initial centers chosen uniformly at random.

1b Build the co-association matrix C, where $c_{i,j} = n_{i,j}/H$, with $n_{i,j}$ the number of times the pair (y_i, y_j) is assigned to the same cluster among the H partitions.

1c Apply the MUS algorithm to the matrix C and find the pivots $y_{i_1}, \dots, y_{i_K}$. For each group, set the initial center $\mu_k = y_{i_k}$.

2 For each k, $k = 1, \dots, K$, set the cluster S_k to be the set of points in $\mathscr{Y}$ that are closer to μ_k than they are to μ_j for all $j \neq k$.

3 For each k, $k = 1, \dots, K$, set μ_k to be the center of mass of all points in S_k: $\mu_k = \frac{1}{|S_k}|\sum_{y \in S_k} y$, where $|S_k|$ is the cardinality of S_k.

4 Repeat Steps 2 and 3 until $\mathscr{S}$ no longer changes.

3 Simulation results

A preliminary simulation study is carried out in order to explore the performance of the methodology proposed in Sect. 2. One of the drawbacks of K-means is its inefficiency in distinguishing between groups of unbalanced sizes. For this reason, two different scenarios in which the classical approach may fail to identify the ‘natural’ groups are considered in the following. In particular, the two simulated datasets reproduce two clusterings in two dimensions, with three and two groups, respectively. For illustration purposes, the results from a single simulation are shown in Fig. 1. The left panel (top) displays the first simulated scenario, where the input data consist of three clusters drawn from bivariate Gaussian distributions with 20, 100 and 500 observations, respectively. The partitions obtained from the classical K-means algorithm using multiple random seeds and from MUSK-means are plotted in the top central and right panel of Fig. 1, respectively. As can be seen, classical K-means tends to split the cluster with the highest density in two separate clusters; conversely, the cluster composition identified by MUSK-means shows a greater agreement with the true partition, and the final centers are close to the pivotal units used for the seeding. The second configuration (see the bottom panel of Fig. 1) consists of data with ‘two-sticks’ shaped groups of 30 and 370 observations, respectively. Classical K-means fails in recognizing the true pattern, and the final centers both belong to the largest cluster. Clustering based on our pivotal units seems to correctly identify the simulated groups, since two well separated pivots are identified and set as initial group centers.

In order to evaluate and compare the performance of classical K-means and MUSK-means, a common measure of the similarity between two partitions, namely,

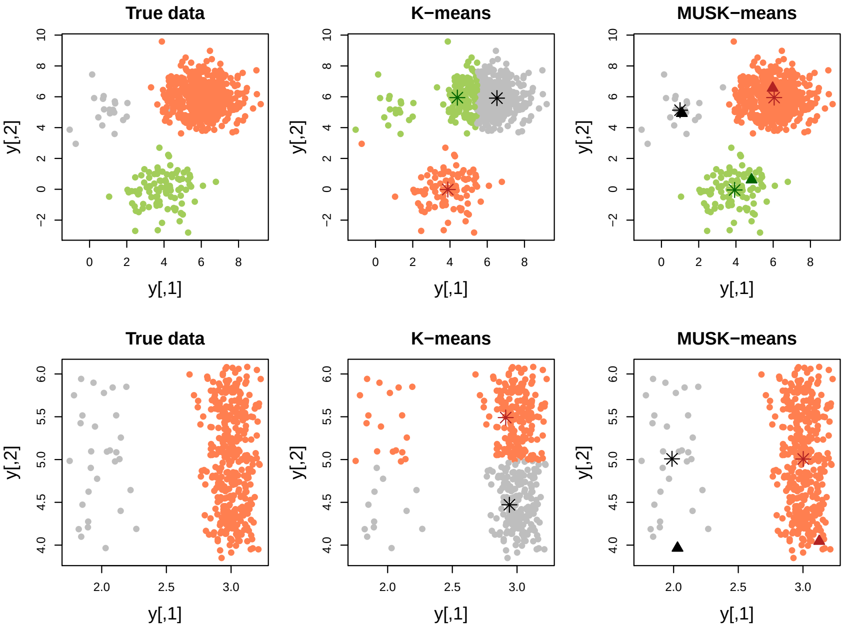


Fig. 1: From left to right. Input data generated from a mixture of three Gaussian distributions (620 samples) (top) and 'stick' data (400 samples) with two groups (bottom) of unequal sample sizes; clustering solutions obtained via classical K-means and MUSK-means algorithms. Each cluster identified is shown in a different color, with final group centers and pivots marked via asterisks and triangles symbols, respectively.

the Adjusted Rand Index (ARI), is computed at each iteration between the resulting clustering and the true data partition. The number of replications in the simulation study is set equal to 1000. Fig. 2 shows the comparison in terms of ARI, for the first scenario considered. As may be noted, MUSK-means gives overall good results; in fact, it yields higher values for 60% of the replications, whereas the two procedures yield the same value of the index in 36% of cases. Concerning the second scenario characterized by 'two-sticks' data, the ARI for K-means is always approximately equal to 0; the same index for MUSK-means is about zero in 43% of cases, while for the remaining 57% it outperforms classical K-means giving an ARI equal to 1, denoting a perfect agreement with the true partition.

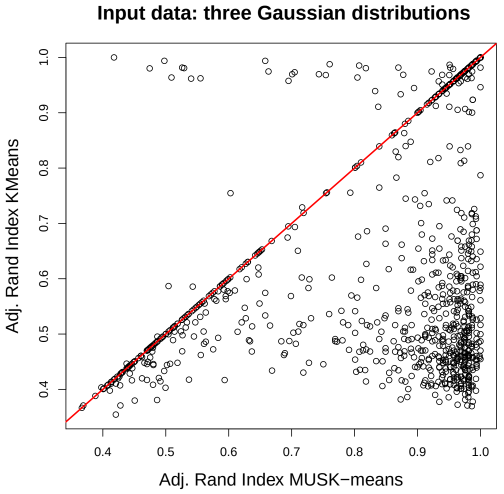


Fig. 2: Comparison between the ARI obtained via classical K-means and `MUS`K-means algorithms for the input Gaussian data over 1000 replications.

4 Discussion

We propose a modified K-means algorithm which exploits a pivotal-based phase seeding. Despite the limited study, preliminary results seem to be promising in terms of clusters' identification. It is worth noting that the proposed algorithm is in general computationally more demanding that the standard procedure, and the complexity grows with the size of the dataset. On the other hand, similarly to clustering ensemble, our method takes advantage of the construction of a co-association matrix, whose information has been only partially exploited so far. Further work is needed to investigate the use of such matrix and the pivotal units for inferring the optimal number of groups.

References

1. Arthur, D., Vassilvitskii, S.: `k-means++`: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, pp. 1027-1035 (2007)
2. Dongkuan, X., Yingjie, T.: A comprehensive survey of clustering algorithms. Annals of Data Science **2**(2), 165–193 (2015)
3. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Maxima Units Search (MUS) algorithm: methodology and applications. In: Perna, C. , Pratesi, M., Ruiz-Gazen A. (eds.) Studies in Theoretical and Applied Statistics, Springer Proceedings in Mathematics & Statistics 227, pp. 71–81 (2018)

4. Egidi, L., Pappadà, R., Pauli, F., Torelli, N.: Relabelling in Bayesian mixture models by pivotal units. Stat. Comput. **28**(4), 957–969 (2018)
5. Everitt, B.S.: Cluster Analysis. Heinemann, London, United Kingdom (1981)
6. Fred, A. L., Jain, A. K.: Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(6), 835–850 (2005)
7. Jain, A.: Data clustering: 50 years beyond K-means. Patt. Recog. Lett. **31**(8), 651–666 (2010)
8. Lourenco, A., Bulò, S. R., Fred, A., Pelillo, M.: Consensus clustering with robust evidence accumulation. In: International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 307-320. Springer, Berlin, Heidelberg (2013)
9. Strehl, A., Joydeep G.X: Cluster ensembles–a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2002)
10. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence **25**(3), 337–372 (2011)

4. Sollicited Sessions

Advances in Discrete Latent Variable Modelling

A joint model for longitudinal and survival data based on a continuous-time latent Markov model

Un modello congiunto per dati longitudinali e di sopravvivenza basato su un processo Markoviano latente in tempo continuo

Alessio Farcomeni and Francesco Bartolucci

Abstract A shared-parameter approach for jointly modeling longitudinal and survival data is proposed, which allows for time-varying random effects that enter both in the longitudinal and survival processes. The distribution of these random effects is modeled according to a continuous-time hidden Markov chain, so that latent transitions may occur at any time point. Our formulation allows for: (*i*) informative drop-out with precise time-to-event outcomes, while existing approaches are all based on drop-out at longitudinal measurement times and (*ii*) completely non-parametric treatment of unequally spaced intervals between consecutive measurement occasions (even not in the presence of drop-out). For maximum likelihood estimation we propose an algorithm based on coarsening. The resulting estimator is studied by simulation. The approach is illustrated by an application to data about patients suffering from mildly dilated cardiomyopathy.

Abstract *Si propone un modello congiunto per dati longitudinali e di sopravvivenza. La distribuzione degli effetti casuali condivisi è modellata sulla base di un processo latente in tempo continuo e spazio degli stati discreto. La formulazione proposta permette di prevedere (i) drop-out informativo in presenza di tempo-a-evento precisamente misurato, mentre gli approcci attualmente disponibili sono basati su indicatori al momento della misurazione del dato longitudinale, e (ii) un trattamento non-parametrico per il caso di intervalli non-omogenei tra occasioni di misura, anche in assenza di drop-out. Per la stima di massima verosimiglianza sviluppiamo un algoritmo basato sul coarsening. Lo stimatore risultate é valutato per simulazione. L'approccio é illustrato tramite una applicazione a dati su pazienti cardiomiopatia dilatativa di grado lieve.*

Key words: Baum-Welch recursion, Informative drop-out, Unequally spaced times

A. Farcomeni
Sapienza - University of Rome, e-mail: alessio.farcomeni@uniroma1.it

F. Bartolucci
University of Perugia e-mail: francesco.bartolucci@unipg.it

1 Introduction

Informative drop-out in longitudinal studies is often treated by linking a model for time to drop-out and one for the longitudinal outcome. Many models devised for informative drop-out assume that subject-specific parameters are time-constant. This is a limitation as unobserved factors affecting the outcomes and the relationship between longitudinal and survival outcomes might evolve over time in an unpredictable way, especially when the follow-up is relatively long. One exception is [1], who propose a discrete-time event-history approach based on latent Markov models [2], which naturally accommodate time-dependent unobserved heterogeneity. Nevertheless, the approach in [1] has two limitations: (*i*) the event-history component models drop-out, by a conditional logit model, as occurring within a time interval, therefore ignoring precise follow-up time information; (*ii*) latent transitions, as common in latent Markov models, are based on a discrete-time stochastic process and hence transitions may only occur at visit times. In terms of interpretation, assuming that transitions may occur only at certain time occasions is rather unrealistic and the explicit use of an hazard function is preferable to that of a conditional logit model.

In order to overcome the above limitations, we propose a shared-parameter model characterized by the following features. First of all the time-varying unobserved heterogeneity is accounted for by a continuous-time discrete-state hidden Markov model [5], parameterized by an initial probability vector and an infinitesimal transition matrix. In this respect our approach can be seen as a complete generalization of the relevant work by [3], which is limited to $k = 2$ and to missing-at-random data. Second, for the survival time we assume a Weibull model with hazard function depending on the (entire) trajectory of the continuous-time latent variable. A latent class model (with time-constant subject-specific parameters) is obtained whenever the infinitesimal transition matrix is constrained to have all elements equal to 0.

For model fitting we introduce a novel method that provides maximum likelihood estimates. The approach is based on a time discretization, in a certain number of windows of arbitrary length, and on an extension of the Baum-Welch recursions. It converges in an accurate and stable way. This algorithm also represents an advance in the literature about estimation of continuous hidden Markov models in general, with respect to computational demands, ease of implementation, and stability.

In the following we illustrate in some detail the assumptions of the proposed model, then we describe the approach to parameter estimation and, finally, we outline an application based on data about patients suffering from Mildly Dilated CardioMyopathy (MDCM).

2 Shared-parameter continuous-time latent Markov and survival models

Consider a sample of n individuals and for individual i, with $i = 1,\dots,n$, let $T_i = \min(T_i^*, C_i)$ be the survival time taken as the minimum between the true event time T_i^* and the censoring time C_i. Furthermore, let Δ_i be the corresponding event indicator defined by $\Delta_i = I(T_i^* \leq C_i)$, where $I(\cdot)$ is the indicator function equal to 1 if its argument is true and to 0 otherwise. The outcome $Y_i(t)$, which arises from a natural exponential family, is repeatedly observed at arbitrary time points t_{ij}, $j = 1,\dots, j_i$, where j_i is the number of observations; also let $Y_{ij} = Y_i(t_{ij})$. We assume that the longitudinal process is associated with T_i^*, namely with the true event time, but, as customary in survival analysis, is independent of the censoring time C_i. In general, realizations of random variables are denoted by small letters, so that, for instance, t_i is the observed value of T_i and δ_i is the observed value of Δ_i.

We denote by w_i a row vector of (time-fixed) baseline covariates to be used in modeling the survival process. For the longitudinal process, we denote by $x_i(t)$ a vector of predictors at time t and we also let $x_{ij} = x_i(t_{ij})$, $j = 1,\dots, j_i$.

The proposed model is based on two equations. Specifically, the model for the longitudinal outcome is formulated along the usual lines of mixed-effects models and the model for the time-to-event outcome is based on a subject-specific hazard function as in Cox-type models. More formally, we assume that

$$\begin{aligned} g(\mu_{ij}) &= \alpha_i(t_{ij}) + x'_{ij}\beta, \quad j = 1,\dots, j_i, \\ h(t_i^*) &= h_0(t_i^*)\exp\{\alpha_i(t_i^*)\phi + w'_i\psi\}, \end{aligned}$$

where $g(\cdot)$ is a link function of the conditional expectation of Y_{ij} denoted by μ_{ij} and $h(\cdot)$ is the hazard function, with $h_0(\cdot)$ being a baseline hazard. In this paper we will assume a Weibull parametric form for $h_0(\cdot)$, that is, $h_0(t) = \nu t^{\nu-1}$, resulting in an Accelerated Failure Time (AFT) model for the survival part. Other parametric choices, or even a non-parametric specification, are possible. We assume that $\alpha_i(t)$ follows a time-continuous Markov process, whereas β and ψ are fixed parameter vectors for the covariates, and ϕ is a parameter for the effect of the latent process on the survival process. Note that several generalizations, including the case of more than one parameter being time-dependent according to the latent process, are straightforward. Regarding the distribution of Y_{ij}, our model has the same degree of flexibility as generalized linear models. It is worth also stressing that the hazard function depends on the *entire* trajectory of the random effect $\alpha_i(t)$, and not only on $\alpha_i(t_{ij})$.

Unlike usual formulations, random intercepts are assumed to be time varying. This greatly enhances model flexibility. In particular, as already mentioned, we assume that the random effects follow a continuous-time (discrete-state) Markov chain [5], with state-space $\{\xi_1,\dots,\xi_k\}$ having elements collected in the column vector ξ. We assume that the transition function of the latent chain satisfies the Chapman-Kolmogorov equations, and specify its Q-matrix based on positive off-diagonal elements q_{uv} for $u, v = 1, \dots, k$ and $v \neq u$. By definition, the diagonal elements are given by $-q_u$, with $q_u = \sum_{v=1, v\neq u}^{k} q_{uv}, u = 1, \dots, k$. Accordingly, for the longitudinal outcome transitions from time t to time $t+s$ are collected in the $k \times k$ matrix $\Pi_s = e^{sQ}$, where e denotes the *matrix exponential* operator, that is, $e^{sQ} = \sum_{n=0}^{\infty} \frac{s^n Q^n}{n!}$. Note that irregularly spaced time occasions are directly accommodated, and hence our model also generalizes [3] simply by restricting it to the first equation. We also define the *jump matrix* R as a matrix with off-diagonal elements $r_{uv} = q_{uv}/q_u$, and collect initial probabilities π_u in the column vector π.

The latent process captures the time-varying unobserved heterogeneity linking the longitudinal and survival outcomes. The shared-parameter formulation is in the spirit of copula models [6]. We also recall that, according to this process, the sojourn time in each state u has an exponential distribution with parameter q_u, denoted as $\text{Exp}(q_u)$, whereas the probability of moving at the end of the sojourn time to state v is equal to the suitable element of the jump matrix R.

3 Estimation

It is straightforward to check that the complete likelihood of our proposed model depends on the entire trajectory of the continuous-time latent process, through the integrals involved in the time-to-event component. This makes it hard to efficiently compute the observed likelihood: the classical Baum-Welch recursion is not directly applicable, even after their extension to continuous time processes, due to lack of certain conditional independence statements. To derive inference we build a sequence of equally spaced windows corresponding to time fixed points $\bar{t}_1, \dots, \bar{t}_M$, with $\bar{t}_1 = 0$. The first window spans the time interval from $\bar{t}_1$ to $\bar{t}_2$, the second from $\bar{t}_2$ to $\bar{t}_3$, and so on. These time points are chosen so that each observation time t_{ij} corresponds to one of them. Let $\bar{y}_{im}$ denote the observation at time $\bar{t}_m$ for individual i, which may be missing for certain time occasions, $\bar{x}_{im}$ be the corresponding vector of covariates, and $\bar{U}_{im}$ the corresponding latent variable. Also let $\bar{y}_{i,\leq m}$ be vector of observations available until time $\bar{t}_m$.

The following forward recursion can now be used. Consider the joint density $f_{im}(u) = f(\bar{y}_{i,\leq m}, t_i \geq \bar{t}_m, \bar{U}_{im} = u)$ referred the observation availably until time $\bar{t}_m$ for individual i, latent state at the same time occasion, and for the event that the individual survives time $\bar{t}_m$. We have that $f_{i1}(u) = \pi_u f(\bar{y}_{i1}|\bar{U}_{i1} = u), \quad u = 1, \dots, k$, and $f_{im}(v) = f(\bar{y}_{im}|\bar{U}_{i1} = v) \sum_{u=1}^{k} \pi_{v|u} f_{i,m-1}(u) S_{m-1}(\bar{t}_m, u), \quad m = 1, \dots, m_i, v = 1, \dots, k$, where, in general, we have $S_m(\bar{t}, u) = \exp\{-H_m(\bar{t}, u)\}$ with

$$H_m(\bar{t}, u) = \int_{\bar{t}_m}^{\bar{t}} \exp(\xi_u \phi + w_i' \psi) \nu t^{\nu-1} dt = \exp(\xi_u \phi + w_i' \psi)(\bar{t}^{\nu} - \bar{t}_m^{\nu}), \quad u = 1, \dots, k,$$

$f(\bar{y}_{im}|\bar{U}_{i1} = v)$ is set equal to 1 if the observation is not available at time t_m, and m_i be largest value of m such that $\bar{t}_m \leq t_i$. For individual i we have the contribution to the likelihood given by $f(y_i, t_i, d_i) = \sum_{u=1}^{k} f_{im_i}(u) h(t_i)^{\delta_i} S_{m_i}(u, t_i)$. Regarding the transition probabilities $\pi_{v|u}$, note that these are the elements of the $k \times k$ matrix $\mathit{\Pi}_a$ obtained as $\exp(aQ)$, where $a = \bar{t}_{m+1} - \bar{t}_m$ is the length of each time window.

The log-likelihood function to be maximized is then $\ell(\theta) = \sum_{i=1}^n \log f(y_i, t_i, d_i)$. In order to maximize this function we also need a backward recursion. In particular, let $g_{im}(u) = f(\bar{y}_{i,>m}, t_i, \delta_i | t_i > t_m, U_{im} = u)$. For $m = m_i$ have that $g_{im_i}(u) = h(t_i)^{\delta_i} S_{m_i}(t_i, u)$ and $g_{im}(u) = S_m(\bar{t}_{m+1}, u) \sum_{v=1}^k \pi_{v|u} g_{i,m+1}(v) f(\bar{y}_{i,m+1}, U_{i,m+1} = v)$ for $m < m_i$. From this recursion we can obtain two posterior distributions used to update the parameters π and $\mathit{\Pi}_a$. In particular, we have that

$$p(U_{im} = u | y_i, t_i, \delta_i) = \frac{f_{im}(u) g_{im}(u)}{f(y_i, t_i, d_i)}, \quad m = 1, \ldots, m_i,\ u = 1, \ldots, k.$$

Moreover, we have that

$$p(U_{im} = u, U_{i,m+1} = v | y_i, t_i, \delta_i) = \frac{f_{im}(u) S(u, \bar{t}_{m+1}) \pi_{uv} f(\bar{y}_{i,m+1}, U_{i,m+1} = v) g_{i,m+1}(v)}{f(y_i, t_i, d_i)}, \\ m = 1, \ldots, m_i - 1,\ u, v = 1, \ldots, k.$$

Then, we update these parameters as $\pi_u = \frac{1}{n} \sum_{i=1}^n p(U_{i1} = u | y_i, t_i, \delta_i)$, $u = 1, \ldots, k$, and

$$\pi_{uv} = \frac{\sum_{i=1}^n \sum_{m=1}^{m_i - 1} p(U_{im} = u, U_{i,m+1} = v | y_i, t_i, \delta_i)}{\sum_{i=1}^n \sum_{m=1}^{m_i - 1} \sum_{\bar{v}=1}^k p(U_{im} = u, U_{i,m+1} = \bar{v} | y_i, t_i, \delta_i)}, \quad u, v = 1, \ldots, k.$$

The infinitesimal transition matrix Q is obtained from $\mathit{\Pi}_a$ by inverting $\exp(aQ)$.

To update the other parameters, we explicit the expected value of the complete log-likelihood. In particular, regarding the third component about the survival process we have

$$\mathrm{E}\{\ell_3(\theta)\} = \sum_{i=1}^n \mathrm{E}\left\{ \delta_i \log h_0(t_i | U_{im_i}) - \sum_{m=2}^{m_i} H_{m-1}(\bar{t}_m, U_{i,m-1}) - H_m(t_i, U_{im_i}) \right\}.$$

Regarding the derivative of $\ell(\theta)$ with respect to the model parameters, let τ denote any of the elements of θ apart from those involved in the latent process; we apply the general rule

$$\ell(\theta) = \sum_{i=1}^n f(y_i, t_i, d_i)^{-1} \frac{\partial f(y_i, t_i, d_i)}{\partial \tau}.$$

4 Application to MDCM data

We illustrate now the proposed approach using an original study on a cohort of patients affected by MDCM, a primary myocardial disease characterized by left ventricular systolic dysfunction and dilation. For more details on the pathology, see [4].

Prognostic measurements were taken at basal time for $n = 642$ patients, who were followed-up until urgent heart transplant or death occurred. There were 212 events during follow-up, which lasted up to 25 years. If censoring (administrative or due to the event) did not occur, measurement of longitudinal biomarkers were taken at visits scheduled at months 6, 12, 24, 48, 72, 96, and 120. Hence each patient has a maximum of 8 longitudinal measurements, with 79 patients having complete records.

The longitudinal outcome is the New York Hearth Association (NYHA) classification, a direct measure of discomfort caused by the disease. Specifically, for each subject at each follow-up occasion we measure an indicator of being in NYHA class III or IV, indicating the presence of strong limitations to physical activity, and the occurrence of dyspnea and discomfort during ordinary activities or even at rest. For the longitudinal model we parameterize probability of high NYHA class as a function of $t > 0$ (indicating medical treatment according to international guidelines), an indicator of history of heart disease in the family, and the left ventricular Ejection Fraction (EF). The latter is a measure of the proportion of blood that is pumped out of the left ventricle at each heart beat.

It is natural to expect that a continuous-time model is more appropriate for the data at hand than any model assuming latent transitions occurring at visit times. In fact, latent states shall be interpreted as patients' frailty beyond that summarized by the predictors, and changes in disease status (and hence propensity to event and/or change of NYHA class) obviously can occur at any time point and not necessarily on the day of the visit by the doctor. Further, a strong dependence between NYHA class and the event is expected, with patients in NYHA classes III or IV being at higher risk of death.

For interpretability reasons, EF has been centered at 30 (which is believed to be a significant threshold, where EF < 30 indicates patients at risk of heart failure).

For our model fitting procedure we evaluate several values for M, and end up fixing $M = 200$, which is well above values guaranteeing stability of estimates. Using the Bayesian information criterion we select $k = 3$. In order to estimate standard errors we perform a non-parametric bootstrap procedure based on $B = 1000$ replicates. In Table 1 we report parameter estimates for the manifest distribution, along with an indication of significance at the 5% level.

The estimate of Q is better understood after computation of the time-specific transition matrix. For this purpose, we report Figure 1 where the inhomogeneous transition matrix at each time t is reported.

The results indicate an important role of all predictors, with the exception of history of hearth disease for survival. Comparing $k = 1$ with $k > 1$, it is clear that taking into account unobserved heterogeneity leads to a more clear identification of the roles of EF and family history for NYHA classification. The effect of family history doubles when passing from $k = 1$ to $k = 3$, while the effect of each percentage point of EF is almost three times larger. Hence, based on our results, doctors should probably pay more attention to EF and family history than expected when assessing

Table 1 *MDCM data: parameter estimates for the manifest distribution, different values of k. An asterisk indicates statistical significance at the 5% level.*

		k			
	Effect	1	2	3	4
logit NYHA	ξ_1	-0.967*	-4.745*	-4.556*	-6.354*
	ξ_2	-	-0.164	-1.446*	-2.182*
	ξ_3	-	-	2.891*	-1.481*
	ξ_4	-	-	-	2.902*
	$t > 0$	1.047*	2.289*	0.920*	0.847*
	history	0.611*	0.724*	1.169*	1.126*
	EF	0.056*	0.094*	0.134*	0.137*
survival	ϕ	0.000	-0.475*	-0.314*	-0.310*
	history	-0.125	0.031	-0.019	-0.001
	EF	-0.058*	-0.048*	-0.049*	-0.058*
	ν	0.799*	0.781*	0.762*	0.789*

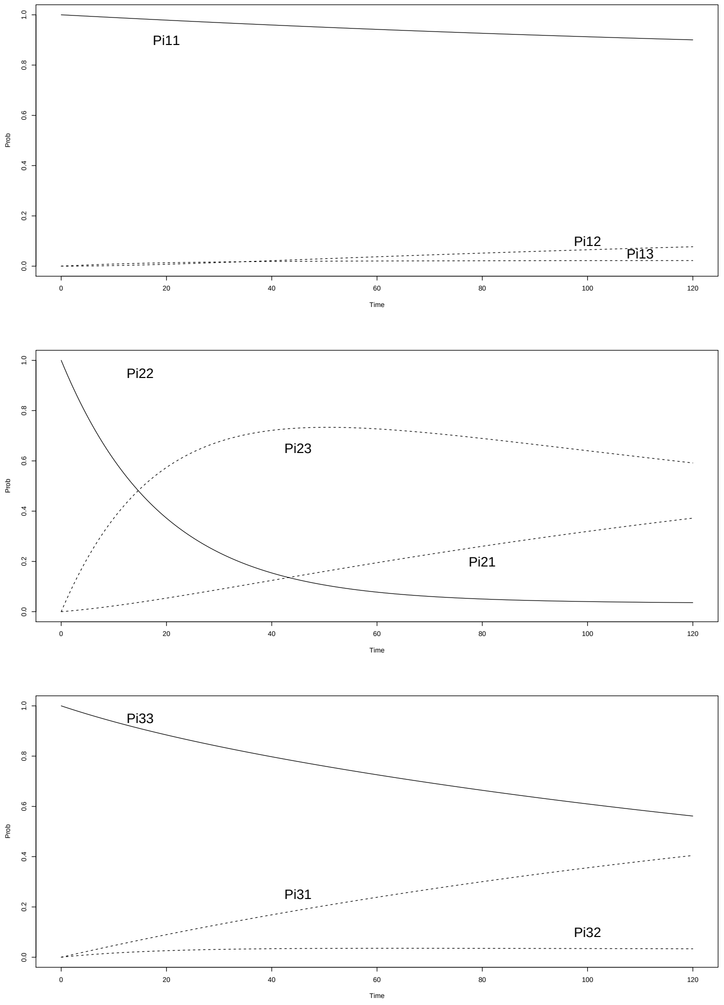
prognosis to high NYHA classes. The estimate of ϕ is negative and significant in all cases, indicating as expected that subjects with, for instance, dyspnea during ordinary activities are at higher risk of death than patients without clear signs of heart insufficiency.

When $k = 3$ three clearly separate groups of patients are identified. Even when they have the same history, EF and timing configuration, patients might be different due to unobserved factors. A group of patients (about 30% at baseline time) is at a very low risk. From Figure 1 it can be observed that this group of patients is slightly stable, with low probability of transitions to different states during follow-up. A second group (about 60%) is at slightly larger propensity to high MDCM at baseline. These patients are very likely to change state during follow-up, with many switchings to an even higher risk (especially in the period 15-40 months from time zero) and the rest switchings to the low risk first latent state (possibly due to successful medical treatment). Finally, a third group of patients is at very high risk of high NYHA class at baseline time. Most of them remain at high risk during follow-up, but a slight proportion switches to better propensity states; surprisingly more often to state 1 than to state 2. This might be due to increased medical attention given to high risk patients.

References

1. Bartolucci, F., Farcomeni, A.: A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. Biometrics **71**, 80–89 (2015)
2. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. Chapman & Hall/CRC Press, Boca Raton, FL (2013)
3. Böckenholt, U.: A latent Markov model for the analysis of longitudinal data collected in continuous time: States, durations, and transitions. Psychological Methods **10**, 65–83 (2005)

Fig. 1 *MDCM data: estimated transition matrix as a function of time t when $k = 3$*



4. Gigli, M., Stolfo, D., Merlo, M., Barbati, G., Ramani, F., Brun, F., Pinamonti, B., Sinagra, G.: Insights into mildly dilated cardiomyopathy: Temporal evolution and long-term prognosis. European Journal of Heart Failure **19**, 531–539 (2017)
5. Liggett, T.M.: Continous Time Markov Processes. American Mathematical Society, Providence (2010)
6. Rizopoulos, D., Verbeke, G., Molenberghs, G.: Shared parameter models under random effects misspecification. Biometrika **95**, 63–74 (2008)

Modelling the latent class structure of multiple Likert items: a paired comparison approach

Modellazione della struttura a classi latenti di item multipli su scala Likert: un approccio basato sul confronto a coppie

Brian Francis, Lancaster University, UK

Abstract The modelling of the latent class structure of multiple Likert items measured on the same response scale can be challenging. The standard latent class approach is to model the absolute Likert ratings, where the logits of the profile probabilities for each item have an adjacent category formulation (DeSantis et al., 2008). We instead propose modelling the relative orderings, using a mixture model of the *relative* differences between pairs of Likert items. This produces a paired comparison adjacent category log-linear model (Dittrich et al., 2007; Francis and Dittrich, 2017), with item estimates placed on a (0,1) "worth" scale for each latent class. The two approaches are compared using data on environmental risk from the International Social Survey Programme, and conclusions are presented.
Abstract *La modellazione della struttura a classi latenti di item multipli misurati sulla stessa scala Likert pu essere problematica. Lapproccio classico a classi latenti si basa sulla modellazione dei punteggi assoluti tramite una formulazione per categorie adiacenti dei logit delle probabilit dei vari profili di risposta per ciascun item (De Santis et al, 2008). In questo contributo si propone, invece, di modellare gli ordinamenti degli item tramite un modello mistura specificato sulle differenze relative tra coppie di item. Il modello risultante un modello log-lineare per categorie adiacenti basato su confronti a coppie (Dittrich et al, 2007; Francis and Dittrich,2017), dove le stime degli item sono poste su una scala di merito (0, 1) per ogni classe latente. I due approcci vengono confrontati e debitamente commentati utilizzando dati relativi al rischio ambientale, provenienti dallInternational Social Survey Programme.*

Key words: latent class, ordinal data, multiple Likert items, paired comparisons

Brian Francis
Department of Maths and Statistics, Fylde College. Lancaster University, Lancaster, LA1 4YF, UK
e-mail: B.Francis@Lancaster.ac.uk

1 Introduction

Collections of multiple Likert items in questionnaires are very common, and are usually used to measure underlying constructs. Scale from the Likert items can be built either through simply adding the item score or through using an IRT model such as a graded response model to build a score. This approach assumes that there is a single underlying construct to the items. The current paper, in contrast, takes a different view. It proposes that there is a latent class structure to the Likert items, with different classes having different patterns of high and low responses. In this approach, score building is not the aim; instead the aim is to understand the various patterns of responses that might exist in the population.

The standard latent class approach to multiple ordinal indicators essentially constructs a polytomous latent class model (Linzer and Lewis, 2011), and constrains the latent class profile probabilities, imposing a linear score ordinal model on them (Magidson and Vermunt, 2004; DeSantis et al., 2008). This results in a latent class adjacent category ordinal model. The method however uses the *absolute* responses, and this has been criticised by some authors, as they state that each respondent has their own way of interpreting the Likert scale. Such interpretation may itself be culturally determined, or may depend on other covariates such as age, gender and so on. For example younger people and males may be more likely to express a firm opinion, using the end categories of a unipolar Likert scale, than older people and females. The alternative is to take a relative approach. While one method of doing this is to standardise the items for each respondent, subtracting the respondent mean. This is unsatisfactory as it ignores the categorical nature of the data. In this paper we instead develop a paired comparisons approach, which produces a worth scale for each latent class, ranking the items in order of preference. The paper compares the two methods and discusses the advantages and disadvantages of each method.

Some common notation is introduced which will be used to develop both models. The Likert items are assumed to be measured on the same response scale with identical labelling; it is assumed that there are H possible ordered response categories taking the values $1,\ldots,H$ for each of the J Likert items indexed by j, and with N respondents indexed by i. $y_{ij}; \quad y_{ij} \in 1,2,\ldots,H$ is defined to be the (ordinal) response given by respondent i to item j. A set of H indicators for each item and respondent with the indicator z_{ijh} taking the value 1 if $y_{ij} = h$ and 0 otherwise.

2 The ordinal latent class model

We first introduce the ordinal latent class model, which models the absolute responses. Let y_{ij} be the ordinal response of respondent i to item j. It is assumed that there are K latent classes. The item response vector for respondent i is

$$\mathbf{y_i} = (y_{i1}, y_{i2}, \cdots, y_{iJ}),$$

Then the ordinal latent class model is defined by:

$$
\begin{aligned}
P(\mathbf{y_i}) &= \sum_{k=1}^{K} \pi(k) P(\mathbf{y_i}|\mathbf{k}) \\
&= \sum_{k=1}^{K} \pi(k) \prod_{j} P(y_{ij}|k) \qquad \text{under conditional independence.}
\end{aligned}
$$

We write

$$
P(y_{ij}|k) = \prod_{h=1}^{H} p_{jkh}^{z_{ijh}}
$$

where p_{jkh} is the probability of observing the ordinal response h for indicator j given membership of latent class k - these are sometimes called the latent class profile probabilities.

Ordinality is imposed by using an adjacent categories ordinal model and we parameterise the model through regression parameters on the logit scale, which separates out the intercept parameter β_{jh} and the class specific parameters β_{jkh} for each item and response category.

$$
\begin{aligned}
\text{logit}(p_{jkh}) &= \beta_{jh} + \beta_{jkh} \\
&= \beta_{jh} + h\beta_{jk} \qquad \text{under a linear score model}
\end{aligned}
$$

The likelihood L is then given by

$$
L = \prod_{i} \sum_{k=1}^{K} \pi(k) P(\mathbf{y_i}|\mathbf{k}).
$$

Model fitting is usually carried out by using the EM algorithm - details are given in Francis et al. (2010) and Aitkin et al. (2014). Determination of the optimal number of classes is commonly achieved by choosing that model which minimises an information criterion, although a wide variety of other methods have been proposed. We have used the BIC in this paper.

3 The latent class ordinal paired comparison model

An alternative to the absolute latent class approach is to work on a **relative scale**. This perhaps is of greater interest. We take a paired comparison approach, using the difference in the ordinal likert responses. This allows the development of a “worth” scale between 0 and 1 awith items placed on this scale. The sum of the item scores is defined to be 1. This section proceeds by developing the ordinal paired comparison model, and then extends that model by adding a mixture or latent class process to the model.

3.1 The ordinal paired comparison model

This model starts by constructing a set of paired comparisons - taking all possible pairs of items and comparing them in turn (Dittrich et al., 2007). For respondent i and for any two items $j = a$ and $j = b$, let

$$x_{i,(ab)} = \begin{cases} \quad h \text{ if item } a \text{ preferred by } h \text{ steps to item } b & = y_{ia} - y_{ib} \\ \quad 0 \qquad\quad \text{if Likert ratings are equal} & = 0 \\ -h \text{ if item } b\text{preferred by } h \text{ steps to item } a & = y_{ia} - y_{ib} \end{cases}$$

The probability for a single PC response $x_{i,(ab)}$ is then defined by

$$p(x_{i,(ab)}) = \begin{cases} \mu_{ab} \left(\frac{\pi_a}{\pi_b}\right)^{x_{i,(ab)}} & : \text{ if } x_{i,(ab)} \neq 0 \\ \mu_{ab}\, c_{ab} & : \text{ if } x_{i,(ab)} = 0 \end{cases}$$

The πs represent the worths or importances of the items, c_{ab} represents the probability of no preference between items a and b and μ_{ab} is a normalising quantity for the comparison ab. Over all items, we now form a pattern vector $\boldsymbol{x}_i$ for observation i with $\boldsymbol{x}_i = (x_{i,(12)}, x_{i,(13)}, \ldots, x_{i,(ab)}, \ldots, x_{i,(J-1,J)})$ and count up the number of responses n_ℓ with that pattern. The **probability** for a certain pattern ℓ is

$$p_\ell = \triangle^* \prod_{a<b} p(x_{ab})$$

where $\triangle^*$ is a constant (the same for all patterns). A log-linear model can now be constructed with observed counts n_ℓ. The **expected counts for a pattern** ℓ **are** defined as $\boxed{m_\ell = n p_\ell}$ where n is the total number of respondents defined by $n = n_1 + n_2 + \cdots + n_\ell + \cdots + n_L$ and where L is the number of all possible patterns.

Taking natural logs, the **log expected counts** are obtained by

$$\ln m_\ell = \alpha + \sum_{a<b} x_{ab}(\lambda_a - \lambda_b) + \mathbf{1}_{x_{ab}=0}\, \gamma_{ab}$$

For $x_{ab} = \quad h$ this is $\quad h(\lambda_a - \lambda_b)$, for $x_{ab} = -h$ this is $h(-\lambda_a + \lambda_b)$ and for $x_{ab} = \quad 0$ this is γ_{ab}.

To show that this is an adjacent categories model, the log odds of a pair for any two adjacent categories on the ordinal scale can be examined - say h and $h+1$. Then, as $m_\ell = np_\ell$, we have

$$\begin{aligned} \ln\left(\frac{m_\ell(h)}{m_\ell(h+1)}\right) &= \ln(\mu_{ab}) + h(\lambda_a - \lambda_b) - \ln(\mu_{ab}) - (h+1)(\lambda_a - \lambda_b) \\ &= \lambda_a - \lambda_b \end{aligned}$$

which is true for any h as long as h or $h+1$ are not zero.

The worths π_j are calculated from the λ_j through the formula

$$\pi_j = \frac{\exp(2\lambda_j)}{\sum_{j=1}^{J}\exp(2\lambda_j)}$$

.

3.2 Extending the model to incorporate latent classes

As before, we assume that there are K latent classes with different preference patterns (the lambdas). The likelihood L becomes:

$$\mathrm{L} = \prod_{\ell}\Big(\sum_{k=1}^{K} q_k \; n \; p_{\ell k}\Big) \quad \text{where } \sum_{\ell} p_{\ell k} = 1 \quad \forall\, k \text{ and } \sum_{k} q_k = 1.$$

$$\ln p_{\ell k} = \alpha + \sum_{a<b} x_{ab}(\lambda_{ak} - \lambda_{bk}) + \mathbf{1}_{x_{ab}=0}\, \gamma_{ab}$$

λ_j is replaced in the model byλ_{jk}, and we now have to additionally estimate the q_k. q_k is the probability of belonging to class k (the mass points or class sizes). Again, we use the EM algorithm to maximise the likelihood, and use the BIC to determine the number of classes. Typically, we need to use a range of starting values to ensure an optimal solution.

4 An Example

Six question items on the topic of environmental danger were taken from the 2000 sweep of the International Social Survey Programme , which focused on issues relating to the environment. As part of this survey, the respondents assessed the environmental danger of a number of different activities and items. The question is reproduced below; each question used the same response scale. The six Likert items are:

c air pollution caused by **c**ars (`CAR`)

t a rise in the world's **t**emperature (`TEMP`)

g modifying the **g**enes of certain crops (`GENE`)

i pollution caused by **i**ndustry (`IND`)

f pesticides and chemicals used in **f**arming (`FARM`)

w pollution of **w**ater (rivers, lakes, ...) (`WATER`)

with the response scale for each of the items as follows:

In general, do you think *item* is

4. extremely dangerous for the environment
3. very dangerous
2. somewhat dangerous
1. not very dangerous
0. not dangerous at all for the environment

Table 1 BIC values from fitting latent class models (a) the standard ordinal LC model and (b) the ordinal PC LC model

	(a) standard ordinal LC model		(b) Ordinal PC LC model	
	absolute		relative	
No. of classes K	BIC	no of parameters	BIC	no of parameters
1	24207.04	24		
2	22680.48	31	6823.11	26
3	22153.75	38	6359.56	32
4	22112.70	45	6204.76	38
5	22097.07	52	6303.71	44
6	22084.99	59		
7	22083.33	66		

Both absolute and relative latent class models are fitted to this data. The standard ordinal latent class model (absolute) was fitted using Latent Gold 5.1 (Vermunt and Magidson, 2013), and the paired comparison ordinal latent class model (relative) was fitted using an extension to the `prefmod` package in R (Hatzinger and Maier, 2017). Both approaches used 20 different starting values to ensure that the global maximum of the likelihood was reached. Table 1 shows the BIC values for both models, for a range of values of K. It can be seen that the standard latent class approach needs either six or seven classes (six classes is chosen here), whereas the paired comparison latent class model gives a minimum BIC for $K = 4$. The smaller number of classes found for the paired comparison approach is perhaps to be expected, as the standard approach needs to model both the absolute level of the Likert responses as well as the differences.

We examine the mean Likert rating for each of the items within each of the latent classes for the standard ordinal latent class model. In contrast, the worths provide the interpretation of the latent classes in the paired comparison LC model. Both plots are shown in Figure 1, which are oriented so that greater dangerousness (or greater danger worth) is towards the top of the plots.

It can be seen that for the standard ordinal latent class model, the first three classes - Class 1 (51%), Class 2 (24%) and Class 3 (12%) - all show little difference between the items, but differ according to their absolute level. The three remaining classes, in contrast, show considerable differences between the items. The paired comparison solution gives a similar story. The largest class shows little difference between the items, with the three remaining classes showing large differences in

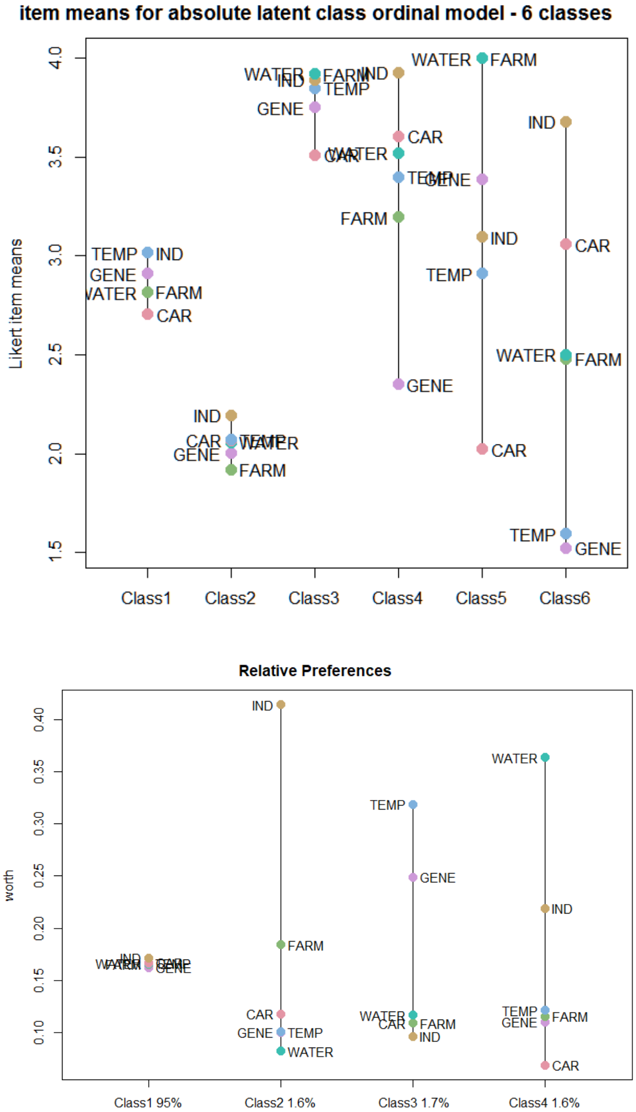


Fig. 1 Item worths for (top) standard ordinal LC model and (bottom) ordinal paired comparison LC model

dangerousness between items. Although the item rankings show some minor differences between the two methods, the results are similar.

5 Discussion and conclusions

This paper has demonstrated that the paired comparison ordinal model can be useful to understand the relative ordering of items in multiple Likert responses when the absolute level of the response is not of interest. The method leads to simpler models, which makes interpretation simpler. There are however some restrictions in using the model. The most important is that all Likert items must be measured on the same response scale. Differences between Likert items only make sense when this is true, and the paired comparison method relies on that. The PC method as currently implemented also assumes equidistance between the Likert categories, and further work is needed to relax this assumption.

References

Aitkin, M., D. Vu, and B. Francis (2014). Statistical modelling of the group structure of social networks. *Social Networks 38*, 74–87.

DeSantis, S. M., E. A. Houseman, B. A. Coull, A. Stemmer-Rachamimov, and R. A. Betensky (2008). A penalized latent class model for ordinal data. *Biostatistics 9*(2), 249–262.

Dittrich, R., B. Francis, R. Hatzinger, and W. Katzenbeisser (2007). A Paired Comparison Approach for the Analysis of Sets of Likert Scale Responses. *Statistical Modelling 7*, 3–28.

Francis, B. and R. Dittrich (2017). Modelling multiple likert items through an adjacent categories ordinal paired comparison model. Presented at the 10th ERCIM comference on computational and Methodological Statistics, London; 16-18 December 2017.

Francis, B., R. Dittrich, and R. Hatzinger (2010). Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do europeans get their scientific knowledge? *The Annals of Applied Statistics 4*(4), 2181–2202.

Hatzinger, R. and M. J. Maier (2017). *prefmod: Utilities to Fit Paired Comparison Models for Preferences*. R package version 0.8-34.

Linzer, D. and J. Lewis (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software 42*(10).

Magidson, J. and J. Vermunt (2004). Latent class analysis. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Chapter 10, pp. 175–198. Thousand Oaks, CA.: Sage Publications.

Vermunt, J. and J. Magidson (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations Inc.

Dealing with reciprocity in dynamic stochastic block models

Analisi della reciprocità in modelli dinamici basati su blocchi stocastici

Francesco Bartolucci, Maria Francesca Marino, Silvia Pandolfi

Abstract For directed relations among a set of nodes with a longitudinal structure, we introduce a dynamic stochastic block model where the blocks are represented by a sequence of latent variables following a Markov chain. Dyads are explicitly modeled conditional on the states occupied by both nodes involved in the relation. We mainly focus on reciprocity and propose three different parameterizations in which: (i) reciprocity is allowed to depend on the blocks of the nodes in the dyad; (ii) reciprocity is assumed to be constant across blocks; and (iii) reciprocity is ruled out. Inference on the model parameters is based on a variational approach. An approximate likelihood ratio test statistic based on the variational approximation is also proposed. This allows us to formally test for both the hypothesis of no reciprocity and that of constant reciprocity with respect to the latent blocks. The proposed approach is illustrated by a simulation study and two applications.

Abstract Si propone un modello dinamico a blocchi stocastici per dati relazionali longitudinali. I blocchi sono identificati da una sequenza di variabili latenti distribuite secondo una catena di Markov. Oggetto dell'analisi è ogni singola diade, la cui distribuzione viene modellata condizionatamente ai blocchi di appartenenza di ciascuno dei nodi coinvolti nella relazione. Particolare enfasi viene posta sullo studio della reciprocità tra i nodi, proponendo tre diverse parametrizzazioni in cui: (i) la reciprocità varia al variare dei blocchi di appartenenza dei nodi, (ii) il livello di reciprocità è constante, (iii) la reciprocità è assente. Per fare inferenze sui parametri del modello si propone l'utilizzo di un approccio variazionale, che rappresenta anche la base per lo sviluppo di un test basato sul rapporto di verosimiglianza ap-

Francesco Bartolucci
Department of Economics, University of Perugia, e-mail: francesco.bartolucci@unipg.it

Maria Francesca Marino
Department of Statistics, Applications, Computer Science, University of Florence, e-mail: mariafrancesca.marino@unifi.it

Silvia Pandolfi
Department of Economics, University of Perugia, e-mail: silvia.pandolfi@unipg.it

prossimato che può essere utilizzato per verificare l'ipotesi di assenza di reciprocità o di reciprocità costante rispetto ai blocchi latenti. L'approccio proposto è illustrato tramite uno studio di simulazione e due applicazioni.

Key words: Dyads, EM algorithm, Hidden Markov models, Likelihood ratio test, Variational inference

1 Introduction

Dynamic Stochastic Block Models (SMBs) [5, 6] represent an important tool of analysis in the dynamic social network literature when the focus is on discovering communities and clustering individuals with respect to their social behavior. According to this specification, nodes in the network can be clustered into k distinct blocks, corresponding to the categories of discrete latent variables which evolve over time according to a first-order Markov chain. The probability of observing a connection between two nodes at a given occasion only depends on their block memberships at the same occasion.

Extending the proposal in [6], we develop an SBM for dynamic directed networks, observed in discrete time, in which the main element of analysis is the *dyad* referred to each pair of nodes. Our main assumption is that of conditional independence between the dyads, given the corresponding latent variables, rather than between univariate responses. This leads to a more flexible specification which does not rely on restrictive assumptions about the reciprocity between nodes. To provide a deeper insight into reciprocity effects, we propose to parametrically specify every dyadic relation by means of a conditional log-linear model. This allows us to effectively distinguish between main and reciprocal effects and improve interpretability of the results. Furthermore, the proposed approach allows us to formulate three different hypotheses: (i) reciprocity may depend on the blocks to which the units involved in the relation belong; (ii) reciprocity is constant across blocks; (iii) reciprocity is absent. Inference on the model parameters is pursued via a variational approach based on a lower bound for the intractable likelihood function. This lower bound also allows us to derive an approximate Likelihood Ratio (LR) test for inferential purposes on the reciprocity parameters.

The reminder of this paper is structured as follows. Section 2 describes the standard dynamic SBM and details the proposed dyadic formulation. In Section 3 we describe the variational approach for model inference. The simulation study and applications are outlined in Sections 4 and 5, respectively. For a detailed description of the proposed approach we remind the reader to [1].

2 Dynamic stochastic block models

Let $Y_{ij}^{(t)}, i, j = 1, \ldots, n, j \neq i$, denote a binary response variable which is equal to 1 if there exists an edge from node i to node j at occasion t, with $t = 1, \ldots, T$, and is equal to 0 otherwise; $y_{ij}^{(t)}$ is used to denote a realization of $Y_{ij}^{(t)}$. We focus on directed networks without self-loops, so that $Y_{ij}^{(t)}$ may differ from $Y_{ji}^{(t)}$ and $Y_{ii}^{(t)}$ is not defined. Moreover, let $\mathbf{Y}^{(t)}$ be the binary adjacency matrix recorded at occasion t, which summarizes the relations between nodes at this occasion and let $\mathscr{Y} = \{\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(T)}\}$ be the set of all network snapshots taken across time occasions.

Standard dynamic SBMs [5, 6] assume that network nodes belong to one of k distinct blocks, which are identified by the node- and time-specific latent variables $U_i^{(t)}$. These are defined on the finite support $\{1, \ldots, k\}$ and are assumed to follow a Markov chain with initial probability vector $\boldsymbol{\lambda} = \{\lambda_u, u = 1, \ldots, k\}$ and transition probability matrix $\boldsymbol{\Pi} = \{\pi_{u|v}, u, v = 1, \ldots, k\}$. A further crucial assumption of dynamic SBMs is that of *local independence*: given the latent variables $U_i^{(t)}$ and $U_j^{(t)}$, the responses $Y_{ij}^{(t)}$ are conditionally independent and follow a Bernoulli distribution with success probability only depending on the blocks of the nodes at occasion t.

We extend the previous formulation by relaxing the local independence assumption and directly accounting for reciprocal effects. For this aim we let $\mathbf{D}_{ij}^{(t)} = (Y_{ij}^{(t)}, Y_{ji}^{(t)})'$ denote the random vector corresponding to the dyad involving nodes i and j at occasion t, with $i = 1, \ldots, n-1$, $j = i+1, \ldots, n$, and $t = 1, \ldots, T$. Conditional on $U_i^{(t)} = u_1$ and $U_j^{(t)} = u_2$, we denote the dyad probabilities by

$$\psi_{y_1 y_2 | u_1 u_2} = p(\mathbf{D}_{ij}^{(t)} = \mathbf{d} \mid U_i^{(t)} = u_1, U_j^{(t)} = u_2),$$

with $u_1, u_2 = 1, \ldots, k$, $y_1, y_2 = 0, 1$, and $\mathbf{d} = (y_1, y_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$. To put emphasis on reciprocity, we use the following log-linear parametrization:

$$\psi_{y_1 y_2 | u_1 u_2} \propto \exp\left[\alpha_{u_1 u_2} y_1 + (\alpha_{u_1 u_2} + \beta_{u_1 u_2}) y_2 + \rho_{u_1 u_2} y_1 y_2\right],$$

where $\beta_{uu} = 0$, for $u = 1, \ldots, k$, $\alpha_{u_1 u_2} = \alpha_{u_2 u_1} + \beta_{u_2 u_1}$, $\beta_{u_1 u_2} = -\beta_{u_2 u_1}$, and $\rho_{u_1 u_2} = \rho_{u_2 u_1}$, for all $u_1 \neq u_2$, to ensure identifiability.

Different versions of the proposed model specification may be obtained by imposing constraints on the $\rho_{u_1 u_2}$ parameters. In particular, under the hypothesis

$$H_I : \rho_{u_1 u_2} = 0, \quad u_1, u_2 = 1, \ldots, k, u_1 \leq u_2,$$

the model directly reduces to the standard dynamic SBM in [6], denoted by M_I, and based on the local independence between the responses $Y_{ij}^{(t)}$. Constant reciprocity effects correspond to the following hypothesis leading to model M_C:

$$H_C : \rho_{u_1 u_2} = \rho, \quad u_1, u_2 = 1, \ldots, k, u_1 \leq u_2.$$

The unconstrained model, with free $\rho_{u_1 u_2}$ parameters, will be denoted by M_U.

3 Variational inference

Let $\mathscr{U} = \{U_i^{(t)},\ i = 1,\dots,n,\ t = 1,\dots,T\}$ denote the overall set of latent variables in the model; based on the assumptions introduced so far, the observed network distribution is obtained by marginalizing out all these latent variables from the joint distribution of $\mathscr{Y}$ and $\mathscr{U}$. This would require the evaluation of a sum over $k^{Tn(n-1)/2}$ terms that, therefore, becomes quickly cumbersome as n, the number of nodes in the network, increases. We then rely on a variational approximation of the intractable likelihood function for making inference on the model parameters.

3.1 Parameter estimation

Let $\boldsymbol{\theta}$ denote the vector of all free model parameters. Following the approach suggested in [5] and [6], we estimate model parameters by a Variational Expectation Maximization (VEM) algorithm [3]. Let $p(\mathscr{U} \mid \mathscr{Y})$ denote the posterior distribution of $\mathscr{U}$ given the observed data $\mathscr{Y}$ and let $Q(\mathscr{U})$ denote its approximation. The VEM algorithm maximizes the following lower bound of the log-likelihood function:

$$\begin{aligned}\mathscr{J}(\boldsymbol{\theta}) &= \log p(\mathscr{Y}) - KL[Q(\mathscr{U}) \,||\, p(\mathscr{U} \mid \mathscr{Y})]\\ &= \sum_{\mathscr{U}} Q(\mathscr{U}) \log p(\mathscr{Y},\mathscr{U}) - \sum_{\mathscr{U}} Q(\mathscr{U}) \log Q(\mathscr{U}),\end{aligned} \qquad (1)$$

where $KL[\cdot \,||\, \cdot]$ stands for the Kullback-Leibler distance. In particular, we use the class of approximate distributions assuming conditional independence between the latent variables in the network given the observed data, namely $Q(\mathscr{U}) = \prod_{i=1}^n \prod_{t=1}^T q(u_i^{(t)}; \boldsymbol{\tau}_i^{(t)})$, where $q(\cdot; \boldsymbol{\tau}_i^{(t)})$ denotes a multinomial probability distribution with parameters 1 and $\boldsymbol{\tau}_i^{(t)} = \{\tau_{iu}^{(t)},\ u = 1,\dots,k\}$. Consequently, function $\mathscr{J}(\boldsymbol{\theta})$ defined in (1) can be rewritten as the sum of the following components:

$$\begin{aligned}\mathscr{J}_1(\boldsymbol{\theta}) &= \sum_{i=1}^n \sum_{u=1}^k \tau_{iu}^{(1)} \log \lambda_u + \sum_{i=1}^n \sum_{t=2}^T \sum_{u=1}^k \sum_{v=1}^k \tau_{iu}^{(t-1)} \tau_{iv}^{(t)} \log \pi_{v|u},\\ \mathscr{J}_2(\boldsymbol{\theta}) &= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{t=1}^T \sum_{u=1}^k \sum_{v=1}^k \tau_{iu}^{(t)} \tau_{jv}^{(t)} \log p(y_{ij}^{(t)}, y_{ji}^{(t)} \mid U_i^{(t)} = u, U_j^{(t)} = v),\\ \mathscr{J}_3(\boldsymbol{\theta}) &= \sum_{i=1}^n \sum_{t=1}^T \sum_{u=1}^k \tau_{iu}^{(t)} \log \tau_{iu}^{(t)}.\end{aligned}$$

To obtain parameter estimates, the VEM algorithm alternates two separate steps until convergence: the E-step and the M-step. At the E-step, we maximize $\mathscr{J}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\tau}_i^{(t)}, i = 1,\dots,n, t = 1,\dots,T$, under the constraint that these quantities are non-negative and $\sum_u \tau_{iu}^{(t)} = 1$. In the M-step, we maximize $\mathscr{J}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Closed form solutions are available for the initial and the transition probabilities of the hidden Markov chain:

$$\lambda_u = \frac{\sum_{i=1}^n \tau_{iu}^{(1)}}{n}, \quad \pi_{v|u} = \frac{\sum_{i=1}^n \sum_{t=2}^T \tau_{iu}^{(t-1)} \tau_{iv}^{(t)}}{\sum_{i=1}^n \sum_{t=2}^T \tau_{iu}^{(t-1)}}.$$

The remaining model parameters are estimated by a standard Netwon-Raphson algorithm for log-linear models.

Two further relevant issues concern the selection of the optimal number of blocks k and the clustering of nodes. Regarding the first aspect, we rely on an Integrated Classification Likelihood (ICL) approach [2]; moreover, nodes may be assigned to one of the k blocks according to a maximum-a-posteriori rule based on the estimated parameters of the multinomial distribution $\hat{\boldsymbol{\tau}}_i^{(t)}$.

3.2 Testing for reciprocity

Reciprocity plays a central role when dealing with directed networks. To test for the absence of reciprocity in the network we propose an approximate LR test based on the lower bound of the likelihood function, $\mathscr{J}(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}_I$, $\hat{\boldsymbol{\theta}}_C$, and $\hat{\boldsymbol{\theta}}_U$ denote the vectors of parameters estimated under models M_I, M_C, and M_U, respectively, with the first model incorporating hypothesis H_I and the second incorporating hypothesis H_C. The proposed test is based on the statistic

$$R_I = -2\left[\mathscr{J}(\hat{\boldsymbol{\theta}}_I) - \mathscr{J}(\hat{\boldsymbol{\theta}}_U)\right].$$

We compare the observed value of this test statistic against a χ^2 distribution with a number of degrees of freedom equal to the number of free parameters in $\boldsymbol{\rho}$, that is, $k(k+1)/2$. In fact, we consider R_I as an approximation of the LR statistic $-2\left[\ell(\hat{\boldsymbol{\theta}}_I) - \ell(\hat{\boldsymbol{\theta}}_U)\right]$ that, under suitable regularity conditions, has null asymptotic distribution of this type.

For a more detailed analysis we also consider the decomposition $R_I = R_C + R_{CI}$, where

$$\begin{aligned} R_C &= -2\left[\mathscr{J}(\hat{\boldsymbol{\theta}}_C) - \mathscr{J}(\hat{\boldsymbol{\theta}}_U)\right], \\ R_{CI} &= -2\left[\mathscr{J}(\hat{\boldsymbol{\theta}}_I) - \mathscr{J}(\hat{\boldsymbol{\theta}}_C)\right], \end{aligned}$$

with R_C being the approximate LR test statistic for testing the constant reciprocity assumption H_C and R_{CI} begin the approximate LR test statistic for testing H_I against H_C. To perform the test, the first statistic is compared against a χ^2 distribution with $k(k+1)/2-1$ degrees of freedom and the second against a χ^2 distribution with one degree of freedom only.

4 Simulation study

To assess the properties of the approximate LR test statistics under different scenarios, we performed an intensive simulation study. We randomly drew $1,000$ samples from a two state ($k = 2$) dynamic SBM for $n = 20, 50, 100$ units observed at $T = 10$ different time occasions. The initial probability vector $\boldsymbol{\lambda}$ has elements 0.4 and 0.6 and the transition matrix $\boldsymbol{\Pi}$ has diagonal elements equal to 0.7 and 0.8. For the parameterization of the dyad probabilities, we set $\boldsymbol{\alpha} = (-2, -3, -1)'$, $\beta_{12} = 0$, and different values for the reciprocity parameter ranging from -2.5 to 2.5.

To evaluate the performance of the proposed inferential procedure, for each simulated scenario we considered the distribution of the approximate LR test statistics R_I and R_{CI}, which allow us to compare the independence model (M_I) against the unconstrained model (M_U) and the constant reciprocity model (M_C), respectively. Results are reported in Tables 1 and 2. The tables also report the simulated type I error probability/power of the above test statistics.

Table 1 *Mean $(\bar{R}_I)$, variance $(Var(R_I))$, and simulated type I error probability/power of the test statistic R_I (p) under different scenarios.*

	$n = 20$			$n = 50$			$n = 100$		
ρ	$\bar{R}_I$	Var(R_I)	p	$\bar{R}_I$	Var(R_I)	p	$\bar{R}_I$	Var(R_I)	p
-1.50	35.76	200.22	0.993	229.01	1167.33	1.000	922.82	5912.34	1.000
-1.00	21.15	109.27	0.975	129.17	557.34	1.000	523.86	2605.36	1.000
-0.50	7.53	28.92	0.737	42.13	178.39	1.000	167.06	728.66	1.000
-0.25	2.89	9.25	0.272	12.71	49.36	0.922	47.97	192.24	1.000
0.00	1.00	2.02	0.052	0.93	1.83	0.045	1.02	2.07	0.052
0.25	3.15	13.70	0.297	14.68	57.46	0.957	57.44	260.67	1.000
0.50	10.65	46.61	0.861	62.66	275.15	1.000	251.68	1130.22	1.000
1.00	45.71	220.82	1.000	293.87	1574.97	1.000	1180.30	7741.51	1.000
1.50	114.84	598.56	1.000	736.43	4668.70	1.000	2977.89	29305.76	1.000

Results confirm our conjecture that, when simulating data from model M_I, both approximate test statistics have a distribution reasonably close to a χ^2 distribution, leading to the rejection of H_I in about 5% of the simulated samples. On the other hand, under the homogeneity assumption for the reciprocity effects, we observe that the power increases as much as ρ deviates from 0. Moreover, the power of the test increases as the sample size n increases.

We also explored the performance of the proposed method for clustering units across time. For this aim, we evaluated the agreement between the estimated and the true latent structure in terms of adjusted rand index [4], obtaining rather encouraging results in comparison to alternative approaches.

Table 2 *Mean $(\bar{R}_{CI})$, variance $(Var(R_{CI}))$, and simulated type I error probability/power of the test statistic R_{CI} (p) under different scenarios.*

	$n = 20$			$n = 50$			$n = 100$		
ρ	$\bar{R}_{CI}$	Var(R_{CI})	p	$\bar{R}_{CI}$	Var(R_{CI})	p	$\bar{R}_{CI}$	Var(R_{CI})	p
-1.50	38.11	200.09	0.988	231.12	1177.18	1.000	924.85	5930.55	1.000
-1.00	23.67	117.15	0.959	131.21	563.74	1.000	525.90	2617.52	1.000
-0.50	10.21	37.30	0.601	44.10	182.99	1.000	169.01	733.51	1.000
-0.25	5.42	14.42	0.227	14.77	54.87	0.825	49.89	194.19	1.000
0.00	3.51	7.52	0.078	3.00	6.04	0.055	2.94	6.73	0.051
0.25	5.76	19.62	0.244	16.55	60.25	0.883	59.50	268.90	1.000
0.50	13.06	53.14	0.743	64.70	281.35	1.000	253.65	1128.36	1.000
1.00	47.91	224.45	1.000	296.14	1623.78	1.000	1182.33	7749.08	1.000
1.50	116.90	603.89	1.000	738.37	4665.33	1.000	2979.98	29281.81	1.000

5 Empirical applications

5.1 Newcomb Fraternity network

The network at issue consists of 14 network snapshots on preference rankings (coded from 1 to 16) from 17 students. Data were collected longitudinally over 15 weeks between 1953 and 1956 among students living in an off-campus (fraternity) house at the University of Michigan. For the purpose of the analysis, we considered the binary socio-matrices $\mathbf{Y}^{(t)}$ derived from these data that are freely available as part of the R package `networkDynamic`. In each network snapshot, $Y_{ij}^{(t)} = 1$ if student i states a ranking for student j equal to 8 or less at time occasion t.

For these data, we estimated the proposed dynamic SBM with $k = 1,\dots,5$, considering the different model specifications corresponding to different hypotheses on reciprocity. The ICL criterion leads to selecting $k = 3$ latent blocks, regardless the chosen model specification. This criterion also identified M_C as the optimal model specification.

Based on the LR test statistic with $k = 3$, we observe that R_I is statistically significant and, therefore, leads to prefer M_U to M_I. A significant test statistic is also observed when comparing M_I against M_C, again with a p-value smaller than 0.001. On the other hand, we conclude that the assumption of constant reciprocity, H_C, cannot be rejected based on the observed data because $p(\chi^2_5 > R_C) = 0.102$, confirming the result based on the comparison of the ICL values.

The parameter estimates suggest the presence of significant mutual relations between students, irrespective to the cluster they belong to ($\hat{\rho} = 1.044$). Regarding the remaining parameters, we observe that students in block 1 are likely to declare a non-reciprocated friendship with nodes belonging to the same block ($\hat{\alpha}_{11} = 1.203$), while null within-group relations are mainly observed for students belonging to block 2 ($\hat{\alpha}_{22} = -1.069$). A non-significant value is observed for α_{33}. Regarding the estimated initial and transition probabilities of the hidden Markov chain, cluster 2 is the most likely at the beginning of the observation period ($\hat{\lambda}_2 = 0.48$). Moreover, estimated transitions show quite persistent hidden states.

5.2 Enron email network

The second example is based on a dynamic network derived from the Enron corpus, consisting of a large set of email messages that was made public during the legal investigation concerning the Enron corporation. The dataset concerns 184 Enron employees; we considered communications recorded between April 2001 and March 2002 and we built an email network for each month, so that the dynamic network has 12 time points. In this application, $Y_{ij}^{(t)} = 1$ if user i sent at least one email message to user j during the t-th month of the analyzed time window, with $i = 1,\ldots,183$, $j = i+1,\ldots,184$, and $t = 1,\ldots,12$.

We estimated a dynamic SBM with a varying number of blocks ($k = 1,\ldots,7$). ICL values lead to selecting a model with $k = 6$ hidden states for all considered parameterizations. Based on the same index, we selected the unconstrained model M_U, with reciprocity parameters depending on the latent blocks. Even in this case, we may validate the results by comparing the values of the approximate LR statistics. From this comparison, when $k = 6$, we observe that the hypothesis of absence of reciprocity, H_I, is strongly rejected by both tests based on R_I and R_{CI}. Moreover, the observed value of the test statistic R_C allows us to confirm that the unconstrained model has to be preferred to the other model specifications, due to a very low p-value. Accordingly, in this application, we conclude that reciprocal relations are statistically significant, and that they depend on the latent blocks of the nodes.

References

1. Bartolucci, F., Marino, F., Pandolfi, S.: Dealing with reciprocity in dynamic stochastic block models. Comput. Stat. Data An. **123**, 86–100 (2018)
2. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE T. Pattern Anal., **22** 719–725 (2000)
3. Daudin, J.-J., Picard, F., Robin, S.: A mixture model for random graphs. Stat. Comput. **18**, 173–183 (2008)
4. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
5. Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. J. R. Stat. Soc. B **79**, 1119–1141 (2017)
6. Yang, T., Chi, Y., Zhu, S., Gong, Y., Jin, R.: Detecting communities and their evolutions in dynamic social networks - a Bayesian approach. Mach. Learn. **82**, 157–189 (2011)

Causality patterns of a marketing campaign conducted over time: evidence from the latent Markov model

Effetti causali di una campagna di marketing protratta nel tempo attraverso un modello con processo di Markov latente

Fulvia Pennoni, Leo Paas and Francesco Bartolucci

Abstract Many statistical methods currently employed to evaluate the effect of a marketing campaign in dealing with observational data advocate strong parametric assumptions to correct for endogeneity among the participants. In addition, the assumptions compromise the estimated values when applied to data in which the research expects endogeneity but this is not realized. Based on the recent advances in the literature of causal models dealing with data collected across time, we propose a dynamic version of the inverse-probability-of-treatment weighting within the latent Markov model. The proposal, which is based on a weighted maximum likelihood approach, accounts for endogeneity without imposing strong restrictions. The likelihood function is maximized through the Expectation-Maximization algorithm which is suitably modified to account for the inverse probability weights. Standard errors for the parameters estimates are obtained by a nonparametric bootstrap method. We show the effects of multiple mail campaigns conducted by a large European bank with the purpose to influence their customers to the acquisitions of the addressed financial products.

Keywords: causal latent Markov model, customer relationship management, direct marketing, Expectation-Maximization algorithm

Fulvia Pennoni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy, e-mail: fulvia.pennoni@unimib.it

Leonard J. Paas
School of Communication, Journalism and Marketing, Massey University, Auckland, New Zealand, e-mail: L.J.Paas@massey.ac.nz

Francesco Bartolucci
Department of Economics, University of Perugia, Perugia, Italy, e-mail: francesco.bartolucci@unipg.it

1 Background

Firms operationalizing relationship marketing strategies often require insights into long-term developments of interactions with customers. Marketing response models have been developed to relate the customer decisions to the marketing efforts (Manchanda et al., 2004). For example, Li et al. (2005, 2011) propose to assess the consumer financial product portfolios at a bank, by considering key customer lifetime value indicators, such as customer profitability and retention. Latent Markov (LM) models (Bartolucci et al., 2013) have been extensively applied to incorporate switching of customers between segments over time (see, among others, Paas et al., 2007; Schweidel et al., 2011). They have been employed to potentially identify cross-sell opportunities.

In this context, the portfolios development can be influenced by the marketing activities operated by the firm or the bank that may include, among many strategies, also the direct mail channel. This kind of activity is often strategically based on customer characteristics or behaviors. When they are incorporated as covariates in an LM model, endogeneity can occur, since the bank is using the information on, for example, customer demographic characteristics and product ownership when targeting customers for campaigns. Among the models which have been applied to cope with this feature, the Gaussian copulas (Park and Gupta, 2012) and the latent instrumental variable models (Ebbes et al., 2005) are suitable to cope with endogeneity. However, these alternative models need strong assumptions which are not often satisfied. The first one assumes a Gaussian distribution for the errors and non-Gaussian distribution for the endogenous regressors. The latent instrumental variable models assume a discrete component among the endogenous regressors.

We propose an alternative model which is an extension of that proposed by Bartolucci et al. (2016) to address causal effects within observational studies. Consistently with the potential outcome (PO) framework (Rubin, 2005), the causal effects are considered under a counterfactual scenario because individuals cannot be observed under multiple treatments, but only for the treatment they effectively receive. We apply the inverse-probability-of-treatment weighting (IPW), based on the reciprocal of the probability for an individual of receiving the treatment s/he effectively received to alleviate the potential bias due to endogeneity (see also, Skrondal and Rabe-Hesketh, 2014). This accounts for dynamic counterfactuals and sequential endogeneity since treatments at different time occasions may induce different responses. Therefore, we propose to apply a weighted maximum likelihood approach to obtain a consistent estimator for the causal effect of interest.

The remainder of the paper proceeds as follows. In Section 2 we illustrate the proposed causal LM model. In Section 3 we present the empirical application related to a repeated direct mail campaign of an important bank to enlarge the customers' financial product portfolios. In Section 4 we report the results and in Section 5 we offer some concluding remarks.

2 The proposed causal LM model

We consider an observational longitudinal study involving n customers of a firm or a bank. Let $\boldsymbol{Y}_{it}$ be the observed binary vector of responses for customer i, $i = 1, \ldots, n$, at each time occasion t, $t = 1, \ldots, T$, which is a vector indicating the product portfolio of the customer owned at time t. Let $\boldsymbol{X}_{it}$ be a column vector of the available time-varying customer's covariates, some of which may have been selected by the bank to establish the marketing campaign.

Under the potential outcome framework we define a discrete variable to denote the time-varying treatment Z_{it} for customer i and time occasion t that is an ordinal variable with levels in $\{1, \ldots, l\}$. We denote by $H_{it}^{(z)}$ the discrete latent variable that is customer- and time-specific when s/he has received treatment z. Its distribution has support $\{1, \ldots, k\}$ and the support points are discrete finite values as each customer has many potential outcomes which are "potential versions" associated with each marketing intensity z administered at each time occasion. We also allow that the treatment can be absent when the first level of z is set equal to zero. For instance, $H_{it}^{(3)}$ corresponds to the latent state of prospect i at time t if s/he has received the treatment intensity equal to three marketing stimuli up to time t. The time sequence of these variables for customer i is collected in vector $\boldsymbol{H}_i^{(z)} = (H_{i1}^{(z)}, \ldots, H_{iT}^{(z)})'$.

In the following, we denote by lower case letters the realized values of the variables. Accordingly, the time-specific vectors of responses are denoted by $\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}$ and we assume that every customer has a positive probability $p(Z_{it} = z|\boldsymbol{x}_{it})$ of receiving a marketing stimulus at each time period. We also assume that the marketing stimuli at each time period are independent of the potential outcomes given the pre-treatment covariates, that is, $Z_{it} \perp\!\!\!\perp H_{it}^{(z)}|\boldsymbol{X}_{i,t-1}$ for $i = 1, \ldots, n$, $t = 1, \ldots, T$.

The distribution of each $H_{it}^{(z)}$ variable is defined according to a of first-order Markov chain, where the initial probabilities of the latent chain define the potential products portfolios according to the received treatment at the first occasion. These are modeled by the following baseline-category logit model because the latent states do not show a precise order:

$$\log \frac{p(H_{i1}^{(z)} = h)}{p(H_{i1}^{(z)} = 1)} = \alpha_h + \boldsymbol{d}(z)'\boldsymbol{\beta}_h, \quad h = 2, \ldots, k. \tag{1}$$

In the previous expression, α_h is the state specific intercept, $\boldsymbol{\beta}_h = (\beta_{h2}, \ldots, \beta_{hl})'$ is a column vector of $l-1$ parameters, and $\boldsymbol{d}(z)$ is a column vector of $l-1$ zeros with the $(z-1)$-th element equal to 1 if $z > 1$. The element β_{hz} of $\boldsymbol{\beta}_h$ for $z > 1$, corresponds to the effect at the first occasion of the z-th treatment with respect to the effect of the first level of the treatment ($z = 1$). This coefficient is the average effect of the corresponding treatment for the population of interest at the beginning of the campaign.

A similar parameterization, which is not reported here, is employed for the transition probabilities. The parameters referred to the distribution of every observed response variable conditional to the POs identify the products owned by customers at each time occasion and they are useful to cluster the customers into homogenous segments. If required, the latter assumption may be relaxed in different ways (see Bartolucci et al., 2013, for further details).

In order to introduce the IPW estimator, we estimate the following multinomial logit model

$$\log \frac{p(Z_{it} = z|\boldsymbol{x}_{it-1})}{p(Z_{it} = 1|\boldsymbol{x}_{it-1})} = \eta_z + \boldsymbol{x}'_{it-1}\boldsymbol{\lambda}_z, \quad z = 1, \dots, l, \quad t = 1, \dots, T,$$

where η_z and $\boldsymbol{\lambda}_z$ are the intercept and the regression parameters referred to each treatment level. On the basis of the parameter estimates, we compute the customer weights referred to each time period as

$$\hat{w}_{it} = n\frac{1/\hat{p}(z_{it}|\boldsymbol{x}_{it-1})}{\sum_{i=1}^{n} 1/\hat{p}(z_{it}|\boldsymbol{x}_{it-1})}, \quad i = 1, \dots, n, \quad t = 1, \dots, T,$$

and the overall individual weights are provided by products of the weights concerning each year of the campaign

$$\hat{\boldsymbol{w}}_i = \prod_{t=1}^{T} \hat{w}_{it} \quad i = 1, \dots, n.$$

Given the observed data, model estimation is performed by maximizing the weighted log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \hat{\boldsymbol{w}}_i \log \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \log p(\boldsymbol{y}_{i1}, \dots, \boldsymbol{y}_{iT}|z_{it}),$$

where $\boldsymbol{\theta}$ is the overall vector of parameters. This log-likelihood is maximized by using a modified version of the Expectation-Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977). The algorithm is based on the complete data log-likelihood $\ell^*(\boldsymbol{\theta})$ which involves the weighted frequencies of each latent states (see also Bartolucci et al., 2013, for details). The steps of the algorithm are the following:

- **E-step**: it computes the conditional expected value of each frequency involved in the complete data log-likelihood. It requires to compute the posterior probabilities of the latent variable given the weights, the observed responses, and the treatment sequence;
- **M-step**: it maximizes the complete-data log-likelihood where the frequencies are obtained by the corresponding expected values calculated at the E-step.

A non-parametric bootstrap algorithm (Davison and Hinkley, 1997) is employed to obtain the standard errors for the model parameters. We resample customers, with their observed pre-treatment covariates and outcomes, a suitable number of times; then we estimate the model for each generated sample in accordance with the estimated customer's time-varying weights. The number of latent states is selected by the Bayesian Information Criterion (BIC; Schwarz, 1978).

3 Application

We analyze a sample of 49,967 customers aged 18 years and older, provided by a large anonymous European bank that conducted a multiple direct mail campaign over a long period of time. We evaluate the effectiveness of the efforts made by the bank in terms of developments of the customers's portfolios. The following products could be affected by the marketing campaign: *loans, credit cards, checking accounts, investment products, mortgages, savings accounts, and a paid phone service enabling customers to gain insights into their account balances*.

Direct mail was the dominant channel for making product offers to customers in the 2000-2003-period to which the data refer. Table 1 shows the proportions of customers involved in each year of the campaign. We consider four treatment levels varying from none to more than six mails sent to each customer. As shown in Table 1, the sequential treatments have been administered according to an increasing intensity to the customers.

Direct mail intensity	2001	2002	2003
none	0.317	0.237	0.186
1-2	0.311	0.323	0.248
3-5	0.222	0.211	0.245
≥ 6	0.150	0.229	0.321

Table 1 *Observed proportions of direct mail intensity by year.*

The available time-varying covariates influencing the treatment probability are the following: customer's age, money s/he has transferred each year on the account, number of transactions made annually, and annual bank' profits on each customer. It is important to note that the endogeneity arises mainly because the customer selection to which address the campaign made by the managers is not done by randomized methods. Instead the bank's managers choose by simple common sense or by employing logistic regression models or tree-based algorithms.

4 Results

The BIC index favors the model with seven latent states when we estimate the causal LM model for a number of latent states ranging from 1 to 8. The model-based approach defines the following customer segments corresponding to the latent states and characterized on the basis of the conditional response probabilities given the latent states:

1. "none" (0%)
2. "checking account only" (40%)
3. "savers'segment" (5%)

4. “investors” (23%)
5. “phone service customers” (11%)
6. “loans customers” (16%)
7. “actives” (5%)

where the reported percentages are referred to the proportion of customers in the corresponding segment after the first year of the campaign. In this way, we also identify a segment (denoted with 1) as that of individuals which have churned the bank during the first period.

In Table 2 we show the estimated initial probabilities which are averaged according to the intensity of the treatment (see Table 1). They are obtained by considering the estimated average treatment effect on the initial segments as in equation (1). From these results we notice that if the campaign is not conducted, or it is conducted with a low mail intensity, there is a higher probability for a customer to be allocated in segment 2 “checking account only” at the beginning of the period respect to a more intensive campaign. Note that, except for segment 2, all the probabilities referred to intensities (3-5) and (≥ 6) are higher than those of lower treatment levels. This indicates that the treatment enhances the probability to be more active at bank, especially to become “investor” (segment 4).

	Latent state (h)						
Direct mail intensity	1	2	3	4	5	6	7
none	0.00	0.48	0.21	0.03	0.10	0.04	0.15
1-2	0.00	0.44	0.22	0.04	0.10	0.04	0.15
3-5	0.00	0.38	0.25	0.05	0.11	0.05	0.15
≥6	0.00	0.18	0.30	0.11	0.13	0.07	0.21

Table 2 *Estimated initial probabilities for each treatment level under the causal LM.*

	Latent state (h)						
$\bar{h}$	1	2	3	4	5	6	7
1	1.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.072	0.808	0.030	0.000	0.056	0.020	0.015
3	0.017	0.031	0.894	0.000	0.042	0.003	0.014
4	0.023	0.000	0.000	0.967	0.006	0.000	0.004
5	0.011	0.005	0.001	0.000	0.976	0.000	0.009
6	0.004	0.014	0.004	0.000	0.004	0.972	0.002
7	0.009	0.006	0.005	0.001	0.006	0.000	0.974

Table 3 *Estimated transition probabilities of the causal LM model with 7 latent states.*

In Table 3 we report the estimated transition probabilities for customers in the highest treatment level (intensity ≥6 mails). The other three tables concerning the transition probabilities of the latent states of lower mails intensity are not reported here. Based on all relevant tables, we notice that when treatment is intensive, the customers in segment 2 “checking account only” show a relative high probability to switch towards the segments in which customers own multiple products, that is, 5.6% for switching into latent state 4 “investment inclined customers”, 2.0% for switching into segment 6 “loan customers”, and 1.5% for switching into segment 7 of “active customers”.

From the other estimated matrices relative to a low mail intensity or no treatment at all, we notice that mainly for the customers in segment 5 “loan customers ” and 6 “mortgage customers” there is a higher probability to switch into segment 1 of “inactive customers” or segment 2 of “checking account only” compared to those observed in Table 3.

5 Conclusions

The findings gained by the results illustrated in Section 4 provide a number of salient managerial implications. We provide a short summary through the following three advices:

- ensure that each customer receives at least one direct mailing each year to reduce churn probabilities;
- perform an intensive campaign towards customers in segment 6 “lenders” to reduce the probability of terminating the usage of the loan at the same bank;
- send at least six direct mails each year to customers in segment 2 “checking account only”, to enhance their probability to switch into the more active states, emphasizing the loan and the online phone service, as the acquisition of these financial products is most strongly influenced by the direct mail channel.

Direct mail mostly affects forward switching probabilities for customers with only a checking account (segment 2) into multiple segments not characterized by high ownership of savings accounts and mortgages. This leads to two additional suggestions: (*i*) assess through experiments whether innovative direct mailings can enhance switching probabilities forward that are currently low. Also, other channels can be employed to assess whether these low transition probabilities can be affected (e.g., the personal sales channel); (*ii*) find how the direct mail channels or other marketing communication channels can be employed to effectively market mortgages.

The innovative approach we propose to address endogeneity has important features which may help managers to make better decisions on how many direct marketing mailings each individual customer should receive. Therefore, the proposed causal LM model may be fruitfully applied in many other potential marketing contexts over that illustrated in the applicative example.

References

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC press, Boca Raton.

Bartolucci, F., Pennoni, F., and Vittadini, G. (2016). Causal latent Markov model for the comparison of multiple treatments in observational longitudinal studies. *Journal of Educational and Behavioral Statistics*, 41:146–179.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, MA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society - Series B*, 39:1–38.

Ebbes, P., Wedel, M., Böckenholt, U., and Steerneman, T. (2005). Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics*, 3:365–392.

Li, S., Sun, B., and Montgomery, A. L. (2011). Cross-selling the right product to the right customer at the right time. *Journal of Marketing Research*, 48:683–700.

Li, S., Sun, B., and Wilcox, R. T. (2005). Cross-selling sequentially ordered products: An application to consumer banking services. *Journal of Marketing Research*, 42:233–239.

Manchanda, P., Rossi, P. E., and Chintagunta, P. K. (2004). Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research*, 41:467–478.

Paas, L. J., Vermunt, J. K., and Bijmolt, T. H. (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society - Series A*, 170:955–974.

Park, S. and Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31:567–586.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling and decisions. *Journal of the American Statistical Association*, 100:322–331.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Schweidel, D. A., Bradlow, E. T., and Fader, P. S. (2011). Portfolio dynamics for customers of a multiservice provider. *Management Science*, 57:471–486.

Skrondal, A. and Rabe-Hesketh, S. (2014). Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *Journal of the Royal Statistical Society - Series C*, 63:211–237.

Complex Spatio-temporal Processes and Functional Data

Clustering of spatio-temporal data based on marked variograms

Clustering di dati spazio-temporali basato su marked variograms

Antonio Balzanella and Rosanna Verde

Abstract This paper deals with the clustering of data generated by spatio-temporal point processes. The interest on this topic is motivated by the recent availability of spatio-temporally indexed data in several applicative fields like seismology, climatology, economics, social sciences. The data we analyse is a collection of instantaneous events, each occurring at a given spatial location. We introduce a strategy which finds a partition of the individuals into homogeneous clusters considering the space-time interactions. We transform the spatio-temporal point process into two marked point processes, considering the times as marks of the spatial point process and the locations as marks of the times. This allows to use the marked variograms to describe the second-order characteristics of the individuals, in time and space. We propose a k-means like algorithm which uses the marked variograms as cluster representative and performs the allocation to clusters evaluating the contribution of each individual to the definition of the marked variograms. This allows to get clusters of individuals which are homogeneous in terms of space-time interactions.
Abstract *Il presente articolo è incentrato sulle tecniche di clustering per dati generati da processi di punto spazio-tempo. L'interesse sulla tematica è dovuto alla crescente disponibilità di dati spazio-tempo in ambiti applicativi quali la sismologia, la climatologia, le scienze economiche, le scienze sociali. La tipologia di dati che si propone di analizzare è una collezione di eventi istantanei verificatisi, ciascuno, in una specifica locazione spaziale. In tale contesto, si propone una strategia di clustering finalizzata all'individuazione di una partizione degli individui in cluster omogenei, considerando le interazioni spazio-tempo. Si propone di trasformare il processo di punto spazio-tempo in due processi di punto marcati nei quali si considera il tempo, come marcatore di un processo di punto spaziale e lo spazio, come marcatore di un processo di punto temporale. Ciò consente di utilizzare i var-*

Antonio Balzanella
Università della Campania L. Vanvitelli, e-mail: antonio.balzanella@unicampania.it

Rosnna Verde
Università della Campania L. Vanvitelli e-mail: rosanna.verde@unicampania.it

iogrammi marcati per descrivere le caratteristiche di secondo ordine degli individui nel tempo e nello spazio. Nella strategia proposta, i variogrammi marcati vengono utilizzati come centroidi di un algoritmo di tipo k-means che effettua l'allocazione ai cluster sulla base del contributo di ciascun individuo alla definizione della struttura di variabilità spazio-temporale sintetizzata dai variogrammi marcati. Tale approccio consente di ottenere cluster di individui che sono omogenei in termini di interazioni spazio-tempo.

Key words: clustering, spatio-temporal point process, mark variogram

1 Introduction

In this paper we deal with the statistical analysis of a collection of data generated by a spatio-temporal point process. Single observations are instantaneous events, each occurring at a given spatial location with a given associated time stamp. Typical applications are the analysis of seismic events, the monitoring of crimes or diseases.

Recent proposals in [2][4] provide, a review of the statistics for analysing such kind of data and new methods for exploring the spatio-temporal data structure by means of suitable second-order statistics.

We focus on clustering the data points into homogeneous groups keeping into consideration their spatio-temporal interactions.

Consistently with [4], we transform the spatio-temporal point process into two marked point processes. We can consider the times as marks of the spatial point process of point locations and the locations as marks of the times.

For evaluating the second-order characteristics of the two marked point processes, we use the mark variograms. The latter, provide a measure of the data interaction over the whole geographic area, for the considered time window.

In this paper, we propose a decomposition of the mark variograms into a set of local measures of spatio-temporal interaction (LMSTI), which allow to evaluate the contribution of each data point to the definition of the mark variogram.

Since we use two mark variograms for evaluating the second order characteristics of the spatio-temporal point process, we will associate each data point to two LMSTI, the first one uses the time as mark, the second one uses the location as mark.

Our idea, is to describe each data point through the two mark LMSTI and to perform a *k*-means like algorithm on this new data description. This allows to get clusters of data points which interact similarly with the other data points, accounting both for the space and time dimension.

2 Main notations and definitions

We consider a spatio-temporal point process $X = \{(s_1,t_1),\ldots,(s_i,t_i),\ldots,(s_n,t_n)\}$, where $s_i \in D \subseteq \Re^2$ is a spatial location and $t_i \in T \subseteq \Re^+$ the corresponding time. We assume that the point process is orderly, that is, coincident points cannot occur.

Similarly to [4], the random process X can be transformed into two mark point processes X_t and X_s using, respectively, the time and the locations as marks.

This allows to use the classic mark variogram [3] for measuring the interactions in the two processes. In particular, if we consider the time, as mark, the variogram γ_{sp} is:

$$\gamma_{sp}(h) = \frac{1}{2}\mathbf{E}\left[(t_i - t_j)^2 \ \mid t_i, t_j \in T\right] \tag{1}$$

where $h = \left\|s_i - s_j\right\|$.

Similarly, if we consider the spatial location as mark, the variogram γ_{ti} is:

$$\gamma_{ti}(v) = \frac{1}{2}\mathbf{E}\left[(s_i - s_j)^2 \ \mid s_i, s_j \in D\right] \tag{2}$$

where $v = \left\|s_i - s_j\right\|$.

Similarly to the classic variogram used in geostatistics ([1]), the mark variogram gives information on the correlations in the marked point process.

In order to estimate $\gamma_{sp}(h)$ we can use the expression in [4]:

$$\hat{\gamma}_{sp}(h) = \frac{\sum_{s_i,s_j \in D} \frac{1}{2}(t_i - t_j)^2 k_e(\left\|x_i - x_j\right\| - h)}{\sum_{s_i,s_j \in D} k_e(\left\|x_i - x_j\right\| - h)} \tag{3}$$

where k_e is is a one-dimensional kernel function with bandwidth e.

Similarly, $\gamma_{ti}(v)$ can be estimated by:

$$\hat{\gamma}_{ti}(v) = \frac{\sum_{t_i,t_j \in T} \frac{1}{2}(s_i - s_j)^2 k_g(\left\|t_i - t_j\right\| - v)}{\sum_{t_i,t_j \in T} k_g(\left\|t_i - t_j\right\| - v)} \tag{4}$$

where k_g is a one-dimensional kernel function with bandwidth g.

In order to introduce our clustering algorithm, we associate to each data point two functions which measure the interaction of the data point (s_i,t_i) with the other data points. The two functions Δ^i_{sp} and Δ^i_{ti} are obtained by decomposing the mark variograms γ_{sp} and γ_{ti}:

$$\Delta^i_{sp}(h) = \sum_{s_j \in D} \frac{1}{2}(t_i - t_j)^2 k_e(\left\|x_i - x_j\right\| - h) \tag{5}$$

$$\Delta^i_{ti}(v) = \sum_{t_j \in T} \frac{1}{2}(s_i - s_j)^2 k_g(\left\|t_i - t_j\right\| - v) \tag{6}$$

where $h = \|s_i - s_j\|$ and $v = \|s_i - s_j\|$.

3 Clustering of local measures of spatio-temporal interaction

By means of Δ^i_{sp} and Δ^i_{ti} we get a new description of the observed data points which we use in a k-means like algorithm for obtaining a partitioning P of the data points (s_i, t_i) into K clusters C_k (with $k = 1, \ldots, K$).

In order to reach our aim, we propose to minimize the following objective function:

$$J(P, L) = \sum_{k=1}^{K} \sum_{(s_i,t_i)\in C_k} d^2(\Delta^i_{sp}(h); \overline{\Delta^k_{sp}}(h)) + d^2(\Delta^i_{ti}(v); \overline{\Delta^k_{ti}}(v)) \tag{7}$$

where:

$\overline{\Delta^k_{sp}}(h)$ and $\overline{\Delta^k_{ti}}(v)$ are the prototypes of the cluster C_k;
L is the matrix of prototypes;
$d^2(.)$ is the squared euclidean distance.

The minimization of the objective function is performed by an iterative algorithm which starts from an initial random partitioning of the data points and, then, alternates a representation step and an allocation step until a stable value of $J(P, L)$ is reached.

In the representation step, the prototype $\overline{\Delta^k_{sp}}$, $\overline{\Delta^k_{ti}}$ of each cluster are computed by:

$$\overline{\Delta^k_{sp}} = \frac{\sum_{(s_i,t_i)\in C_k} \Delta^i_{sp}}{|C_k|} \tag{8}$$

$$\overline{\Delta^k_{ti}} = \frac{\sum_{(s_i,t_i)\in C_k} \Delta^i_{ti}}{|C_k|} \tag{9}$$

that is, they are the average of the LMSTI functions allocated to a cluster and can be seen as a summary of the spatio-temporal interaction structure in each cluster.

In the allocation step, the data points (s_i, t_i) are allocated to the cluster C_k, if the squared Euclidean distance between LMSTI functions Δ^i_{sp} and Δ^i_{ti} and the cluster prototype $\overline{\Delta^k_{sp}}$, $\overline{\Delta^k_{ti}}$ is minimum, through the following rule:

$$d^2\left(\Delta^i_{sp}(h); \overline{\Delta^k_{sp}}(h))\right) + d^2\left(\Delta^i_{ti}(v); \overline{\Delta^k_{ti}}(v))\right) < \\ d^2\left(\Delta^i_{sp}(h); \overline{\Delta^{k'}_{sp}}(h))\right) + d^2\left(\Delta^i_{ti}(v); \overline{\Delta^{k'}_{ti}}(v))\right) \quad \forall k \neq k' \tag{10}$$

Since the centroids of each cluster are the average of the allocated items, similarly to the classic k-means, the optimized criterion decreases with the iterations until it reaches a stable value.

4 Preliminary results on real data

In this section, we introduce some preliminary results on real data. The dataset collects earthquakes in Japan from 1926 to 2005. The dataset contains 948 epicenters of earthquakes with a magnitude higher than 4.5. The data was obtained from `http://users.jyu.fi/~penttine/ppstatistics/data/Earthquakes_Fig_6_21.txt`. We consider D as a planar rectangle of side lengths $a = 543km$ and $b = 556km$, and a time interval T of length 12000 days.

We show in figure 3 the results of the partitioning into $K = 4$ clusters of the dataset of seismic events.

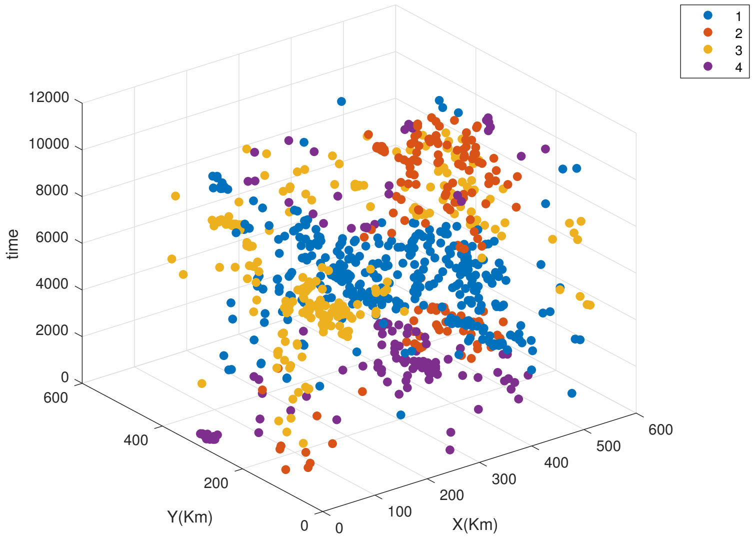


Fig. 1 Partitioning of the seismic events into $K = 4$ clusters.

Still, we show the LMSTI functions and the their average (cluster prototype) for each cluster and for the space and time domain in fig.

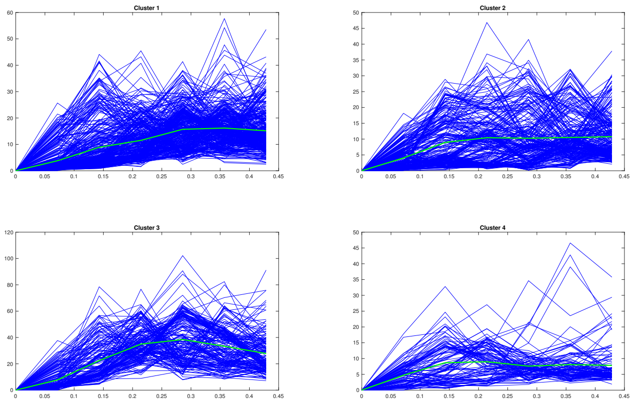


Fig. 2 LMSTI functions and cluster centroid in Space domain

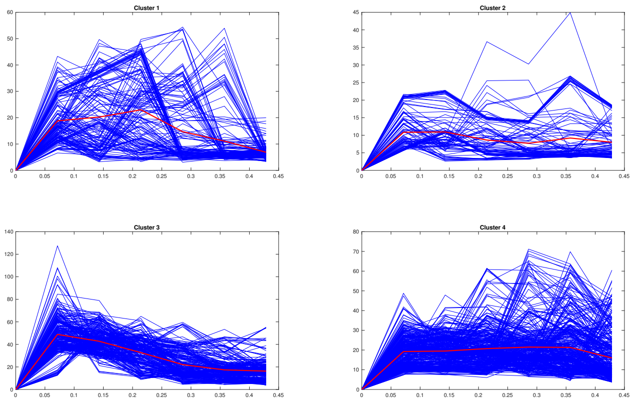


Fig. 3 LMSTI functions and cluster centroid in Time domain

References

1. Cressie N. Statistics for Spatial Data. John Wiley & Sons, New York (1993)
2. González J. A., Rodríguez-Cortés F.J., Cronie O., Mateu J.: Spatio-temporal point process statistics: A review. Spatial Statistics, **18**, Part B, 505–544 (2016) doi: https://doi.org/10.1016/j.spasta.2016.10.002.
3. Illian J., Penttinen A., Stoyan H., Stoyan D. In: Statistical Analysis and Modelling of Spatial Point Patterns, John Wiley & Sons, Chichester (2008)
4. Stoyan D., Rodríguez-Cortés F. J., Mateu J., Gille W.:Mark variograms for spatio-temporal point processes. Spatial Statistics, **20**, 125–147 (2017) doi: https://doi.org/10.1016/j.spasta.2017.02.006.

Space-time earthquake clustering: nearest-neighbor and stochastic declustering methods in comparison

Clustering spazio-temporale di terremoti: i metodi nearest-neighbor e di declustering stocastico a confronto

Elisa Varini, Antonella Peresan, Renata Rotondi, and Stefania Gentili

Abstract Earthquakes do not occur randomly in space and time; rather, they tend to group into clusters that can be classified according to their different properties, presumably related to the specific geophysical properties of a seismic region. Two methods for detection of earthquake clusters are considered in order to take advantage of different descriptions of the seismic process and assess consistency with the obtained clusters: the former is based on "nearest-neighbor distances" between events in space-time-energy domain; the latter is a stochastic method based on a branching point process, named Epidemic-Type Aftershock-Sequence (ETAS) model, which provides different plausible clustering scenarios by simulation. Both methods allow for a robust data-driven identification of seismic clusters, and permit to disclose possible complex features in the internal structure of the identified clusters. We aim at exploring the spatio-temporal features of earthquake clusters in Northeastern Italy, an area recently affected by low-to-moderate magnitude events, despite its high seismic hazard attested by historical destructive earthquakes.

Abstract *I terremoti non avvengono in modo casuale nello spazio-tempo, tendono piuttosto a raggrupparsi in cluster che possono essere classificati secondo le loro diverse proprietà, verosimilmente legate alle specifiche proprietà geofisiche della regione dove essi accadono. Due metodi per l'individuazione di cluster di terremoti sono stati presi in esame al fine di benificiare di descrizioni diverse del processo sismico e valutare la coerenza dei cluster ottenuti: il primo è un metodo basato sulla distanza "nearest-neighbor" tra coppie di eventi nel dominio spazio-tempo-energia; il secondo è un metodo stocastico basato su un modello di processi di punto di tipo branching noto come modello ETAS (Epidemic-Type Aftershock-Sequence)*

Elisa Varini and Renata Rotondi
CNR-Istituto di Matematica Applicata e Tecnologie Informatiche, via Corti 12 - 20133 Milano (I), e-mail: elisa@mi.imati.cnr.it, reni@mi.imati.cnr.it

Antonella Peresan and Stefania Gentili
OGS-Centro di Ricerche Sismiche, Via Treviso 55 - 3100 Udine (I) e-mail: aperesan@inogs.it, sgentili@inogs.it

che, attraverso tecniche di simulazione, è in grado di fornire diversi scenari plausibili di raggruppamento in clusters. Entrambe i metodi consentono una identificazione robusta dei cluster sismici, fondamentalmente guidata dai dati, e permettono di studiare la complessità nella struttura interna dei cluster stessi. Analizziamo le caratteristiche spazio-temporali dei cluster di terremoti avvenuti nell'Italia nord-orientale, un'area ad alta pericolosità sismica, come attestato da alcuni forti terremoti storici, e con una sismicità di magnitudo medio-bassa negli ultimi decenni.

Key words: earthquake clustering, simulation, stochastic declustering, ETAS model, nearest-neighbor distance

1 Some methods for earthquake clustering

Earthquake clustering is a prominent feature of seismic catalogs, both in time and space. Several methodologies for earthquake cluster identification have been proposed in the literature with at least a twofold scope: (1) characterization of the clustering features and their possible relation to physical properties of the crust; (2) declustering of earthquake catalogs which, by removing events temporally and spatially dependent on the mainshocks, allows for spatio-temporal analysis of the background (independent) seismicity. Nevertheless the application of different (de)clustering methods may lead to diverse classifications of earthquakes into main events and secondary events; consequently, the definition of mainshock is not univocal, but strictly related to the different physical/statistical assumptions underlying each method. Therefore we consider different declustering techniques to investigate classification similarities which might provide strong support for some clustering features, and classification differences which might highlight strength and lack of the clustering methods. Two clustering techniques are applied: the nearest-neighbor approach (Zaliapin and Ben-Zion, 2013) and the stochastic declustering approach (Zhuang et al., 2004). Both methods can be satisfactorily applied to decompose the seismic catalog into background seismicity and individual sequences (clusters) of earthquakes; moreover, they are data-driven and allow studying the internal structure of the clusters.

1.1 Nearest-neighbor method (NN)

Bak *et al.* [2] show that the waiting times between earthquakes in California follow a unified scaling law valid at different time scales (from tens of seconds to tens of years), by varying the magnitude threshold M and the linear size L of the studied area. The unified scaling law is obtained by rescaling the distribution $F(t)$ of the waiting times in such a way that t is replaced by $t10^{-bM}L^{d_f}$ and $F(t)$ by $t^{\alpha}F(t)$. Parameters α, b, and d_f correspond to empirical power laws which establish general relations among frequency, waiting times, magnitude, and epicentre locations of the earthquakes: α is the interval exponent of the Omori-Utsu law, b is the b-value of the Gutenberg-Richter distribution of the magnitude, and d_f is the spatial fractal dimension of the epicentre distribution. Thus the unified scaling law is a combination of scale invariant power functions that can be interpreted as statistical evidence of the self-organized critical behaviour of the earth dynamics, which results into a hierarchical organization of earthquakes in time, space, and magnitude. Based on this concept [1], the nearest-neighbor distance between two events in the space-time-energy domain is defined by:

$$\eta_{ij} = t_{ij}(r_{ij})^{d_f} 10^{-bM_i} \ , \tag{1}$$

where t_{ij} denotes the inter-occurrence time between events i and j ($i < j$), r_{ij} is the spatial distance between their epicentres, M_i is the magnitude of i-th event. Event i^* is the nearest-neighbor of event j if $i^* = argmin_{\{i:i<j\}} \eta_{ij}$; in other words, event j is an offspring of event i^*, and also i^* is the parent of j.

By connecting each event with its nearest-neighbor, one obtains a time-oriented tree where each event has a unique parent and may have multiple offspring. By setting a threshold distance η_0, a link between events i and j is removed if $\eta_{ij} > \eta_0$, for all i and j such that $i < j$. The removal of weak links leads to the identification of clusters of events; in each cluster, we define the largest magnitude event as mainshock, the events preceding the mainshock as foreshocks, and the events following the mainshock as aftershocks.

According to Zaliapin and Ben-Zion [6], the histogram of the distances between every pair of events clearly shows a bimodal distribution that can be approximated as a mixture of two Gaussian distributions, one associated with the Poissonian background activity (independent events) and the other with the clustered populations. Thus, threshold distance η_0 can be chosen as equal to the intersection point of the two estimated Gaussian distributions.

Parameters α, b, and d_f are estimated by the Unified Scaling Law for Earthquakes (USLE) method [3, 5].

1.2 Stochastic declustering method (SD)

The stochastic declustering approach is based on the space-time epidemic-type aftershock sequence (ETAS) model [4], a branching point process controlled by its intensity function $\lambda^*(t,x,y,M \mid \mathscr{H}_t)$ conditional on the observation history $\mathscr{H}_t$ up to time t.

Let (t_j, x_j, y_j, M_j) denote occurrence time, epicentral coordinates and magnitude of j-th event. Under the assumption of stationarity, ergodicity, and independence of the magnitude, the general expression of the conditional intensity function of ETAS model is given by:

$$\lambda^*(t,x,y,M \mid \mathscr{H}_t) = J(M)\lambda(t,x,y \mid \mathscr{H}_t) =$$
$$= J(M)\left[\mu(x,y) + \sum_{\{k:t_k<t\}} k(M_k)g(t-t_k)f(x-x_k,y-y_k \mid M_k)\right] ,$$

which, in the formulation of this study, decomposes into:

$$J(M) = be^{-b(M-M_c)} , \tag{2}$$
$$\mu(x,y) = \nu u(x,y) , \tag{3}$$
$$k(M) = Ae^{-\alpha(M-M_c)} , \tag{4}$$
$$g(t) = \begin{cases} (p-1)c^{p-1}(t+c)^{-p} & \text{for } t>0 \\ 0 & \text{otherwise} \end{cases} , \tag{5}$$
$$f(x,y \mid M) = \frac{(q-1)D^{2(q-1)}e^{\gamma(q-1)(M-M_0)}}{\pi[x^2+y^2+D^2e^{\gamma(M-M_0)}]^q} . \tag{6}$$

where M_c is the magnitude threshold of the catalog; Eq.(2) is the distribution of earthquake magnitude; the background rate in Eq.(3) is assumed to be constant in time; the expected number of events triggered from an event of magnitude M is expressed by Eq.(4); the probability density function of the occurrence times and the location distribution of the triggered events are given by Eqs.(5-6), respectively. It is also worth defining the total spatial intensity $m(x,y) = \lim_{T\to\infty} \int_0^T \lambda(t,x,y \mid \mathscr{H}_t)dt/T$ and the clustering spatial intensity $\gamma(x,y) = m(x,y) - \mu(x,y)$. The model parameters are denoted by ν, A, c, α, p, D, q, and γ.

In ETAS model, background earthquakes independently occur at a Poisson rate constant in time, triggering other events with a spatio-temporal decay modelled by the Omori-Utsu law; triggered events have, in turn, the ability to trigger other events. The following expressions respectively provide the probability that event j is triggered by previous event i, the probability that event j is triggered in general, and the probability that event j is generated by the background process:

$$\rho_{ij} = \frac{k(M_i)g(t_j-t_i)f(x_j-x_i,y_j-y_i \mid M_i)}{\lambda(t_j,x_j,y_j \mid \mathscr{H}_{t_j})} \tag{7}$$
$$\rho_j = \sum_{i:t_i<t_j} \rho_{ij} \tag{8}$$
$$\varphi_j = 1-\rho_j \tag{9}$$

By simulating according to these probabilities, the dataset splits into two subsets which are realizations of the background process and the triggered process, respectively: the former is the declustered catalog and the latter identifies a set of clusters each starting from a background event.

The estimation of ETAS parameters is performed by an iterative algorithm that simultaneously estimates the background rate by a variable kernel method and the model parameters by the maximum likelihood method; then the branching structure is obtained by simulation [7, 8, 9].

2 Case study: Northeastern Italy seismicity

The comparative analysis of earthquake clusters is carried out for a sequence of earthquakes occurred in Northeastern Italy. In this area only low-to-moderate magnitude events have been recorded during the last decades, despite its high seismic hazard attested by at least eight historical destructive earthquakes occurred since 1348, the most recent one being the 1976 May 6 M6.4 earthquake, located in the Julian Prealps. A further aim of the clustering analysis is to provide a quantitative basis to understand the role of moderate size earthquakes in the framework of regional seismicity.

2.1 Data

The Italian National Institute of Oceanography and Experimental Geophysics (OGS) started with the monitoring of seismic activities in the Northeastern area of Italy since 1977. We consider the set of earthquakes reported in the OGS bulletins, occurred from 1977 to 2015, with local magnitude at least $M_c = 2$, and ranging from 11.5°E to 14.0°E in longitude and from 45.5°N to 47.0°N in latitude (Fig. 1).
The dataset is considered as complete except for the early 1990s, when data are missing due to a fire accident [5].

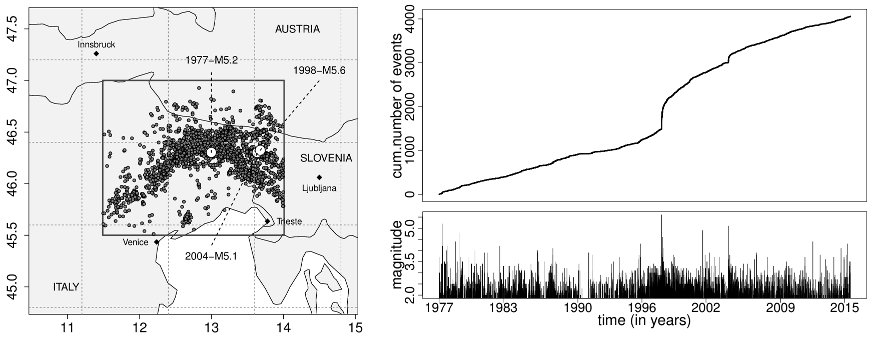


Fig. 1 Epicentres of the earthquakes in the study area: black point if magnitude $M \leq 5$, white points otherwise [left]. Cumulative number of earthquakes versus time (top) and magnitude versus time (bottom) [right].

2.2 Results

NN method univocally splits the dataset into two subsets, a set of background events and a set of triggered events; more details are given in [5]. A tree representation of the identified clusters is provided in order to better highlight the complexity of clusters structure. For example Fig. 2(a) shows the tree of the cluster related to the 1998 earthquake, with magnitude $M5.6$.

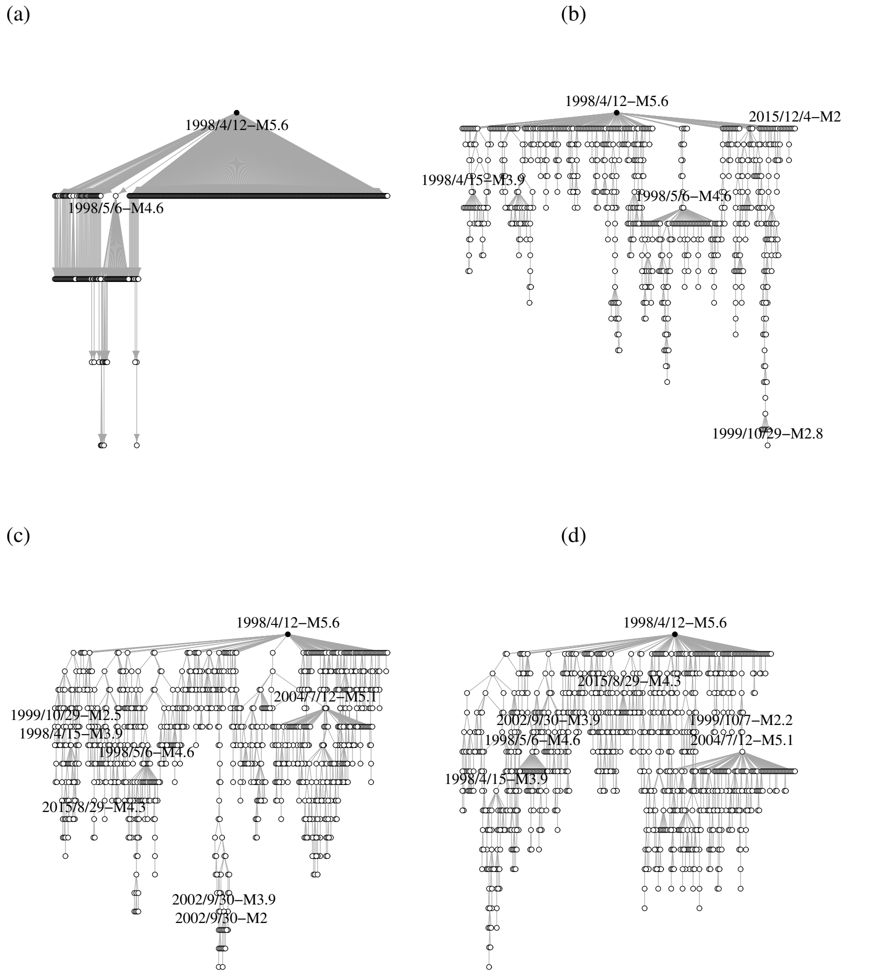


Fig. 2 Cluster trees of the 1998/4/12 earthquake, magnitude M5.6: (a) NN method, (b) SD method, the most probable scenario, and (c-d) SD method, two simulated scenarios, respectively.

As for the SD method, each event j in the dataset is associated with the estimated probability $\hat{\varphi}_j$ ($\hat{\rho}_j$) to be a background (triggered) event. These probabilities provide different clustering scenarios: Fig. 2(b) shows the cluster tree of the 1998 earthquake where each $j-$th event is associated with the most probable ancestor i^* ($i^* = argmax_{i<j}\rho_{ij}$), while Fig. 2(c-d) show the cluster trees of the same earthquake, each derived by simulation according to the estimated background probabilities. For all the spatial coordinates (x,y) in the study area, with reference to SD method, Fig. 3 shows: (a) the estimated background rates $\hat{\mu}(x,y)$, (b) the estimated clustering rate $\hat{\gamma}(x,y)$, (c) the ratio between between estimated clustering rate $\hat{\gamma}(x,y)$ and estimated total rate $\hat{m}(x,y)$, which can be regarded as a smoothed approximation of the triggering probabilities ρ_j. Fig. 3(d) shows the histogram of the estimated background probabilities $\hat{\varphi}_j$, from which the triggering probabilities are given by $\hat{\rho}_j = 1 - \hat{\varphi}_j$: most of the events have high probability of being either background events (21% for $\hat{\varphi}_j \geq 0.9$) or triggered events (42% for $\hat{\varphi}_j \leq 0.1$), and the remaining events have less decisive probabilities (about 37% of the data have background probability ranging from 0.1 to 0.9).

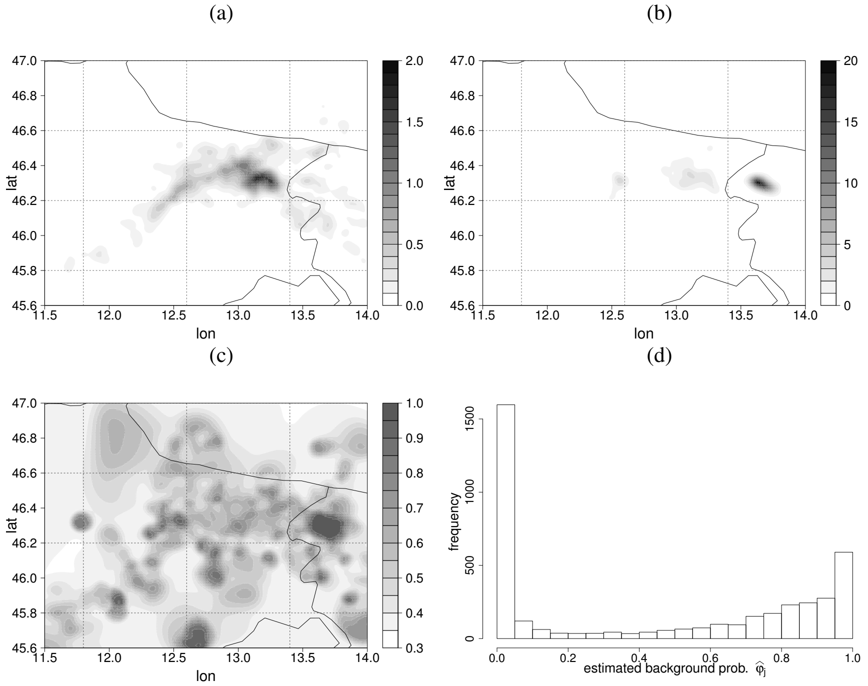


Fig. 3 Estimated background rates (a), estimated clustering rates (b), ratio between estimated clustering rate and estimated total rate (c), and histogram of the estimated background probabilities (d).

A preliminary comparison of results from the two methods shows that the cluster structures produced by NN and SD approaches have comparable trend in terms of spatial extent of seismic clusters. But SD method tends to find some connections between events close in space even if far in time. With reference to 1998 earthquake (Fig. 2), its NN-cluster includes only 480 earthquakes while SD-clusters have more than 900 earthquakes, on average; this is due to the presence in the SD-clusters of earthquakes occurred years later (e.g. in 2002, 2004, or even 2015).

For large sequences, trees obtained from the SD method show a more complex internal structure than trees obtained by the NN method. To quantify topological differences among trees, the average node depth and average leaf depth are considered. The former is the average number of links that connects each node of the tree root; the latter is similarly defined as the average number of links that connects each leaf (node without descendant) to the tree root. According to these scalar measures of internal complexity of a tree, average node depth and average leaf depth are, respectively, 1.30 and 1.31 for the NN-cluster in Fig. 2(a), 5.59 and 5.59 for the SD-cluster in Fig. 2(b), 5.90 and 6.14 for the SD-cluster in Fig. 2(c), 6.20 and 6.43 for the SD-cluster in Fig. 2(d). Greater complexity of the clusters identified by SD method reflects the multilevel triggering property of the ETAS model; we recall that ETAS model assumes that each event is able to generate offspring.

References

1. Baiesi, M., Paczuski, M.: Scale-free networks of earthquakes and aftershocks. Phys. Rev. E **69**, 066106, 1–8 (2004)
2. Bak, P., Christensen, K., Danon, L., Scanlon, T.: Unified Scaling Law for earthquakes. Phys. Rev. Lett. **88**, 178501, 1–4 (2002)
3. Nekrasova, A., Kossobokov, V., Peresan, A., Aoudia, A., Panza, F.: A multiscale application of the Unified Scaling Law for earthquakes in the central Mediterranean area and Alpine region. Pure Appl. Geophys. **168**, 297–327 (2011)
4. Ogata, Y.: Space-time point-process models for earthquake occurrences. Ann. Inst. Stat. Math. **50**, 2, 379–402 (1998)
5. Peresan, A., Gentili, S.: Seismic clusters analysis in Northeastern Italy by the nearest-neighbor approach. Phys. of the Earth and Plan. Int. **274**, 87–104 (2018)
6. Zaliapin, I., Ben-Zion, Y.: Earthquake clusters in southern California I: Identification and stability. J. Geophys. Res. **118**, 6, 2847–2864 (2013)
7. Zhuang, J.: Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. J. J R Stat Soc Series B Stat Methodol. **68**, 4, 635–653 (2006)
8. Zhuang, J., Ogata, Y., Vere-Jones, D.: Analyzing earthquake clustering features by using stochastic reconstruction. J. Geophys. Res. **109**, B5, B05301 (2004)
9. Zhuang, J., Ogata, Y., Vere-Jones, D.: Stochastic declustering of space-time earthquake occurrences. J. Am. Stat. Assoc. **97**, 369–380 (2002)

Advanced spatio-temporal point processes for the Sicily seismicity analysis

Processi puntuali spatio-temporali avanzati per l'analisi della sismicità in Sicilia

Marianna Siino and Giada Adelfio

Abstract Due to the complexity of the generator process of seismic events, we study under several aspects the interaction structure between earthquake events using recently developed spatio-temporal statistical techniques and models. Using these advanced statistical tools, we aim to characterise the global and local scale cluster behaviour of the Easter Sicily seismicity considering the catalogue data since 2006, when the Italian National Seismic Network was upgraded and earthquake location was sensibly improved. Firstly, we characterise the global complex spatio-temporal interaction structure with the space-time ETAS model where background seismicity is estimated non-parametrically, while triggered seismicity is estimated by MLE. After identifying seismic sequences by a clustering technique, we characterise their spatial and spatio-temporal interaction structures using other advanced point process models. For the characterisation of the spatial interactions, a version of hybrid of Gibbs point process models is proposed as method to describe the multiscale interaction structure of several seismic sequences accounting for both the attractive and repulsive nature of data. Furthermore, we consider log-Gaussian Cox processes (LGCP), that are relatively tractable class of empirical models for describing spatio-temporal correlated phenomena. Several parametric formulation of spatio-temporal LGCP are estimated, by the minimum contrast procedure, assuming both separable and non-separable parametric specification of the correlation function of the underlying Gaussian Random Field.
Abstract *Il processo generatore dei fenomeni sismici è caratterizzato da una certa complessità. In questo lavoro studieremo sotto diversi punti di vista la struttura di interazione tra i terremonti utilizzando tecniche e modelli statistici di tipo spazio temporale. In particolare, le analisi si pongono come obiettivo quello di caratteriz-*

Marianna Siino
Università degli studi di Palermo, Dipartimento di Scienza Economiche, aziendali e statistiche e-mail: marianna.siino01@unipa.it

Giada Adelfio
Università degli studi di Palermo, Dipartimento di scienza economiche, aziendali e statistiche e-mail: giada.adelfio@unipa.it

zare il comportamento a piccola e a grande scale degli eventi sismici avvenuti a partire dal 2006 nella Sicilia Orientale. In prima istanza, si caratterizza l'andamento globale della sismicità nell'area di studio utilizzando il modello spazio-temporale ETAS, in cui la sismicità di fondo viene stimata in modo non parametrico mentre la sismicità indotta viene stimata mediante massima verosimiglianza. Dopo aver identificato sequenze sismiche, con un opportuno metodo di clustering, per ciascuna sequenza si studia la struttura di relazione che intercorre tra gli eventi dal punto di vista sia spaziale che spazio-temporale. Per quanto riguarda lo studio delle interazioni spaziali, una versione dei modelli ibridi Gibbs viene proposta per descrivere la struttura di interazione multiscala delle sequenze tenendo conto allo stesso tempo del comportamento attrattivo e repulsivo che intercorre tra i punti. Inoltre, abbiamo considerato la classe dei modelli di tipo log-Gaussian Cox, utili per la descrizione di fenomeni correlati nello spazio e nel tempo. Utilizzando il metodo dei momenti, sono stati stimati e in seguito confrontati diversi modelli parametrici assumendo una struttura di correlazione del processo Gaussiano sottostante sia separabile che non-separabile.

Key words: earthquakes; hybrid of Gibbs process; log-Gaussian Cox processes; minimum contrast method; non-separable covariance function; point process; spatio-temporal pair correlation function

1 Goals of the analysis and description of the study area

To describe and predict seismic events in space and time, a proper characterisation of both the intensity function and the second-order properties of the generator process is a crucial issue. In this work, we give a general overview of advanced statistical models that can be used in the seismic context, showing the structure of the models, the main diagnostic tools and the available code (Siino et al, 2018a,b). In an application example, we aim to characterise under several aspects the interaction structure observed in the Sicily catalogue events using proper point process models.

First, we describe the global complex interaction structure using the space-time Epidemic Type Aftershock Sequence model (Ogata, 1988). Focusing on a local scale, we identify some sequences of events characterising their spatial and spatio-temporal structure. In particular, hybrid of Gibbs point processes (Baddeley et al, 2015) are used to describe the distribution of epicentres that generally exhibit interaction at different spatial scales. On the other hand, the spatio-temporal cluster structure is characterised by spatio-temporal log-Gaussian Cox models (Diggle et al, 2013), where the main issue concerns the specification of the moments of the underlying Gaussian Random Field (GRF). Moreover, in the remain of the paragraph we give a general overview of the study area. It is focused on eastern Sicily: it extends from 36.5° to 39° Lat. N and from 14° to 16° Long. E. East Sicily and South Calabria is the area with greater deformation rate and seismic strain release in Italy. We consider the seismic catalogue since 2006, when the Italian National Seismic

Network was upgraded and earthquake location was sensibly improved. The instrumental seismicity recorded in the period from 2006 to 2016, consists of 12356 events; 4170 events (33.7 %) have M between 2.1 and 3.0, 333 events (2.7 %) have M between 3.1 and 4.0, and just 28 events (0.2 %) have $M > 4$. The most of the events (75.4%) are crustal, with hypocentral depth lower than 30 km (Figure 1).

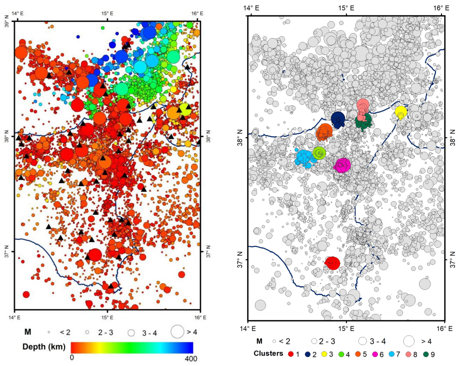


Fig. 1: (left) Seismicity map from 2006 to 2016 period, black triangles indicate the seismic stations. (right) Distribution of the detected sequences.

2 Global scale analysis: Epidemic Type Aftershocks-Sequences model

Let $\mathscr{X}$ be a random countable subset of $\mathbb{R}^2 \times \mathbb{R}^+$ where for a point $(\mathbf{u},t) \in \mathscr{X}$, $\mathbf{u} \in \mathbb{R}^2$ is the location and $t \in \mathbb{R}^+$ is the time of occurrence. In practice, an observed spatio-temporal pattern is a finite set $\{(\mathbf{u}_i,t_i)\}_{i=1}^n$ of distinct points within a bounded spatio-temporal region $W \times T \subset \mathbb{R}^2 \times \mathbb{R}^+$, where usually W is a polygon with area $|W| > 0$ and T a single closed interval with length $|T| > 0$. The space-time point process is defined as a random point pattern, completely characterized by its conditional intensity function (CIF). The CIF of the ETAS model (Ogata, 1988), in time t and location $\mathbf{u}$ conditional to $\mathscr{H}_t$ (the history until t) is defined as the sum of a term describing the long-term variation and one relative to the short-term variation,

$$\lambda_{\boldsymbol{\theta}}(t,\mathbf{u}|\mathscr{H}_t) = \mu f(\mathbf{u}) + \sum_{t_j<t} \frac{\kappa_0 \, e^{(\alpha)\ (m_j-m_0)}}{(t-t_j+c)^p} \left\{||\mathbf{u}-\mathbf{u}_j||^2+d\right\}^{-q} \tag{1}$$

The background activity is characterized by μ, that indicates the general background intensity, and $f(\mathbf{u})$ that is the space density. Instead, for the induced intensity, κ_0 is a constant which measures the aftershocks productivity, c and p are parameters of the modified Omori's law (p characterises the pattern seismicity in the given region indicating the decay rate of aftershocks in time), α measures the influence on the relative weight of each sequence, m_j is the magnitude of the inducing event and m_0 is the completeness threshold of magnitude. The parameters are estimated with the Forward Likelihood-based predictive approach implemented in the `R` package `etasFLP` (Chiodi and Adelfio, 2014). Figure 2 shows the estimated spatial background intensity and the induced one with $m_0 = 2.7$. According to the estimated model (Figure 2), there is quicker decay in space than time since $q > p$, and there is a positive effect of the magnitude of the previous events since $\alpha > 0$.

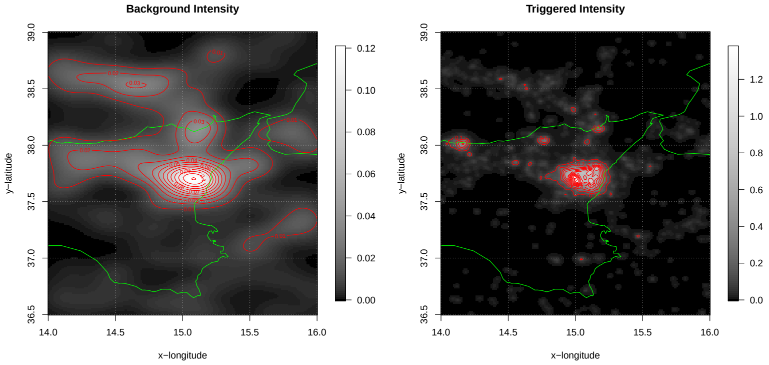


Fig. 2: Estimated background and triggered space intensities, with estimates $\boldsymbol{\theta} = \{0.051, 0.049, 0.005, 1.075, 0.931, 7.510, 1.796\}$ and corresponding standard errors $s.e.(\boldsymbol{\theta}) = \{0.003, 0.017, 0.001, 0.013, 0.113, 1.174, 0.073\}$.

3 Local scale analysis

For the local scale analysis, we report the main results obtained for the cluster 5 of Figure 1 occurred close to the village San Salvatore di Fitalia in 2011. The main-shock event was the 23rd of June 2011 with M=4.5 and 7 km depth. For this sequence, the aims are to characterise the spatial multiscale interaction structure using hybrid models and the spatio-temporal evolution using log-Gaussian Cox models.

3.1 Hybrid of spatial Gibbs process

The class of Gibbs processes $\mathscr{X}$ is determined through a probability density function $f : \mathscr{X} \to [0, \infty)$, where $\mathscr{X} = \{\boldsymbol{\upsilon} \subset W : n(\boldsymbol{\upsilon}) < \infty\}$ is a set of point configurations contained in W. In the literature several Gibbs models have been proposed such as the area-interaction, Strauss, Geyer, hard core processes. However Gibbs processes have some drawbacks when points have a strong clustering and show spatial dependence at multiple scales (Baddeley et al, 2013). Baddeley et al (2013) propose hybrid models as a general way to generate multi-scale processes combining Gibbs processes. Given m unnormalized densities $f_1, f_2, \ldots, f_m$, the hybrid density is defined as $f(\boldsymbol{\upsilon}) = f_1(\boldsymbol{\upsilon}) \times \ldots \times f_m(\boldsymbol{\upsilon})$, respecting some properties. For example, the density of the inhomogeneous hybrid process obtained considering m Geyer components (with interaction ranges $r_1, \ldots r_m$ and saturation parameters $s_1, \ldots, s_m$) is

$$f(\boldsymbol{\upsilon}) = \prod_{i=1}^{n(\boldsymbol{\upsilon})} \beta(\boldsymbol{u}_i) \prod_{j=1}^{m} \gamma_j^{min(s_j, t(\mathbf{u}_i, \boldsymbol{\upsilon} \backslash \mathbf{u}_i; r_j))} \tag{2}$$

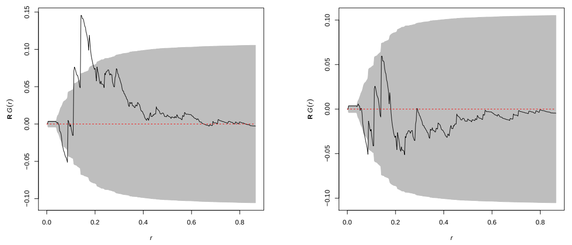
where $t(\mathbf{u}_i, \boldsymbol{\upsilon} \backslash \mathbf{u}_i; r_j) = \sum_i \{\mathbf{1} \| \mathbf{u} - \mathbf{u}_i \| \leq r_j\}$. This density indicates that the spatial interaction between points changes with the distances r_j and the parameters that capture this information are the interaction parameters γ_j. Gibbs models can be fitted to data by pseudolikelihood that is function of the Papangelou conditional intensity $\rho_\phi(\mathbf{u}|\boldsymbol{\upsilon})$ at location $\mathbf{u} \in W$ given $\boldsymbol{\upsilon}$, where ϕ are the parameters to estimate (Baddeley et al, 2015). In hybrid models, the conditional intensity is

$$\rho_\phi(\mathbf{u}, ; \upsilon) = \exp\{B(\mathbf{u}) + \theta_1^T V_1(\mathbf{u}, \eta) + \theta_2^T G(\mathbf{u}, \boldsymbol{\upsilon}, \eta)\}, \tag{3}$$

where $B(\mathbf{u})$ is an offset term, $\theta_1^T V_1(\mathbf{u}, \eta)$ is the first-order potential and $\theta_2^T G(u, \boldsymbol{\upsilon}, \eta)$ accounts for the interaction effects. In the following analysis, it will be a combination of Geyer processes, and in this case the irregular parameter η accounts both for an interaction distance r_j and a saturation parameter s_j for each j-th Geyer component. Usually, to assess the goodness-of fit, the estimated models are compared in terms of AIC, spatial raw residuals and number of simulated points under the estimated model (Baddeley et al, 2015). Furthermore, the diagnostic plots based on the residual K- and G-functions are used to decide which component has to be added at each step to the hybrid model (Baddeley et al, 2015). In the `spatstat` package (Baddeley et al, 2015) of R (R Development Core Team, 2005), there are the most of the functions that have been used for fitting, prediction, simulation and validation of Hybrid models. Given a set of points, the first attempt in model estimation is to fit an inhomogeneous Poisson model with intensity depending on the spatial coordinates in which the points do not interact with each other, that is equal to equation (3) where the term $G(\cdot)$ is null. For the selected sequence, the corresponding AIC and the range of the spatial raw residuals of the fitted inhomogeneous Poisson models with a non-parametric spatial trend are in Table 1. In Figure 3a, the residual G-function for the estimated inhomogeneous Poisson model is reported. For distances up to about 200 meters, it wanders substantially outside the limits showing peaks, so there is a positive association between points unexplained by the Poisson model. Therefore, we consider hybrid of Geyer processes, and the main estimation results are in Table 1. Considering the hybrid model, there is an improvement in terms of AIC, of range of raw residuals, and of residual G-function (Figure 3b), it oscillates around zero and it is inside the envelopes indicating that the multiscale interaction structure between the earthquakes is well described by the hybrid model of two Geyer processes.

Table 1: Inhomogeneous Poisson and the hybrid of Gibbs model: AIC, range of spatial raw residuals, vector of irregular parameter η and interaction parameter for each Geyer component γ_j.

Inhom. Poisson model		Inhom. hybrid of Gibbs processes				
AIC	Range res.	Comp.	η	γ	AIC	Range res.
-961.829	[-0.142 ; 0.531]	G_1	$(r_1;s_1) = (0.060;0.2)$	0.413	-990.565	[-0.258 ; 0.171]
		G_2	$(r_2;s_2) = (0.240;3.5)$	1.157		



(a) (b)

Fig. 3: For the sequence 5, (a) residual G-functions for the inhomogeneous Poisson process model and for the inhomogeneous hybrid model of Geyer processes (b).

3.2 *Spatio-temporal log-Gaussian Cox process*

The class of log-Gaussian Cox processes (LGCPs) (Møller et al, 1998) is a flexible and relatively tractable class of models for describing correlated phenomena specifying the moments of an underlying Gaussian Random Field (GRF). Considering a spatio-temporal point process as in Section 2, $\mathscr{X}$ is said to be a Cox process driven by Λ (a non-negative and locally integrable stochastic process), if the conditional distribution of the point process $\mathscr{X}$ given a realization $\Lambda(\mathbf{u},t) = \lambda(\mathbf{u},t)$ is a Poisson process on $W \times T$ with intensity function $\lambda(\mathbf{u},t)$. Following the homogeneous specification in Diggle et al (2013), the log-Gaussian Cox process for a generic point in space and time has the following intensity $\Lambda(\mathbf{u},t) = \exp\{\beta + S(\mathbf{u},t)\}$ where S is a Gaussian process with $\mathbb{E}(S(\mathbf{u},t)) = \mu = -0.5\sigma^2$ and so $\mathbb{E}(\exp\{S(\mathbf{u},t)\}) = 1$ and with variance-covariance matrix $\mathbb{C}(S(\mathbf{u},t), S(\mathbf{v},s)) = \mathbb{C}(||\mathbf{u}-\mathbf{v}||, |t-s|) = \sigma^2\gamma(r,h)$ under the stationary assumption, where $\gamma(\cdot)$ is the correlation function of the GRF. Following Møller et al (1998), the first-order product density and the pair correlation function of the log-Gaussian Cox process are $\mathbb{E}(\Lambda(\mathbf{u},t)) = \lambda = \exp(\beta)$ and $g(r,h) = \exp(\sigma^2\gamma(r,h))$, respectively.

To describe the spatio-temporal correlation structure of the selected sequence (Figure 1), we consider several formulation of LGCP models changing the specification of the covariance structure of the underline GRF assuming both separable (exponential structure in space and time) and non-separable parametric specifications (Gneiting and Iaco-Cesare families). Table 2 shows the estimated parameters for each model obtained with the minimum contrast method proposed in Siino et al (2018a). To assess the goodness-of-fit of the estimated models, we use the 95% envelopes of the non-parametric spatio-temporal K-function (Figure 4) and a global test based on Monte Carlo procedure to check if the spatio-temporal point pattern is a realisation of a LGCP with the specified parameters. The fitted models assuming a separable structure and the Gneting families seem to do not describe properly the data because the observed spatio-temporal K-function is outside the envelope, see Figures 4a and 4c. This is confirmed by the global p-values in Table 2. A LGCP model with Iaco-Cesare structure for the underlying GRF describes better the spatio-temporal cluster structure according to Figure 4b and the global p-value that is 0.10 (Table 2). According to the estimates for the selected homogeneous LGCP model, there is a decay of the interaction in space and time governed mainly by the scale parameters α and β.

4 Comments

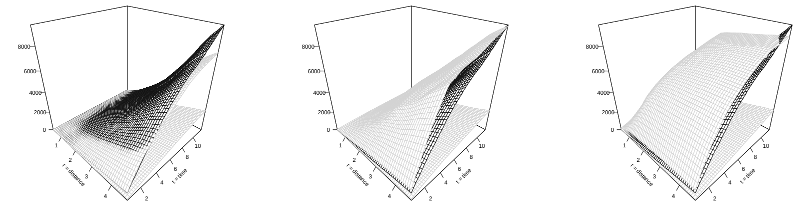
We propose the use of different models for the analysis of seismic data, focussing both on the global and local scales. In the first case the ETAS model is considered, providing a space-time characterization of the seismic activity in the considered area. For the local scale analysis of the sequence, we consider both Hybrid and LGCP models. Their results are not comparable, since the two model formulation focus on different aspects. In particular, with the hybrid model, we focus on a spatial scale and the multiscale interaction structure is properly described by a hybrid of Geyer processes after taking into account the spatial inhomogeneity with a non-parametric kernel. Moreover, there is evidence of a strong clustering structure for short distances up to 200 meters. On the other hand, by using the LGCP model, spatial and temporal evolution are simultaneously described specifying the moments of a GRF, without assuming a deterministic part in the intensity and considering a homogeneous formulation of the model.

References

Baddeley A, Turner R, Mateu J, Bevan A (2013) Hybrids of gibbs point process models and their implementation. Journal of Statistical Software 55(11):1–43

Table 2: Estimates of the LGCP models assuming several covariance parametric families for the GRF. The p-values are shown to assess if the point pattern comes from the assumed spatio-temporal LGCP model.

Cluster	Family	$\hat{\sigma}^2$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}_s$	$\hat{\gamma}_t$	$\hat{\delta}$	p-value
Exp-Exp	$\mathbb{C}(r,h) = \sigma^2 \exp\left(-\frac{r}{\alpha}\right) \exp\left(-\frac{h}{\beta}\right)$	3.075	12.025	14.934				0.01
Iaco-Cesare	$\mathbb{C}(r,h) = \sigma^2 \left(1 + \left(\frac{r}{\alpha}\right)^{\gamma_s} + \left(\frac{h}{\beta}\right)^{\gamma_t}\right)^{-\delta}$	3.237	13.709	15.910	0.983	1.022	1.500	0.10
Gneiting	$\mathbb{C}(r,h) = \frac{\sigma^2}{\left(\left(\frac{h}{\beta}\right)^{\gamma_t}+1\right)^{\delta/\gamma_t}} \exp\left(-\frac{\left(\frac{r}{\alpha}\right)^{\gamma_s}}{\left(\left(\frac{h}{\beta}\right)^{\gamma_t}+1\right)^{\delta/(2\gamma_t)}}\right)$	3.093	2.252	2.955	1.504	1.701	0.354	0.08



(a) Exp-Exp (b) Iaco-Cesare (c) Gneiting

Fig. 4: 95% envelope of the spatio-temporal K-function based on simulated spatio-temporal LGCP patterns according to the estimated parameters in Table 2.

Baddeley A, Rubak E, Turner R (2015) Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press

Chiodi M, Adelfio G (2014) etasflp: Estimation of an etas model. mixed flp (forward likelihood predictive) and ml estimation of non-parametric and parametric components of the etas model for earthquake description. R package version 10

Diggle PJ, Moraga P, Rowlingson B, Taylor BM (2013) Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. Statistical Science 28(4):542–563

Møller J, Syversveen AR, Waagepetersen RP (1998) Log-Gaussian Cox processes. Scandinavian Journal of Statistics 25(3):451–482

Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical Association 83(401):9–27

R Development Core Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-07-0

Siino M, Adelfio G, Mateu J (2018a) Joint second-order parameter estiamation for spatio-temporal log-gaussian cox processes. (submitted)

Siino M, D'Alessandro A, Adelfio G, Scudero S, Chiodi M (2018b) Multiscale processes to describe the eastern sicily seismic sequences. submitted

Spatial analysis of the Italian seismic network and seismicity

Analisi spaziale della rete sismica italiana e della sismicità

Antonino D'Alessandro, Marianna Siino, Luca Greco and Giada Adelfio

Abstract Seismic networks are powerful tools for understanding active tectonic processes in a monitored region. Their numerous applications, from monitoring seismicity to characterizing seismogenic volumes and generated seismicity, make seismic networks essential tools for assessing seismic hazard in active regions. The ability to locate earthquakes hypocenters requires a seismic network with a sufficient number of optimally distributed, stations. It is important to assess existing network geometry, to identify seismogenic volumes that are not adequately monitored, and to quantify measures that will allow network improvement. In this work we have studied the spatial arrangement of the stations of the Italian National Seismic Network by means of several Point Pattern techniques The results of the point patter analysis were compare with the spatial distribution of the historical and instrument seismicity and with the distribution of the well know seismogenetic sources of the Italian peninsula. Some considerations have also been made on some models of seismic hazard of the Italian territory. Our analysis allowed us to identify some critical areas that could require an optimization of the monitoring network.

Abstract *Le reti sismiche permettono di misurare e comprendere i processi tettonici attivi in una regione monitorata. Le informazioni acquisite mediante le reti vengono utilizzate per il monitoraggio della sismicità, per la caratterizzazione dei volumi sismogenetici e della sismicità generata, di fatto sono uno strumento essenziale per la valutazione del rischio sismico. Per localizzare gli ipocentri bisogna disporre di una rete sismica con un numero sufficiente di stazioni distribuite in modo ottimale. È importante valutare la geometria della rete esistente, identificare i volumi sismogenetici che non sono adeguatamente monitorati e quantificare le misure che consentiranno il miglioramento della rete. In questo lavoro abbiamo studiato la*

Antonino D'Alessandro and Luca Greco
Istituto Nazionale di Geofisica e Vulcanologia e-mail: antonino.dalessandro@ingv.it, luca.greco@ingv.it

Marianna Siino and Giada Adelfio
Università degli studi di Palermo, Dipartimento di Scienza Economiche, Aziendali e Statistiche e-mail: marianna.siino01@unipa.it, giada.adelfio@unipa.it

disposizione spaziale delle stazioni della Rete Sismica Nazionale Italiana attraverso diverse tecniche per lo studio dei processi puntuali. I risultati dell’analisi spaziali delle stazioni sono stati confrontati con la distribuzione spaziale della sismicità storica e strumentale e con la distribuzione delle ben note sorgenti sismiche della penisola italiana. Alcune considerazioni sono state fatte anche su alcuni modelli di pericolosità sismica del territorio italiano. La nostra analisi ci ha permesso di identificare alcune aree critiche che potrebbero richiedere un’ottimizzazione della rete di monitoraggio.

Key words: earthquakes; point process; seismic network; spatial correlation

1 Introduction

The Italian seismic network was developed immediately after the Irpinia seismic crisis in 1980. After this disastrous event, the National Seismic Network (NSN) was established. Initially, the NSN consisted of only few stations spread over Italy. In the 90s, the ING (Istituto Nazionale di Geofisica), then became INGV (Vulcanlogia), upgraded progressively the monitoring network leading, in about thirty years, to the current NSN consisting of about 500 seismic stations. Over the years, the spatial and temporal network development continued, depending on the funds availability for the purchase of the instrumentation and the implementation of monitoring nodes. As it is well known, the quality of the estimate of focal parameters depends on the density and geometry of the monitoring stations. Therefore, the development of a seismic network should be carried out according to precise criteria, designed to guarantee a rational development of the monitoring infrastructure. The most correct criterion is to adapt the seismic network, that is to increase the density of monitoring stations, in the areas with the greatest number of seismogenic structures, historical strong earthquakes and with the greater release of seismic energy observed as instrumental seismicity. Unfortunately, the development of the NSN did not follow a precise criterion, uniformly applied to the whole territory. The result is that the quality of the location and the magnitude of completeness is very inhomogeneous and sometimes does not seem to be rational or proportional to the seismic rate (Schorlemmer et al, 2010; D’Alessandro et al, 2011). This clearly can have repercussions on the quality of the seismic monitoring and studies and it is therefore necessary to identify a simple and effective criterion that could be used for the future NSN optimization. In the following, we propose a statistical approach based on the spatial distribution of the NSN, instrumental and historical seismicity, to evaluate the current degree of the network coverage and plan for future optimization.

2 Data description

We integrate seismological, geological and seismic network data in order to address the goals of the analysis, and in this paragraph they are briefly described. We restrict the study window (W) to the Italian peninsula and Sicilian land boundary because only few stations are placed in offshore areas and Sardinia island is not characterised by a high seismicity. As a matter of fact, at this point of the analysis, we want to relate the spatial distribution of the stations with the seismicity and geological information available on the mainland.

More than 500 stations of the National Seismic Network managed by the National Institute of Geophysics and Volcanology and other networks managed by other bodies are used to locate earthquakes, sending data to the central branch in Rome in real time to monitor the seismicity in Italy. The seismic network is composed by 363 stations (Figure 1a).

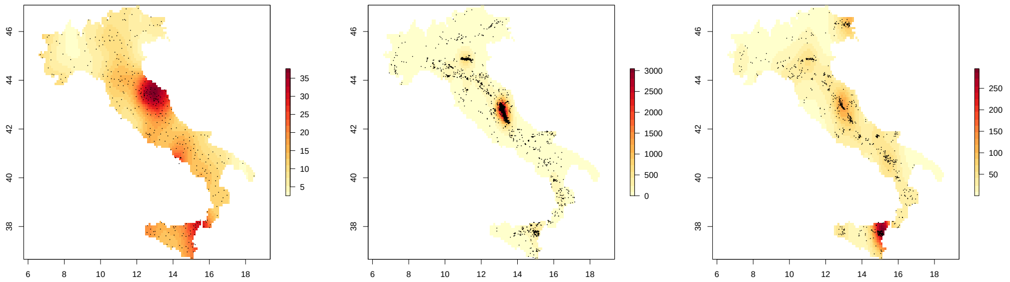
The Italian catalogue contains events since 1985 and we considered the earthquake since 2005, when the network was upgraded and earthquake location was sensibly improved. A subset of the catalogue consisting by 2936 events is analysed, selecting the earthquakes in W with a threshold magnitude equal to 3 and a focal depth less than 50 km (Figure 1b). Furthermore, the historical seismicity is selected starting from the Parametric Catalogue of the Italian Earthquakes, (Rovida et al, 2016) containing 4584 events in the time window 1000-2014; in particular, we select a subset of the events with location, magnitude and depth corresponding to the ones of the catalogue data (Figure 1c).

Additionally, we consider geological information to understand dependence between stations and the sources of earthquakes. The dataset of the Composite Seismogenic Sources (CSS) (Group et al, 2010) is plotted in Figure 2. A composite source represents a complex fault system with an unspecified number of aligned individual seismogenic sources that cannot be separated spatially. In particular for the analysis, we consider the upper edges of the composite sources that for the sake of simplicity are named faults.

The seismic stations, the catalogue events (instrumental seismicity) and the historical seismicity are treated as three spatial point patterns in the region W. Then, we use R (R Development Core Team, 2005) packages for the statistical analysis of spatial patterns of points in two-dimensional space, *spatstat* (Baddeley and Turner, 2005) and *spatstat.local* (Baddeley, 2018).

3 Main results

More formally, a spatial point pattern $\boldsymbol{\upsilon} = \{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is an unordered set of points in the region $W \subset \mathbb{R}^d$ where $n(\boldsymbol{\upsilon}) = n$ is the number of points, $|W| < \infty$ and $d = 2$. The first-order intensity is assumed inhomogeneous ($\lambda(\mathbf{u})$) and for our point patterns is estimated non-parametrically to understand the spatial trend (Figure 1). The usual kernel estimator of the intensity function is (Baddeley et al, 2015)



(a) Stations (b) Earthquakes (c) historical seismicity

Fig. 1: Kernel estimate of intensity, with smoothing bandwidth selected by Scott’s rule for: (a) the seismic monitoring stations, (b) earthquakes occurred between 2005 and 2018 and (c) the historical seismicity. In (a) and (b), the selected events have a magnitude greater than or equal to 3 and a focal depth less than 50 km. Black points are the events.

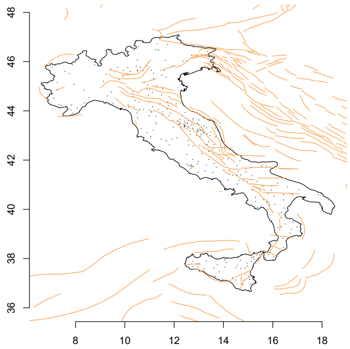


Fig. 2: Faults, upper edges of the composite sources

$$\hat{\lambda}(\mathbf{u}) = 1/e(\mathbf{u}) \sum_{i=1}^{n} k(\mathbf{u} - \mathbf{u}_i, h)$$

where $e(\mathbf{u})$ is the edge corrections and $k(\cdot)$ is a Gaussian density with standard deviation (smoothing bandwidth) equal to h.

From Figure 1a, there is a high concentration of stations in the centre of Italy and in the neighbourhood of the Vesuvio and Etna volcanoes. As for the spatial distribution of instrumental seismicity (Figure 1b), the areas with the higher number of events are in centre of Italy (referring to the Aquila sequence in 2009 and Amatrice-

Norcia-Visso sequence between 2016 and 2018) and in the north centre of Italy for the Emilia sequence in 2012. The historical seismicity (Figure 1c) indicates that most of the seismic activity in Italy is along the alpine and Apennines areas and in Sicily

In this paper, we want to study the relationship between the spatial distribution of the stations with respect to the two types of seismicity (instrumental and historical) under two aspects, their spatial dependence and the global and local characterization. With respect to the first aspect, we compute the inhomogeneous version of the bivariate K-function, to assess if the pair of point patterns are spatially dependent. As it concerns the second aspect, the global and local correlation coefficients are computed to compare the estimated intensities in Figure 1.

Generally for any pair of types i and j, the multitype K-function $K_{ij}(r)$ is the expected number of points of type j lying within a distance r of a typical point of type, i, dividing by the intensity of points of type j. It takes the form

$$K_{ij}(r) = 1/\lambda_j E\left[t(u, r, \mathbf{X}^{(j)})|\mathbf{u} \in \mathbf{X}^{(i)}\right]$$

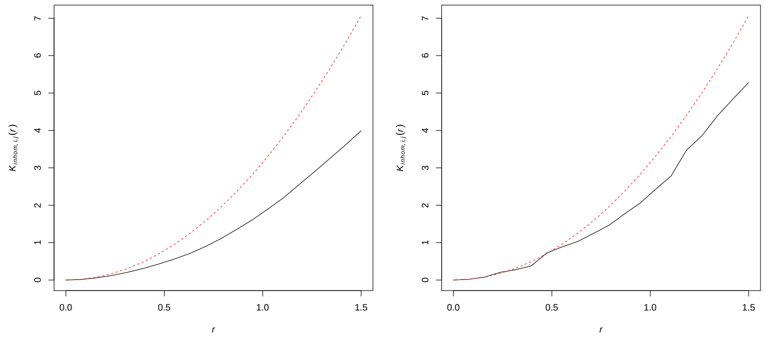
where $\mathbf{X}^{(j)}$ is the sub-process of points of type j, with intensity λ_j, and $t(\cdot)$ is the the number of points in the point pattern $\mathbf{X}^{(j)}$ that lie within a distance r of the location **u**, but not at **u** itself. If the process of type i points are independent of the process of type j points, then $K_{ij}(r)$ is equal to ϕr^2. Deviations between the empirical value of the curve and the theoretical one suggest dependence between the points of types i and j.

For instance, Figure 3a shows the K-cross where the point pattern of stations is type i and the point patterns of earthquake events from 2005 to 2018 is type j. Clearly the two point patterns are dependent and since the observed cross K-function is below the theoretical value, type j events are farther to type i events than it would be expected under complete spatial randomness. Similarly conclusions can be drawn when the K-cross is computed between the point pattern of the stations (type i) and the point pattern of historical seismicity (type j), see Figure 3a. However, for distances up to 0.5 degree, the two distributions (the station and the historical seismic events) seem to be independent.

As expected, the overall correlation between the intensity of stations with respect to the instrumental seismicity and historical seismicity is positive, the Pearson correlation coefficients are 0.423 and 0.643, respectively.

Moreover, considering the estimated intensities as raster data, we want to check if there are variations of correlation in space. Around each cell of the rasters, we define a focal squared area 5x5 and the correlation between the 25 values of each raster in this square is recorded for the central cell, see Figure 4. Comparing the two plots, it seems that there are more areas with negative correlation between the stations and instrumental seismicity (Figure 4a) than the stations and the historical seismicity (Figure 4b). It would suggest the necessity of a development of the current network, not completely concordant with the spatial seismicity evolution.

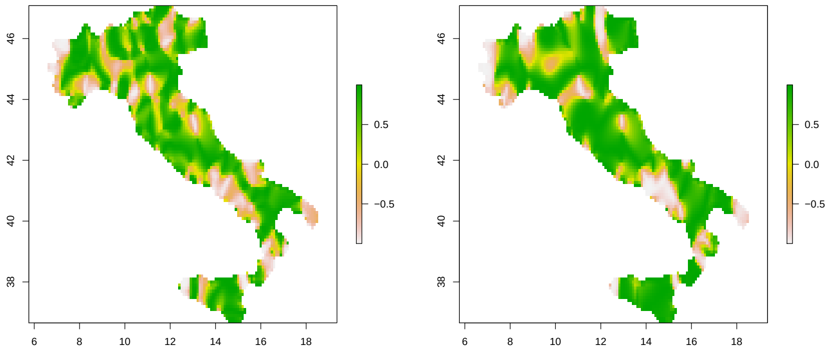
Finally, the main interest lies in deciding whether the spatial arrangement of the stations occurs more frequently near faults. The geological information in Figure 2



(a)

(b)

Fig. 3: (a) The black line is the inhomogeneous bivariate K-function for the seismic station point pattern (type i) and the earthquake events from 2005 and 2018 (type j). (b) The black line is the bivariate inhomogeneous K-function for the seismic station point pattern (type i) and the historical seismicity (type j). In both the plots, the red line corresponds to the theoretical value πr^2.



(a)

(b)

Fig. 4: (a) Local correlation coefficient for the raster objects, between the stations and the instrumental seismicity. (b) Local correlation coefficient for the raster objects, between the stations and the historical seismicity. In both cases, the focal squared area has a dimension 5x5.

is transformed into a spatial variables defined at all locations $\mathbf{u} \in W$, namely $D(\mathbf{u})$ distance to the nearest fault. Assuming an inhomogeneous Poisson model with a parametric log-linear form with respect to the covariate D, we estimate the following intensity

$$\lambda(\mathbf{u}) = exp(\beta_0 + \beta_1 D(\mathbf{u}))$$

where the estimates are $\hat{\beta}_0 = 2.84$ and $\hat{\beta}_1 = -1.69$.

Moreover, we obtain with local inference (Baddeley, 2017) spatially-varying estimates of the parameters of the previous inhomogeneous Poisson process model,

$$\lambda(\mathbf{u}) = exp(\beta_0(\mathbf{u}) + \beta_1(\mathbf{u})D(\mathbf{u}))$$

This approach has the potential to detect and model gradual spatial variation of the parameters that govern the intensity of the stations and the estimates are in Figure 5.

Generally increasing the distance to the nearest fault the station intensity decreases, however according to the spatially varying slope coefficient this reduction is higher in the centre and north-east of Italy.

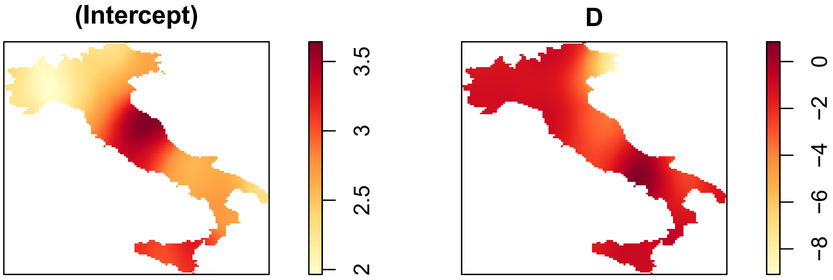


Fig. 5: Spatially varying estimates of intercept (Left) and slope coefficient (Right) from the local likelihood fit of the log-linear model to the seismic station data where the covariate ($D(\mathbf{u})$) is distance to the nearest fault.

4 Conclusions

In this paper we use statistical methods and tools ior the description and characterization of the current degree of the network coverage based on the spatial distribution of the NSN and of the instrumental and historical seismicity, in order to suggest directions for planning future optimization.

As observed from Figure 4, a further upgrade of the network allocation is necessary, in order to get homogeneous and positive correlation in all the Italian area.

In particular, for instance, along the Apennines as well as the West Sicily area, the different correlation between the NSN and instrumental and historical seismicity respectively, suggests the necessity of a network strengthening in those areas.

References

Baddeley A (2017) Local composite likelihood for spatial point processes. Spatial Statistics 22:261–295

Baddeley A (2018) spatstat.local: Extension to ’spatstat’ for Local Composite Likelihood. URL https://CRAN.R-project.org/package=spatstat.local, r package version 3.5-7

Baddeley A, Turner R (2005) Spatstat: An r package for analyzing spatial point patterns. Journal of Statistical Software 12(i06)

Baddeley A, Rubak E, Turner R (2015) Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press

D’Alessandro A, Luzio D, D’Anna G, Mangano G (2011) Seismic network evaluation through simulation: An application to the italian national seismic network. Bulletin of the Seismological Society of America 101(3):1213–1223

Group DW, et al (2010) Database of individual seismogenic sources (diss). In: Version 3.1. 1: A Compilation of Potential Sources for Earthquakes Larger than M 5.5 in Italy and Surrounding Areas, © INGV 2010—Istituto Nazionale di Geofisica e Vulcanologia

R Development Core Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-07-0

Rovida AN, Locati M, Camassi RD, Lolli B, Gasperini P (2016) Cpti15, the 2015 version of the parametric catalogue of italian earthquakes. Istituto Nazionale di Geofisica e Vulcanologia doi:http://doiorg/106092/INGVIT-CPTI15

Schorlemmer D, Mele F, Marzocchi W (2010) A completeness analysis of the national seismic network of italy. J Geophys Res 115:B04,308

Dimensional Reduction Techniques for Big Data Analysis

Clustering Data Streams via Functional Data Analysis: a Comparison between Hierarchical Clustering and K-means Approaches

(*Classificazione di Data Stream con l'Analisi dei Dati Funzionali: un Confronto tra Cluster Gerarchica e Metodo delle K-medie*)

Fabrizio Maturo, Francesca Fortuna, and Tonio Di Battista

Abstract Recently, the analysis of web data, has become essential in many research fields. For example, for a large number of companies, corporate strategies should be based on the analysis of customer behaviour in surfing the world wide web. The main issues in analysing web traffic and web data are that they often flow continuously from a source and are potentially unbounded in size, and these circumstances inhibit to store the whole dataset. In this paper, we propose an alternative clustering functional data stream method to implement existing techniques, and we address phenomena in which web data are expressed by a curve or a function. In particular, we deal with a specific type of web data, i.e. trends of google queries. Specifically, focusing on top football players data, we compare the functional k-means approach to the functional Hierarchical Clustering for detecting specific pattern of search trends over time.

Abstract Recentemente, l'analisi dei dati web é diventata essenziale in molti campi di ricerca. Informazioni sulle ricerche nei motori di ricerca, sul numero di visite, sul tempo trascorso sul sito web, sulla provenienza degli utenti e sul successo di una pubblicitá online, sono essenziali per prevedere vendite future, analizzare le performance passate e monitorare i modelli di acquisto. Tuttavia, lo studio del traffico web presenta alcuni problemi metodologici in quanto questi dati fluiscono continuamente inibendone la totale archiviazione ed ostacolando l'applicazione di tecniche di analisi standard. Questo articolo propone un metodo alternativo di classificazione rispetto a quelli noti in letteratura affrontando situazioni in cui i dati web sono

Fabrizio Maturo
"G. d' Annunzio" University, DEA, Pescara, Italy, e-mail: f.maturo@unich.it

Francesca Fortuna
"G. d' Annunzio" University, DISFPEQ, Pescara, Italy, e-mail: francesca.fortuna@unich.it

Tonio Di Battista
"G. d' Annunzio" University, DISFPEQ, Pescara, Italy, e-mail: dibattis@unich.it

espressi da una curva o da una funzione. L'analisi é condotta su un tipo specifico di dati web, cioé l'andamento delle ricerche su Google. Nello specifico, concentrandoci sui dati dei migliori giocatori di calcio, l'obiettivo é quello di confrontare l'approccio delle k-medie con quello della cluster gerarchica per dati funzionali al fine di rilevare modelli di comportamento nelle tendenze di ricerca attraverso il tempo.

Key words: clustering football players, data streaming, google query, google trends, FDA

1 Introduction

In recent decades, the analysis of web-data has attracted the attention of many companies dealing with marketing and advertising in the Sports sector, and particularly in the football industry. Due to the massive diffusion of high speed network and internet technology, the analysis of consumers' behavior in surfing the web has become a fundamental data for understanding consumer preferences. Data collected on the web are referred to as data streams, whose main characteristics are that they continuously flow. Since data streams are potentially infinite, the identification of common patterns through clustering techniques become a crucial issue. However, clustering methods suffer from many problems related to the nature of the data, such as clusters robustness over time, the difficulty of exploring data at different portions of the stream, and the computational time-consuming issue. Because each stream can be seen as a function in a continuous domain, i.e. time, this paper proposes to analyze web-data through the functional data analysis (FDA) approach [Ramsay and Silverman, 2005] and discusses the benefits of this method focusing on the clustering data streams problem. The main advantage provided by FDA approach is the drastic dimensionality reduction of data streams, which is gained by converting discrete observations into functions. Moreover, the use of additional functional tools may sometimes reveal crucial information better than the original data [Ramsay and Silverman, 2005, Ferraty and Vieu, 2006, Maturo and Di Battista, 2018, Fortuna and Maturo, 2018, Maturo, 2018, Maturo et al., 2018]. Specifically, in this paper, the FDA approach is proposed for clustering data streams using either hierarchical clustering and K-means approaches. In this study, we focus our attention on a semimetric based on the functional principal components (FPCs) decomposition for both clustering technique. The methodological approach is implemented for analyzing a real data set concerning the Google queries regarding 20 top football players. The final aim of this contribution is to propose the use of functional clustering methods for the analysis of web-queries, and to compare the results of the main two functional clustering approaches.

2 Materials and Methods

The basic idea behind a functional interpretation of web-data is that each stream is a sample from a smooth function of time. Because in real applications curves are observed at a finite set of sampling points, the functional form of a stream, say $X(t)$, can be reconstructed and represented by a basis expansion as follows: $X_i(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t) \quad i = 1,....,n$ where $X_i(t)$ is the reconstructed function for the i-th unit; $\phi_k(t)$ are linearly independent and known basis functions; and a_{ik} are coefficients that link each basis function together in the representation of $X_i(t)$. Various basis systems can be adopted, depending on the characteristics of curves [Ramsay, 1991]. In this work, we consider the least squares approximation with B-splines basis for the functional representation of web-data. To achieve an optimal representation of curves into a function space of reduced dimension, the functional principal component analysis (FPCA) can be adopted [Ferraty and Vieu, 2006]. In particular, let us assume that the observed curves are centered so that the sample mean is equal to zero. Then, for each unit ($i = 1,..,n$), the j-th principal component score is given by $\xi_{ij} = \int_T x_i(t) f_j(t)\,dt$, where the weight functions or loadings $f_j(t)$ are the solutions of the eigenequation where $c(t,s)$ is the sample covariance function and $\lambda_j = Var[\xi_j]$ [Aguilera and Aguilera-Morillo, 2013]. Then, the sample curves admit the following principal component decomposition: $x_i(t) = \sum_{j=1}^{p} \xi_{ij} f_j(t) \quad i = 1,...,n$ where p denotes the total number of functional principal components. By truncating this representation in terms of the first q principal components ($q << p$), we can obtain an approximation of the sample curves whose explained variance is given by $\sum_{j=1}^{q} \lambda_j$. If we assume that the observed functions are expressed in terms of B-splines, then the weight functions $f_j(t)$, admit the following basis expansion: $f_j(t) = \sum_{k=1}^{K} b_{jk}\phi_k(t),\ j = 1,2,...,q$. To identify specific common patterns of data-streams, clustering of functions is carried out in combination with dimension reduction in order to remove the effect of irrelevant functional information. In particular, the distance among functional data is computed using the FPCA method according to the following semi-metric [Ferraty and Vieu, 2006]: $d^{(q)}\left(X_i(t), X_{i'}(t)\right) = \left[\sum_{j=1}^{q} \left(\xi_{ij} - \xi_{i'j}\right)^2 \left\|f_j^{(q)}\right\|\right]^{1/2} \quad i \neq i'$ where $\left\|f_j^{(q)}\right\| = \int_T f_j(t)^2\,dt$, and q denotes the reduced dimensional space at q components [Ferraty and Vieu, 2006, Febrero-Bande and de la Fuente, 2012], chosen according to the criterion of their explained variability.

The basic idea of this unsupervised clustering approach is to find a partition for which the variability within clusters is minimized. The most used algorithm, in this context, is the k-means. Starting from n functional observations, this method aims to assembly units into $G \leq n$ groups, $C_1, C_2, ..., C_G$ so as to minimize the within-cluster sum of squares. The first step of this iterative procedure consists in fixing G initial centroids, $\psi_1^{(0)}(t), ..., \psi_G^{(0)}(t)$. Then, each function is assigned to the cluster whose centroid, at the previous iteration $(m-1)$ is the nearest according to the chosen distance. Once all the functions have been assigned to a cluster, the cluster means are updated as follows: $\psi_g^m(t) = \sum_{x_i(t)\in C_g} \frac{x_i(t)}{n_g}$ where n_g is the number of functions in the g-th cluster, C_g.

Despite this approach could be extended to first and second derivatives or other functions derived from the original ones [Fortuna et al., 2018, Fortuna and Maturo, 2018], in this context, we limit our attention to the original b-spline approximation. Effectively, the aim of this paper is to propose thise approach and compare its results to the agglomerative hierarchical method. In this setting, the classification strategy consists of a series of partitions, which may run from a single cluster containing all the functions (divisive methods), to n clusters, each containing a single function (agglomerative methods). In order to determine which groups should be merged (for agglomerative approach) or divided (for divisive approach), different metrics and linkage methods can be used. In this context, the agglomerative approach with the average linkage method is used. In addition, the gap statistic [Tibshirani et al., 2001] and silhouette plot are adopted for estimating the optimal number of clusters [Kassambara and Mundt, 2017].

3 Application

The proposed method has been applied to the number of queries collected by Google trends over two months (from the beginning of January 2018 to the end of February 2018) regarding the 19 football players with highest market values ($\geq$50 million euros), valued on December 19, 2017 (https://www.transfermarkt.it/); and participation in 2017-2018 UEFA Champions League. Table 1 shows the full list of statistical units considered in the study. Figures 1 and 2 illustrate top player queries over time and the corresponding spline approximation, respectively. Figure 3 presents the functional principal components decomposition: the first three FPCs explain 56.93%, 16.11%, and 8.99% of the total variability, respectively. Figure 4 exhibits the results of the k-means functional clustering of the 19 top players' queries over time whereas Figures 5 and 6 presents the hierarchical clustering findings. According to the gap statistic 7 and silhouette analysis 8, the optimal number of groups is two. Effectively, Figure 4 strongly highlights the presence of two distinct groups: the groups n.1 is characterized by higher variability and elevated levels of queries over time; conversely, the group n.2 presents low variability over time and a minor average level of queries. Table 1 specifies in detail the group membership according to the two different methods. 16 soccer players are distinguished by the same pattern of groups' membership. Only Ronaldo, Hazard, and Kroos change group in passage from k-means to hierarchical clustering.

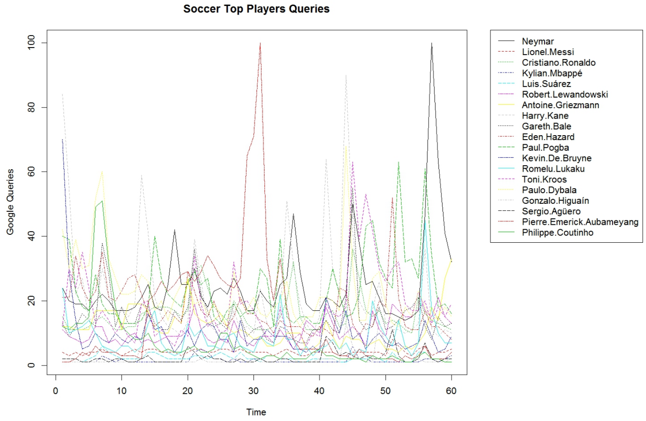


Fig. 1: Top Player Queries Over Time

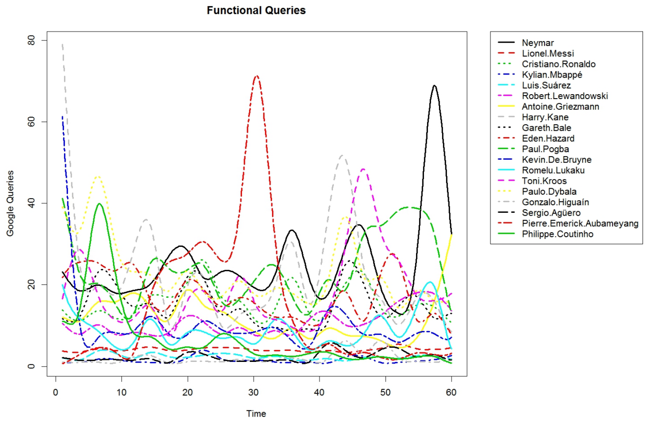


Fig. 2: Spline Approximation of Top Player Queries Over Time

4 Conclusion

The basic idea of this study is to use FDA to analyse datastreaming and in particular the number of queries on search engines. Hence, the originality of this research lies in the application, and in particular we have focused on cluster analysis using the two best known methods: the k-means and hierarchical clustering. Naturally, starting from this idea, many possible developments can be done in the field of functional regression, prediction, and supervised classification. In future research, it would be interesting to analyze in depth what are the reasons that lead to conflicting results between the two types of clustering methods. To achieve this goal, we should

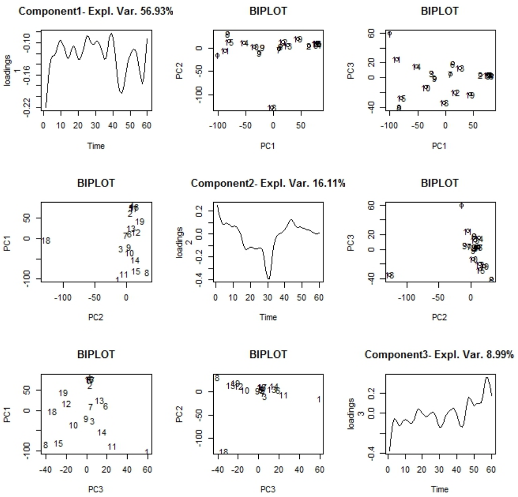


Fig. 3: Functional Principal Components Decomposition of Spline Approximation of Top Player Queries Over Time

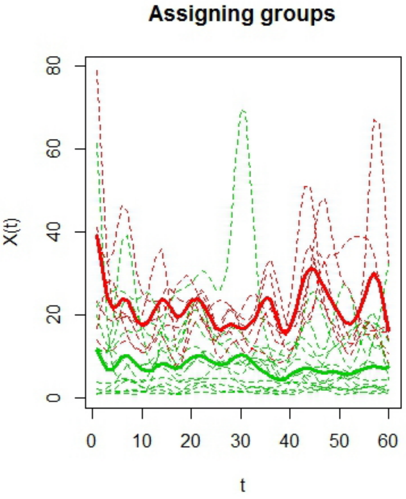


Fig. 4: K-Means Clustering of Top Player Queries Over Time (Red=Group n.1, Green=Group n.2)

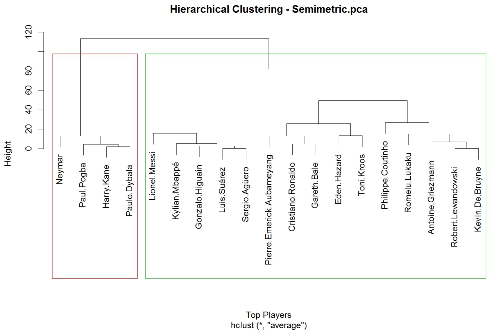


Fig. 5: Hierarchical Clustering of Top Player Queries Over Time (Red=Group n.1, Green=Group n.2)

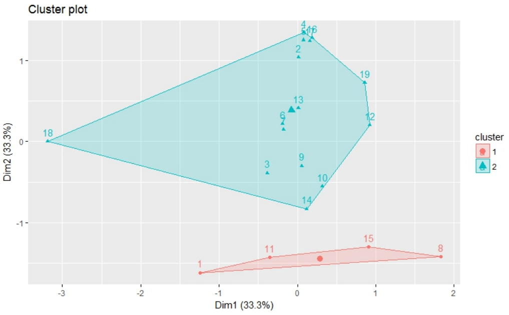


Fig. 6: Plot of Hierarchical Clustering of Top Player Queries Over Time

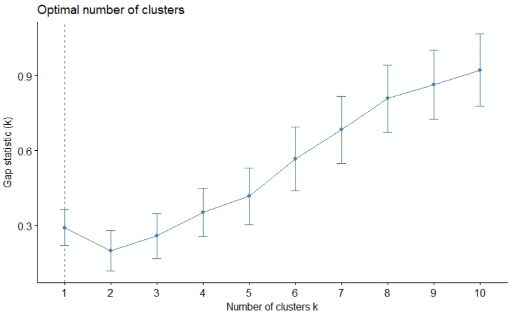


Fig. 7: Gap Statistic for Selecting Groups Number

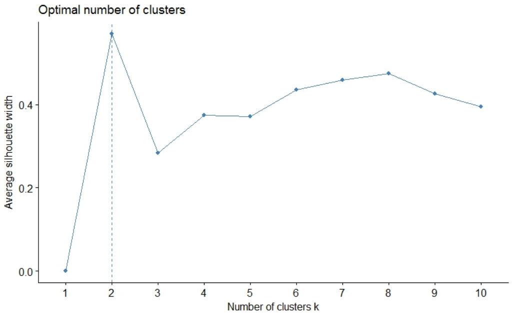


Fig. 8: Silhouette Analysis for Selecting Groups Number

No.	Player name	Age	Football club	Group (K-Means)	Group (Hierarchical)
1	Neymar	25	FC Paris Saint-G.	1	1
2	Messi	30	FC Barcelona	2	2
3	Ronaldo	30	Real Madrid CF	1	2
4	Mbappé	18	FC Paris Saint-G.	2	2
5	Suárez	30	FC Barcelona	2	2
6	Lewandowski	29	FC Bayern Monaco	2	2
7	Griezmann	26	Atlético de Madrid	2	2
8	Kane	24	Tottenham Hotspur	1	1
9	Bale	28	Real Madrid CF	2	2
10	Hazard	26	Chelsea FC	1	2
11	Pogba	24	Manchester United	1	1
12	De Bruyne	26	Manchester City	2	2
13	Lukaku	24	Manchester United	2	2
14	Kroos	27	Real Madrid CF	1	2
15	Dybala	24	Juventus FC	1	1
16	Higuaín	30	Juventus FC	2	2
17	Ag*ü*ero	29	Manchester City	2	2
18	Aubameyang	28	Borussia Dortmund	2	2
19	Coutinho	25	FC Liverpool	2	2

Table 1: Group Membership of 19 Football Top Players according to two functional clustering methods.

also draw this study on the use of further semi-metrics, e.g. based on first and second derivatives, and different types of hierarchical clustering approaches, e.g. single-linkage and complete-linkage.

References

A. Aguilera and M. Aguilera-Morillo. Penalized PCA approaches for b-spline expansions of smooth functional data. *Applied Mathematics and Computation*, 219 (14):7805–7819, mar 2013. doi: 10.1016/j.amc.2013.02.009.

M. Febrero-Bande and M. de la Fuente. Statistical computing in functional data analysis: The r package fda.usc. *Journal of Statistical Software, Articles*, 51(4): 1–28, 2012. doi: 10.18637/jss.v051.i04.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer, New York, 2006.

F. Fortuna and F. Maturo. K-means clustering of item characteristic curves and item information curves via functional principal component analysis. *Quality & Quantity*, mar 2018. doi: 10.1007/s11135-018-0724-7.

F. Fortuna, F. Maturo, and T. Di Battista. Unsupervised classification of soccer top players based on google trends. *Quality and Reliability Engineering International*, 2018. doi: 10.1002/QRE-17-0561.

A. Kassambara and F. Mundt. factoextra: Extract and visualize the results of multivariate data analyses. 2017. URL `https://cran.r-project.org/web/packages/factoextra/index.html`.

F. Maturo. Unsupervised classification of ecological communities ranked according to their biodiversity patterns via a functional principal component decomposition of hill's numbers integral functions. *Ecological Indicators*, 90:305–315, jul 2018. doi: 10.1016/j.ecolind.2018.03.013.

F. Maturo and T. Di Battista. A functional approach to Hill's numbers for assessing changes in species variety of ecological communities over time. *Ecological Indicators*, 84(C):70 – 81, 2018. doi: 10.1016/j.ecolind.2017.08.016.

F. Maturo, F. Fortuna, and T. Di Battista. Testing equality of functions across multiple experimental conditions for different ability levels in the IRT context: The case of the IPRASE TLT 2016 survey. *Social Indicators Research*, apr 2018. doi: 10.1007/s11205-018-1893-4.

J. Ramsay. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56:611–630, 1991.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis, 2nd edn*. Springer, New York, 2005.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, may 2001. doi: 10.1111/1467-9868.00293.

Co-clustering algorithms for histogram data
Algoritmi di Co-clustering per dati ad istogramma

Francisco de A.T. De Carvalho and Antonio Balzanella and Antonio Irpino and
Rosanna Verde

Abstract One of the current big-data age requirements is the need of representing groups of data by summaries allowing the minimum loss of information as possible. Recently, histograms have been used for summarizing numerical variables keeping more information about the data generation process than characteristic values such as the mean, the standard deviation, or quantiles. We propose two co-clustering algorithms for histogram data based on the double *k*-means algorithm. The first proposed algorithm, named ”distributional double Kmeans (DDK)”, is an extension of double Kmeans (DK) proposed to manage usual quantitative data, to histogram data. The second algorithm, named adaptive distributional double Kmeans (ADDK), is an extension of DDK with automated variable weighting allowing co-clustering and feature selection simultaneously.

Abstract *Una delle principali esigenze nell'era dei big-data è quella di rappresentare gruppi di dati attraverso strumenti di sintesi che minimizzano la perdita di informazione. Negli ultimi anni, uno degli strumenti maggiormente utilizzati a tal scopo è l'istogramma. Esso fornisce una sintesi della distribuzione che genera i dati risultando più informativo delle classiche sintesi quali la media, la deviazione standard, o i quantili. Nel presente articolo, si propongono due algoritmi di co-clustering per dati rappresentati da istogrammi che estendono il classico double k-means algorithm (DK). La prima proposta chiamata ”distributional double Kmeans (DDK)”, è un'estensione dell'algoritmo DK a dati ad istogramma. La seconda pro-*

Francisco de A.T. De Carvalho
Centro de Informatica, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes s/n - Cidade Universitaria, CEP 50740-560, Recife-PE, Brazil, e-mail: fatc@cin.ufpe.br

Antonio Balzanella
Università della Campania L. Vanvitelli, e-mail: antonio.balzanella@unicampania.it

Antonio Irpino
Università della Campania L. Vanvitelli, e-mail: antonio.irpino@unicampania.it

Rosanna Verde
Università della Campania L. Vanvitelli e-mail: rosanna.verde@unicampania.it

posta, chiamata adaptive distributional double Kmeans (ADDK), è un'estensione dell'algoritmo DDK che effettua la ponderazione automatica delle variabili consentendo di effettuare simultaneamente il co-clustering e la selezione delle variabili.

Key words: Co-clustering, Histogram data

1 Introduction

Histogram data are becoming very common in several applicative fields. For example, in order to preserve the individuals' privacy, data about groups of customers transactions are released after being aggregated; in wireless sensor networks, where the energy limitations constraint the communication of data, the use of suitable synthesis of sensed data are necessary; official statistical institutes collect data about territorial units or administrations and release them as histograms.

Among the exploratory tools for the analysis of histogram data, this paper focuses on co-clustering, also known as bi-clustering or block clustering. The aim is to cluster simultaneously objects and variables of a data set [1, 2, 5, 6].

By performing permutations of rows and columns, the co-clustering algorithms aim to reorganize the initial data matrix into homogeneous blocks. These blocks also called co-clusters can therefore be defined as subsets of the data matrix characterized by a set of observations and a set of features whose elements are similar. They resume the initial data matrix into a much smaller matrix representing homogeneous blocks or co-clusters of similar objects and variables. Refs. [4] presents other types of co-clustering approaches.

This paper proposes at first the DDK (Distributional Double K-means) algorithm whose aim is to cluster, simultaneously, objects and variables on distributional-valued data sets. Then, it introduces the ADDK (Adaptive Distributional Double K-means) algorithm which takes into account the relevance of the variables in the co-clustering optimization criterion.

Conventional co-clustering methods do not take into account the relevance of the variables, i.e., these methods consider that all variables are equally important to the co-clustering task, however, in most applications some variables may be irrelevant and, among the relevant ones, some may be more or less relevant than others.

ADDK and DDK use, respectively, suitable adaptive and non-adaptive Wasserstein distances aiming to compare distributional-valued data during the co-clustering task.

2 Co-clustering algorithms for distributional-valued data

Let $E = \{e_1, \ldots, e_N\}$ be a set of N objects described by a set of P distributional-valued variables denoted by $Y_j \, (1 \leq j \leq P)$. Let $\mathscr{Y} = \{Y_1, \ldots, Y_P\}$ be the set of P

distributional-valued variables and let

$$\mathbf{Y} = \left(y_{ij}\right)_{\substack{1\le i\le N\\ 1\le j\le P}}$$

be a distributional-valued data matrix of size $N \times P$ where the distributional data observed on the Y_j variable for the i-th object is denoted with y_{ij}. Our aim consists in obtaining a co-clustering of $\mathbf{Y}$, i.e, in obtaining simultaneously a partition $\mathscr{P} = \{\mathscr{P}_1, \ldots, \mathscr{P}_C\}$ of the set of N objects into C clusters and a partition $\mathscr{Q} = \{\mathscr{Q}_1, \ldots, \mathscr{Q}_H\}$ of the set of P distributional-valued variables into H clusters.

The co-clustering can be formulated as the search for a good matrix approximation of the original distributional-valued data matrix $\mathbf{Y}$ by a $C \times H$ matrix

$$\mathbf{G} = \left(g_{kh}\right)_{\substack{1\le k\le C\\ 1\le h\le H}}$$

which can be viewed as a summary of the distributional-valued data matrix $\mathbf{Y}$ (see, for example, Refs. [2, 5, 6]).

Each element g_{kh} of $\mathbf{G}$ is also called a prototype of the co-cluster

$$\mathbf{Y}_{kh} = \left(y_{kj}\right)_{\substack{e_i\in\mathscr{P}_k\\ Y_j\in\mathscr{Q}_h}}$$

Moreover, each g_{kh} is a distributional data, with a distribution function G_{kh} and a quantile function $Q_{g_{kh}}$

In order to obtain a co-clustering that is faithfully representative of the distributional-valued data set $\mathbf{Y}$, the matrix $\mathbf{G}$ of prototypes, the partition $\mathscr{P}$ of the objects and the partition $\mathscr{Q}$ of the distributional-valued variables are obtained iteratively by means of the minimization of an error function J_{DDK}, computed as follows:

$$J_{DDK}(\mathbf{G}, \mathscr{P}, \mathscr{Q}) = \sum_{k=1}^{C}\sum_{h=1}^{H}\sum_{e_i\in\mathscr{P}_k}\sum_{Y_j\in\mathscr{Q}_h} d_W^2(y_{ij}, g_{kh}) \tag{1}$$

where the d_W^2 function is the non-adaptive (squared) L_2 Wasserstein distance computed between the element y_{ij} of the distributional data matrix $\mathbf{Y}$ and the prototype g_{kh} of co-cluster $\mathbf{Y}_{kh}$

In order to propose an adaptive version of the DDK algorithm which evaluates the relevance of each variable in the co-clustering process, we still propose the local minimization of the following criterion function:

$$J_{ADDK}(\mathbf{G}, \Lambda, \mathscr{P}, \mathscr{Q}) = \sum_{k=1}^{C}\sum_{h=1}^{H}\sum_{e_i\in\mathscr{P}_k}\sum_{Y_j\in\mathscr{Q}_h} \lambda_j d_W^2(y_{ij}, g_{kh}) \tag{2}$$

where $\Lambda = (\lambda_j)_{j=1,\ldots,P}$ (with $\lambda_j > 0$ and $\prod_{j=1}^{P} \lambda_j = 1$) are positive weights measuring the importance of each distributional-valued variable.

The minimization of the J_{DDK} criterion, is performed iteratively in three steps: representation, objects assignments and distributional-valued variables assignments.

The representation step gives the optimal solution for the computation of the representatives (prototypes) of the co-clusters. The objects assignments step provides the optimal solution for the partition of the objects. Finally, the variables assignments step provides the optimal solution for the partition of the variables. The three steps are iterated until the convergence to a stable optimal solution.

The minimization of the J_{ADDK} criterion, requires a further weighting step which provides optimal solutions for the computation of the relevance weights for the distributional-valued variables.

Representation step in DDK and ADDK

In the representation step, DDK algorithm aims to find, for $k = 1,\dots,C$ and for $h = 1,\dots,H$, the prototype g_{kh} such that $\sum_{e_i \in \mathscr{P}_k} \sum_{Y_j \in \mathscr{Q}_h} d_W^2(y_{ij}, g_{kh})$ is minimal.

According to [3], the quantile function associated with the corresponding probability density function (*pdf*) g_{kh} $(1 \leq k \leq C; 1 \leq h \leq H)$ is:

$$Q_{g_{kh}} = \frac{\sum_{e_i \in \mathscr{P}_k} \sum_{Y_j \in \mathscr{Q}_h} Q_{ij}}{n_k n_h} \tag{3}$$

where n_k is the cardinality of $\mathscr{P}_k$ and n_h is the cardinality of $\mathscr{Q}_h$.

The ADDK algorithm aims to find, for $k = 1,\dots,C$ and for $h = 1,\dots,H$, prototype g_{kh} such that $\sum_{e_i \in \mathscr{P}_k} \sum_{Y_j \in \mathscr{Q}_h} \lambda_j d_W^2(y_{ij}, g_{kh})$ is minimal.

Under the constraints $\prod_{j=1}^{P} \lambda_j = 1$, $\lambda_j > 0$, the quantile function associated with the corresponding probability density function (*pdf*) g_{kh} $(1 \leq k \leq C; 1 \leq h \leq H)$ is computed as follows:

$$Q_{g_{kh}} = \frac{\sum_{e_i \in \mathscr{P}_k} \sum_{Y_j \in \mathscr{Q}_h} \lambda_j Q_{ij}}{n_k \sum_{Y_j \in \mathscr{Q}_h} \lambda_j} \tag{4}$$

Objects assignment step in DDK and ADDK

During the object assignment step of DDK, the matrix of co-cluster prototypes $\mathbf{G}$ and the partition of the distributional-valued variables $\mathscr{Q}$ are kept fixed. The error function J_{DDK} is minimized with respect to the partition $\mathscr{P}$ of objects and each object $e_i \in E$ is assigned to its nearest co-cluster prototype.

Proposition 1. *The error function J_{DDK} (Eq. 1) is minimized with respect to the partition $\mathscr{P}$ of objects when the clusters P_k $(k = 1,\dots,C)$ are updated according to the following assignment function:*

$$P_k = \left\{ e_i \in E : \sum_{h=1}^{H} \sum_{Y_j \in \mathscr{Q}_h} d_W^2(y_{ij}, g_{kh}) = \min_{z=1}^{C} \sum_{h=1}^{H} \sum_{Y_j \in \mathscr{Q}_h} d_W^2(y_{ij}, g_{zh}) \right\}$$

The error function J_{ADDK} is minimized with respect to the partition $\mathscr{P}$ and each individual $e_i \in E$ is assigned to its nearest co-cluster prototype.

Proposition 2. *The error function J_{ADDK} (Eq. 2) is minimized with respect to the partition $\mathscr{P}$ of objects when the clusters $P_k\,(k = 1,\ldots,C)$ are updated according to the following assignment function:*

$$P_k = \left\{ e_i \in E : \sum_{h=1}^{H} \sum_{Y_j \in \mathscr{Q}_h} \lambda_j d_W^2(y_{ij}, g_{kh}) = \min_{z=1}^{C} \sum_{h=1}^{H} \sum_{Y_j \in \mathscr{Q}_h} \lambda_j d_W^2(y_{ij}, g_{zh}) \right\}$$

Variables assignment step in DDK and ADDK

During the variables assignment step of DDK, the matrix of prototypes **G** and the partition of the objects $\mathscr{P}$ are kept fixed. The error function J_{DDK} is minimized with respect to the partition $\mathscr{Q}$ of the distributional-valued variables and each variable $Y_j \in \mathscr{Y}$ is assigned to its nearest co-cluster prototype.

Proposition 3. *The error function J_{DDK} (Eq. 1) is minimized with respect to the partition $\mathscr{Q}$ of the distributional-valued variables when the clusters $Q_h\,(h = 1,\ldots,H)$ are updated according to the following assignment function:*

$$Q_h = \left\{ Y_j \in \mathscr{Y} : \sum_{k=1}^{C} \sum_{e_i \in \mathscr{P}_k} d_W^2(y_{ij}, g_{kh}) = \min_{z=1}^{H} \sum_{k=1}^{C} \sum_{e_i \in \mathscr{P}_k} d_W^2(y_{ij}, g_{kz}) \right\}$$

where $d_W^2(y_{ij}, g_{kz})$ is the squared L^2 Wasserstein distance.

Proposition 4. *The error function J_{ADDK} (Eq. 2) is minimized with respect to the partition $\mathscr{Q}$ of the distributional-valued variables when the clusters $Q_h\,(h = 1,\ldots,H)$ are updated according to the following assignment function:*

$$Q_h = \left\{ Y_j \in \mathscr{Y} : \sum_{k=1}^{C} \sum_{e_i \in \mathscr{P}_k} \lambda_j d_W^2(y_{ij}, g_{kh}) = \min_{z=1}^{H} \sum_{k=1}^{C} \sum_{e_i \in \mathscr{P}_k} \lambda_j d_W^2(y_{ij}, g_{kz}) \right\}$$

Weighting step for ADDK

We provide an optimal solution for the computation of the relevance weight of each distributional-valued variable during the weighting step of the ADDK algorithm.

During the weighting step of ADDK, the matrix of prototype vectors **G**, the partition $\mathscr{P}$ of the objects and the partition $\mathscr{Q}$ of the distributional-valued variables are kept fixed. The error function J_{ADDK} is minimized with respect to the weights λ_j.

Proposition 5. *The relevance weights are computed according to the adaptive squared L_2 Wasserstein distance:*

If we assume that $\prod_{j=1}^{P} \lambda_j = 1$, $\lambda_j > 0$, *the P relevance weights are computed as follows:*

$$\lambda_j = \frac{\left\{ \prod\limits_{r=1}^{P} \left(\sum\limits_{k=1}^{C} \sum\limits_{h=1}^{H} \sum\limits_{e_i \in \mathscr{P}_k} \sum\limits_{Y_r \in \mathscr{Q}_h} d_W^2(y_{ir}, g_{kh}) \right) \right\}^{\frac{1}{P}}}{\sum\limits_{k=1}^{C} \sum\limits_{h=1}^{H} \sum\limits_{e_i \in \mathscr{P}_k} \sum\limits_{Y_j \in \mathscr{Q}_h} d_W^2(y_{ij}, g_{kh})} \tag{5}$$

3 Conclusions

In this paper we have introduced two algorithms, based on the double k-means, for performing the co-clustering of a distributional valued data matrix. The main difference between the two algorithms is that ADDK integrates in the optimization criterion the search for a set of weights which measure the relevance of each variable. In order to evaluate the effectiveness of the proposal, we have made some preliminary test on real and simulated data with encouraging results.

References

1. Govaert G.: Simultaneous clustering of rows and columns. In: Control and Cybernetics **24** pp. 437–458 (1995)
2. Govaert G., Nadif M.: Co-Clustering: Models, Algorithms and Applications. Wiley, New York (2015)
3. Irpino, A. and Verde, R. Basic statistics for distributional symbolic variables: a new metric-based approach. In: Advances in Data Analysis and Classification,92, pp. 143–175 Springer Berlin Heidelberg (2015)
4. Pontes R., Giraldez R., Aguilar-Ruiz J.S.: Biclustering on expression data: A review. In: Journal of Biomedical Informatics, 57, pp. 163–180, (2015)
5. Rocci R., Vichi M.: Two-mode multi-partitioning. In: Computational Statistics & Data Analysis, 52 pp.1984–2003 (2008)
6. Vichi M.: Double k-means Clustering for Simultaneous Classification of Objects and Variables. In: Advances in Classification and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 43–52, Springer (2001)

A network approach to dimensionality reduction in Text Mining

Un approccio di rete per la riduzione dimensionale nel Text Mining

Michelangelo Misuraca, Germana Scepi and Maria Spano

Abstract There is an ever-increasing interest in developing statistical tools for extracting information from documental repositories. In a Text Mining frame, a knowledge discovery process usually implies a dimensionality reduction of the vocabulary, via a feature selection and/or a feature extraction. Here we propose a strategy designed for reducing dimensionality through a network-based approach. Network tools allow performing the reduction by considering the most important relations among the terms. The effectiveness of the strategy will be shown on a set of tweets about the 2018 Italian General Election.

Abstract *Lo sviluppo di strumenti statistici per estrarre informazioni da archivi documentali è oggi sempre più importante. Nel Text Mining il processo di knowledge discovery solitamente include uno step di riduzione dimensionale, con procedure di selezione o estrazione dei termini. In questo lavoro si propone una strategia per ridurre la dimensionalità attraverso un approccio di rete. Gli strumenti per lanalisi di rete consentono di operare la riduzione considerando le pi importanti relazioni tra i termini. L'efficacia della strategia è mostrata su un insieme di tweet relativi alle Elezioni Italiane del 2018.*

Key words: vector space model, network analysis, community detection.

1 Introduction

The incredible progress of computer technology, as well as the growth of the Internet, has fastened in recent years the transition from analog to digital data communication and storage. This revolution necessarily involved also the statistical rea-

Michelangelo Misuraca
Università della Calabria - Arcavacata di Rende, e-mail: michelangelo.misuraca@unical.it

Germana Scepi, Maria Spano
Università Federico II di Napoli, e-mail: germana.scepi@unina.it, maria.spano@unina.it

soning. In each domain in which Statistics can help researchers in finding trends, revealing patterns or explaining relations, there is nowadays an over-availability of data. This massive amount of data required the development of mining techniques for discovering and extracting knowledge, because only a part of the collected data is relevant and informative with respect to the phenomena of interest.

This instance is particularly important when we consider texts. Texts express a wide, rich range of information, but encode this information in a form difficult to automatically process. Aiming at analysing a collection of documents written in natural language, it is then necessary to transform the original unstructured data into structured data. In this framework, the most common algebraic model for representing documents is the so-called *vector space model* [14]: a document is a vector in the (extremely sparse) space spanned by the terms. Because of vector space model, each document contains a lot of noise. Documents are seen as *bag-of-words*, i.e. as an unordered set of terms, disregarding grammatical and syntactical roles. The focus is on the presence/absence of terms belonging to the collection's vocabulary, their characterisation and their discrimination power. To reduce space dimensionality, it is possible to consider feature selection procedures and/or feature extraction methods. Feature selection allows filtering a subset of the original terms, by excluding the less informative and discriminative ones or by considering only the most relevant ones, with respect to a given criterion. Feature extraction performs a reduction of the original space by combining the terms into new entities. One of the main differences is that selection techniques retain the original meaning of terms, where extraction techniques require an additional effort in interpreting the results.

To reduce the dimensionality and save the readability of the results, here we propose a new feature extraction strategy based on network analysis. Network analyses allow visualising the relations among the terms of a collection by recovering the context of use. This paper is structured as follows. In Section 2 the reference literature is reviewed. Section 3 introduces the problem and describes the proposed strategy. In Section 4 the effectiveness of the proposal is showed on a set of tweets about the 2018 Italian General Election. Final remarks and some possible future development of the research are discussed in Section 5

2 Background and related work

In the statistical analysis of large collections of documents, one of the main problems is the high-dimensionality of data. Once we have pre-processed the documents via the *bag-of-words* coding, every single term belonging to the collection's vocabulary represents a dimension in the vector space. The *documents* × *terms* lexical table − obtained by juxtaposing the different document-vectors − is usually a large and very sparse matrix. Since only a part of the terms is actually relevant for expressing the informative content of the documents, the noise in the data has to be eliminated because it leads to unreliable results and significantly increases the computational complexity. Two approaches are used to deal with the problem of dimensionality reduction in Text Mining: *feature selection* and *feature extraction*.

Feature selection methods aim at finding a subset of the original terms, by focusing on their importance in determining the content itself. One of the main advantages of these methods is that the selected features retain the original meaning and provides a better understanding of the results. The most common approach is using a *stop-list*. Typically in a stop-list we can find articles, prepositions, conjunctions and interjection. These terms can be considered unuseful, because they have a very general and weak lexical meaning. In addition, stop-lists may include terms related to the main topics listed in the document collection. On the other hand, many methods have been proposed aiming at selecting the most important terms in a collection [1]. The simplest criterion for term selection is the *document frequency*, i.e. the number of documents in which a term occurs. An alternative is the well-known *tf-idf* [13], which jointly considers the frequency of a term in a document and the document frequency of the term in the collection. Other methods for feature selection such as *term strength* [18], *term contribution* [10] and *entropy-based ranking* [3] are based on the concept of documents' similarity. More recently, other two similar methods for feature selection was proposed: *term variance* [9] and *term variance quality* [5]. Term variance evaluates the quality of terms by computing the variance of each term in the collection. It follows the same idea of document frequency, where terms with low frequency are not important. When class labels are available for a document collection, there are also some supervised methods for term selection. Methods such as *information gain*, *mutual information* and *Chi-square* statistics, have been successfully used in text classification [19].

Feature extraction methods aim at minimising the amount of resources required to describe a large collection of documents. The main task of this approach is to obtain the most relevant information from the original features – through some functional mapping – and represent the information in a lower dimensionality space. The dimensionality reduction is reached by constructing linear combinations of the original features that still describe the document collection with sufficient accuracy. The obvious drawback of feature extraction is that the new features may not have a clear physical meaning, therefore the results are difficult to interpret. In textual data analysis, methods like *principal component analysis* (PCA) [7], *lexical correspondences analysis* (LCA) [8], and *latent semantic analysis* (LSA) [4] are widely used. Although they have been developed in different contexts and with different aims, they are based on a common algebraic frame. A low-rank approximation of the lexical table is obtained via a generalised singular value decomposition. The differences among these techniques are related to the weights assigned to the elements, introducing different orthonormalising constraints. LSA is popular in an information retrieval framework for representing the semantic structures in a collection of documents. LCA is generally used to identify the association structure in the lexical table on factorial maps. More interesting is a non-linear approach to the problem, also known as *manifold learning*, that encompass for example the *isometric feature mapping* (ISOMAP) [16] and the *local linear embedding* (LLE) [12].

3 Problem definition and proposed strategy

Basically, both feature selection and feature extraction are carried out on the $documents \times terms$ lexical table. The vector space model ignores the context in which each term is used. It is possible to get back part of the structural and semantic information by constructing a $terms \times terms$ co-occurrence matrix. In general, each element of this latter matrix is the number of times two terms co-occur in the collection. This data structure can be represented as a network, where each vertex is a term and each edge is an element of the matrix different from 0. In this way, we can visualise both single terms and subsets of terms frequently co-occurring together.

Aiming at reducing the original dimensionality by a feature extraction process, in this paper we propose a community detection based strategy. Differently from the methods described in Sec. 2, our strategy preserves the original meaning of the terms and allows better readability. In a network, a community is a set of similar nodes that can be easily grouped. There is a lack of a universally accepted definition of community, but it is well known that most real-world networks display this kind of structures [6]. It is usually thought as a group where vertices are densely inter-connected and sparsely connected to other parts of the network [17]. From a theoretical viewpoint, community detection is not very different from clustering. Several algorithms have been proposed. Traditional approaches are based on the well-known hierarchical or partitional clustering [15]. Divisive approaches do not introduce substantial conceptual advances with respect to traditional ones. The main difference is that instead of removing edges between pairs of vertices with low similarity, inter-cluster edges are removed. The most popular algorithm for community detection was proposed by Newman and Girvan [11]. This work introduced the concept of *modularity* as a stopping criterion for the algorithm. Here in the follow, we consider the *fast-greedy* algorithm [2]. The advantage of using this algorithm is that the problem of choosing a grouping criterion is overcome by the direct use of modularity as optimisation function.

3.1 Basic notation and data structure

Let $\mathbf{T} = \{\mathbf{d_1}, \ldots, \mathbf{d_n}\} \subset \Re^p$ be a set of n document vectors in a term space of dimension p. This set can be represented as a $terms \times documents$ lexical table, where each element t_{ij} represents the number of occurrences of a term i into a document j $(i = 1, \ldots, p; j = 1, \ldots, n)$. Let transform $\mathbf{T}$ into a binary matrix $\mathbf{B}$, where the generic element b_{ij} is equal to 1 if the term i occurred at least once in document j, 0 otherwise. From the matrix $\mathbf{B}$, we derive the $terms \times terms$ co-occurrence matrix $\mathbf{A}$ by the product $\mathbf{A} \equiv \mathbf{B}\mathbf{B}^{\mathsf{T}}$. The generic element $a_{ii'}$ is the number of documents in which the term i and the term i' co-occur $(i \neq i')$. According to network theory, $\mathbf{A}$ is a $p \times p$ undirected weighted adjacency matrix that can be used to visualise the relations existing among the different terms.

3.2 Network-based feature extraction

On the matrix $\mathbf{A}$, we perform a community detection through a fast-greedy algorithm. This algorithm falls in the general family of agglomerative hierarchical clustering methods. As we said above, it is based on the optimisation of a quality function known as modularity. Modularity is the difference between the observed fraction of edges that fall within the given communities and the expected fraction in the hypothesis of random distribution. Suppose that the vertices of matrix $\mathbf{A}$ can be divided into two communities. The membership to one community or the other one is detected by a variable s, assuming values 1 or -1 respectively. The modularity Q is defined as:

$$Q = \frac{1}{2h} \sum_{ii'} \left[a_{ii'} - \frac{\delta_i \delta_{i'}}{2h} \right] s_i s_{i'} \tag{1}$$

where δ_i is the degree of the i-th term, h is the total number of edges in the network, and s_i represents the membership value of the term i to a community. When we consider G communities, eq. 1 can be expressed in terms of additional contribution ΔQ to the modularity. In matrix form we have:

$$\Delta Q = \frac{1}{4h} \mathbf{s}^{\mathsf{T}} \mathbf{M}^{(g)} \mathbf{s} \tag{2}$$

where $\mathbf{M}^{(g)}$ is the modularity matrix referred to the g-th community ($g = 1, \ldots, G$), and $\mathbf{s}$ is the binary column vector indicating the membership of each term to the community. The fast-greedy algorithm, starting with a state in which each term is the sole member of one of G communities, repeatedly joins communities together in pairs choosing in each step the join that results in the greatest increase in modularity. At the end of the detection process, we obtain a $terms \times communities$ matrix $\mathbf{C}$, a complete disjunctive table where the c_{ig} element is 0 or 1 when a term i belongs or not belongs to a community. The result of the dimensionality reduction is a *documents* × *communities* matrix $\mathbf{T}^{\star} \equiv (\mathbf{T}^{\mathsf{T}}\mathbf{C})\mathbf{D}_G{}^{-1}$, where $\mathbf{D}_G{}^{-1}$ is the diagonal matrix obtained from the column marginal distribution of $\mathbf{C}$. Each cell of $\mathbf{T}^{\star}$ contains the proportion of terms belonging to a community detected in the documents.

4 A study on the 2018 Italian Election campaign

The campaign for the 2018 General Election in Italy was different from the previous ones. In the past, streets were lined with political posters and candidates rallied potential voters around the country. But for the first time, there was no public refund to Italian parties for their campaign spending. Social media offered a less expensive but affordable way to reach voters in a largely unregulated forum.

Twitter is one of the most popular – and worldwide leading – social networking service. It can be seen as a blend of instant messaging, microblogging and texting, with brief content and a very broad audience. The embryonic idea was developed considering the exchange of texts like Short Message Service in a small group of users. We decided to focus our study on the official Twitter accounts of the first ten Italian political parties, considering the Election results. By using the *Twitter Archiver add-on* for Google Sheet, we collected 6094 tweets. We referred only to the last three months of the 2018 Election campaign, from January 1st to March 4th, 2018. We decided not filtering the so-called retweets, so that in the collection some texts were replicated more times. In Table 1, it is possible to see the main characteristics of the dataset. About 50% of the tweets were posted by the official account of the *Lega* party. Some of the parties, such as *Movimento 5 Stelle* and *Partito Democratico*, seemed to be less active on Twitter, but they showed a greater consensus both regarding the average number of *retweets* and *likes*.

Table 1 Statistics on the parties' accounts from January 1st to March 4th 2018

Party	# tweets	% tweets	avg. # retweets	avg. # like
+ Europa	417	6.84	37.8	103.2
Civica Popolare	86	1.41	7.5	13.5
Forza Italia	122	2.00	25.0	38.6
Fratelli d'Italia	789	12.95	7.8	18.3
Insieme	201	3.30	10.7	17.0
Lega	3025	49.64	7.3	14.5
Liberi e Uguali	560	9.19	36.3	75.9
Movimento 5 Stelle	134	2.20	232.7	398.3
Partito Democratico	288	4.73	114.5	227.7
Potere al Popolo	472	7.75	44.2	84.7

The pre-processing was performed in two phases, because of the peculiarity of tweets. Firstly, we stripped URLs, usernames, hashtags, emoticons and RT prefixes, and we normalised the tweets by removing other special characters and any separators than the blank spaces. Secondly, we performed a lemmatisation and tagged each term with the corresponding grammatical category. We decided to consider only the substantives and the adjectives because of their significant content-bearing role. Moreover, we deleted the terms appearing in the collection just one time from the vocabulary. The result of this step was a $documents \times terms$ matrix $\mathbf{T}$ with 6094 rows and 3610 columns, and a $terms \times terms$ co-occurrence matrix $\mathbf{A}$.

We performed the community detection procedure on $\mathbf{A}$ to extract the features. For better highlighting the relationships, we fixed a threshold of 15 co-occurrences and deleted the isolated terms. The greedy algorithm detected 20 communities. The high value of modularity ($Q = 0.83$) reveals the effectiveness of the procedure. In Table 2, the size of each community and the characterising terms are showed.

It is interesting to note that the algorithm identifies the communities related to the different parties' electoral manifestos. For instance, *C1* contains terms emphasising

Table 2 Communities in collection of tweets

ID	Size	Terms
C01	16	*fratello, diritto, futuro, italia, grande, unico, paese, ospite, tutto, europa, libertá, civile, proprio, unione, europeo, piúeuropa*
C02	9	*lavoratore, scuola, grasso, aggiunto, account, libero, uguale, cgil, buono*
C03	13	*italiano, marzo, estero, politico, istruzione, simbolo, elezione, melone, giorgia, semplice, palermo, croce, inciucio*
C04	9	*nuovo, comunitá, diretto, forza, momento, pagina, ordine, ufficiale, facebook*
C05	4	*potere, popolo, pubblico, spesa*
C06	9	*programma, partito, voto, utile, governo, pd, sinistro, ambiente, democratico*
C07	7	*anno, euro, popolare, ultimo, prossimo, casa, mila*
C08	2	*piano, natalitá*
C09	6	*centrodestra, berlusconi, lega, salvini, premier, matteo*
C10	2	*milano, duomo*
C11	4	*intervista, live, capitano, replay*
C12	3	*donna, violenza, uomo*
C13	2	*museo, egizio*
C14	2	*fake, news*
C15	3	*campagna, elettorale, legge*
C16	2	*made, italy*
C17	2	*giulia, bongiorno*
C18	2	*flat, tax*
C19	2	*movimento, stella*
C20	2	*sociale, centro*

the role of Italy in the European Union (e.g. *diritto, futuro, grande, unico, paese, libertá,. . .*) and the name of the party that promotes these aspects (+ *Europa*). Community *C6*, as well as *C9*, contains more trivial terms but specifically related to main opposite coalitions (*Partito Democratico* vs *Lega* and *Forza Italia*). Some of the communities are smaller, containing only a couple of terms, but still very important because they identify key themes of the campaign, such as *flat tax* and *piano natalitá*. The community detection procedure helps in reducing the dimensionality also by automatically identifying collocations and multiword expressions such as *fake news*, *centro sociale*, *made italy*.

By selecting only the terms belonging to the different communities, we obtain a 6094×20 matrix $\mathbf{T}^{\star}$, which can be used for further statistical analyses.

5 Some remarks and future developments

The proposed strategy aims at extracting features from a collection of documents by detecting high-level structures, i.e. communities. Each community is a new feature that retains the meaning of the single terms, and it can be seen as a concept/topic relevant for the domains to which the collection is referred. The strategy is suitable when we deal with short texts, as in the case study, but can also be applied to other kind of documents. One of the advantages of this approach, compared with the other proposal in the literature, is that the dimensionality is reduced by detecting collocations, multiwords and other structure. This reduction can also be seen as the first step for other analyses. Future developments of this work are devoted to automatically set a co-occurrence threshold in the community detection step, and to evaluate alternative similarity indices for measuring the relation strength among terms.

Acknowledgements The research work was supported by the *Start-(h)open* project, funded in the frame of the OP ERDF ESF 14/20 by the grant J28C17000380006 of the Regional Government of Calabria (Italy). This paper is dedicated to the memory of our beloved colleague, Simona Balbi.

References

1. Bharti, K. K., Singh, P. K.: A Survey on Filter Techniques for Feature Selection in Text Mining. In: Babu, B. V. *et al.* (eds.) Proc. of the 2^{nd} Int. Conference on Soft Computing for Problem Solving (SocProS12), pp. 1545-1559. Springer (2014)
2. Clauset A., Newman, M. E., Moore, C.: Finding community structure in very large networks. Phys. Rev. E. **70**, 066111 (2004)
3. Dash, M., Liu, H.: Feature Selection for Clustering. In: Proc. of the 4^{th} Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD00), pp.110-121. Springer (2000)
4. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by Latent Semantic Analysis. J. Am. Soc. Inform. Sci. **41**, 391–407 (1990)
5. Dhillon, I., Kogan, J., Nicholas, C.: Feature selection and document clustering. In: Berry, M.W. (ed.) Survey of Text Mining. Clustering, Classification, and Retrieval, pp. 73-100. Springer (2004)
6. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
7. Jolliffe, I. T.: Principal Component Analysis. Springer-Verlag, New York (1986)
8. Lebart, L., Salem, A., Berry, L.: Exploring textual data. Kluwer, Dordrecht (1988)
9. Liu, L., Kang, J., Yu, J., Wang, Z.: A comparative study on unsupervised feature selection methods for text clustering. In: Proc. of the IEEE Int. Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE05), pp. 597-601. IEEE (2005)
10. Liu, T., Liu, S., Chen, Z., Ma, W.: An Evaluation on Feature Selection for Text Clustering. In: Proc. of the 20^{th} Int. Conference on Machine Learning (ICML03), pp. 488-495. ACM (2003)
11. Newman, M. E. J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E. **69**, 026113 (2004)
12. Roweis, S. T., Saul, L. K.: Nonlinear dimensionality reduction by locally linear embedding. Science. **290**, 2323–2326 (2000)
13. Salton, G., Buckely, C.: Term-weighting approaches in automatic text retrieval. Inform. Process. Manag. **24**, 513–523 (1988)
14. Salton, G., Wong, A., Yang, C. S.: A vector space model for automatic indexing. Commun. ACM. **18**, 613–620 (1975)
15. Scott, J.: Social Network Analysis: a handbook. Sage, London (2000)
16. Tenenbaum, J., de Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science. **290**, 2319–2323 (2000)
17. Wasserman, S., Faust, K.: Social network analysis. Cambridge University Press (1994)
18. Wilbur, W. J., Sirotkin, K.: The automatic identification of stop words. J. Inform. Sci. **18**, 45–55 (1992)
19. Yang, Y., Pedersen, J. O.: A comparative study on feature selection in text categorization. In: Proc. of the 14^{th} Int. Conference on Machine Learning (ICML97), pp. 412-420. ACM (1997)

Self Organizing Maps for distributional data

Mappe autorganizzanti per dati distribuzionali

Rosanna Verde and Antonio Irpino

Abstract We present Batch Self Organizing Map (BSOM) algorithms for data described by distributions. As unsupervised classification algorithms, BSOM depend on a suitable choice of a distance measure. Using the $L2$ Wasserstein distance for distributional data and its decomposition, we show how adaptive distances can be exploited in the learning process for describing the structure of the data. Adaptive distances induce a set of relevance weights on the descriptors of the data acting, then, as feature selection method. We present different types of adaptive distances based on different constraints and, using real data, we show the results of the proposed method.
Abstract *Nel presente lavoro proponiamo degli algoritmi di tipo Batch per l'estrazione di Mappe Autorganizzate (BSOM) per dati descritti da distribuzioni. Le BSOM sono da considerarsi degli algoritmi di classificazione non supervisionata basate su un'opportuna scelta di una distanza per confrontare i dati. Utilizzando la distanza $L2$ di Wasserstein per dati distribuzionali che, insieme ad una sua particolare decomposizione, mostriamo come le distanze adattative, nel processo di apprendimento della rete, possano essere sfruttate per una descrizione migliore della struttura dei dati. Le distanze adattative, infatti, inducono una serie di pesi per i descrittori dei dati, che ne valutano la rilevanza nel processo di classificazione. In questo modo, le distanze adattative possono fungere come metodo di selezione dei descrittori. Nel lavoro presentiamo differenti schemi tipi di ditanze adattative basati su diversi vincoli e, utilizzando dei dati reali, mostriamo i risultati del metodo proposto.*

Key words: Self Organizing Maps, Distributional data, Adaptive distances

Rosanna Verde
Dipartimento di Matematica e Fisica, Univesit degli Studi della Campania L. Vanvitelli, e-mail: rosanna.verde@unicampania.it

Antonio Irpino
Dipartimento di Matematica e Fisica, Univesit degli Studi della Campania L. Vanvitelli, e-mail: antonio.irpino@unicampania.it

Enviromental Processes, Human Activities and their Interactions

Estimation of coral growth parameters via Bayesian hierarchical non-linear models

Stima dei parametri di crescita di coralli tramite modelli bayesiani gerarchici non lineari

Crescenza Calculli, Barbara Cafarelli and Daniela Cocchi

Abstract In Ecology, the von Bertalanffy growth function (VBGF) is the standard model used to investigate the body growth of marine species. The parameters of this function are usually estimated by a classical method that might induce bias in the results: this method does not allow to distinguish the variability at individual or population level nor to take into account the contribution of environmental factors. A Bayesian hierarchical nonlinear model for estimating the VBGF parameters is proposed in order to overcome the limitations of the traditional method. The proposal improves both the statistical accuracy and the quantification of uncertainties affecting marine species growth. The proposal is assessed through a case study concerning two Mediterranean corals, *Balanophyllia europaea* and *Leptopsammia pruvoti*.

Riassunto *La funzione di von Bertalanffy (VBGF) è il modello standard usato in Ecologia per studiare la crescita degli individui nelle popolazioni marine. I parametri di questa funzione sono comunemente ottenuti tramite un metodo classico che può portare a errori delle stime, oltre a non permettere di modellare la variabilità disgiuntamente a livello di individuo o di popolazione, né di tenere in considerazione il contributo di fattori ambientali. Si propone un modello bayesiano gerarchico non lineare che permette di superare i limiti del metodo tradizionale e di ottenere un guadagno in termini di accuratezza statistica e di quantificazione delle componenti di incertezza che caratterizzano la crescita delle popolazioni marine. La proposta è valutata tramite un caso di studio riguardante due coralli mediterranei, Balanophyllia europaea e Leptopsammia pruvoti.*

Key words: Bayesian hierarchical models, Growth curves, Von Bertalanffy Growth Function

Crescenza Calculli
Dept of Economics and Finance, University of Bari, Italy, e-mail: crescenza.calculli@uniba.it

Barbara Cafarelli
Dept of Economics, University of Foggia, Italy e-mail: barbara.cafarelli@unifg.it

Daniela Cocchi
Dept of Statistical Sciences, University of Bologna, Italy e-mail: daniela.cocchi@unibo.it

1 Introduction

Marine biologists use individual demographic variables (age, oral disk length and body size) for modeling coral peculiarities, as well as their relationships with the environment. Individual coral age is usually obtained from coral body size using reliable growth models. The VBGF is a popular model for predicting the growth of marine organisms linking their lengths to ages by a non-linear relationship [4, 9]. For estimating the VBGF parameters, marine biologists use several methods that do not exploit statistical reasoning as they might. These methods are based on linear transformations of the VBGF, whose parameters are subsequently estimated by OLS. This approach has some limitations. Linear transformations are often proposed without accounting for measurement errors of observed data. Variability, at individual or population level, is neglected, inducing bias in parameter estimates and variance underestimation.

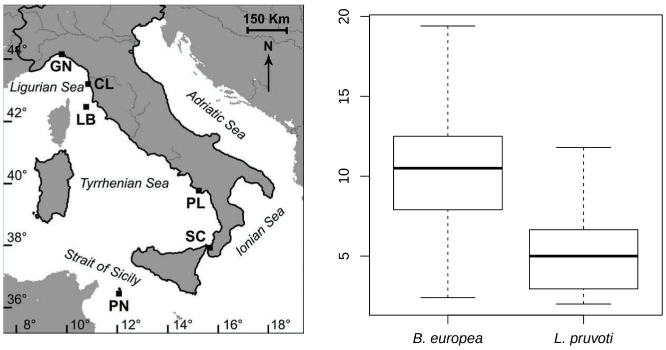
Environmental features at different sites are crucial in affecting coral characteristics. Unfortunately, these features are often unsuitable to this end, since they are measured as synthetic aggregates collected for other purposes: association of site-specific environmental measures to individual species data might be a forcing. The effort for considering environmental information, such as sea temperature, solar radiation, surface ocean acidification and anthropogenic stress, might be costly in terms of modeling and useless when interpreting results. For these reasons, a statistical model with random effects that avoid the explicit consideration of environmental covariates is suitable.

In particular, a feasible approach for estimation of coral growth parameters based on Bayesian hierarchical non-linear mixed effects models is proposed. Hierarchical modeling [2] is equivalent to handle VBGF in a two stage framework, accounting for two different sources of variability: the within-site and the between-site variations. The approach is applied to data of two solitary corals living in the Mediterranean sea, *Balanophyllia europaea* (*B. europaea*) and *Leptopsammia pruvoti* (*L. pruvoti*).

2 Motivating example

In this study, data on two species of solitary Mediterranean corals, *B. europaea* and *L. pruvoti*, are available. Both species live in rocky habitats and are widely distributed in the Mediterranean basin. Their growth and demographic peculiarities have been proved to be sensitive to changes in environmental conditions and, more generally, to the global warming [4, 9].

A dataset of 417 individuals combining 238 specimens of *B. europea* and 179 specimens of *L. pruvoti* is considered. In particular, specimens for the two species were collected during the same time interval (from 9th November 2003 to 30th September 2005) in the same sites as reported in Figure 1a [4, 8]. For each individual the main measurements available are the corallite length (L, in *mm*) which



(a) (b)

Fig. 1: (a) Site Locations: GN, Genova; CL, Calafuria; LB, Elba Isle; PL, Palinuro; SC, Scilla; PN, Pantelleria Isle. (b) Coral species length (in *mm*) distributions

represents the maximum axis of the oral disc and the ages in years obtained counting bands of the skeleton via computerized tomography scans (CTSs).

Table 1: Descriptive features the coral species and annual averages of environmental indicators per site. R: Solar Radiation (from 190 W/m^2); T: Sea Surface Temperature (from 18°C)

	GN		CL		LB		PL		SC		PN	
	B. europea	*L.pruvoti*	*B. europea*	*L.pruvoti*	*B. europea*	*L.pruvoti*	*B. europea*	*L.pruvoti*	*B. europea*	*L.pruvoti*	*B. europea*	*L.pruvoti*
n	42	30	34	29	34	30	54	30	32	30	42	30
Mean age (years)	7.4	6.2	5.5	4.1	4.6	8.1	6.9	5.6	6.2	7.8	5.2	7
Mean length (mm)	11.7	4.8	8.3	3.9	9	6.3	9.9	4.4	9.9	6.3	8.8	4.8
R (W/m^2), annual mean	166.95		170.07		172.74		181.48		187.31		192.95	
T (°C), annual mean	19.56		18.02		18.74		19.14		19.54		19.88	

Figure 1b shows differences in terms of length between the the two species. Summary features of the two species and the different environmental site conditions are reported in Table 1 where some differences among sites are appreciable. The measurement of environmental site conditions, even if in principle very stimulating, is currently performed under very general purposes, not directly linked to coral data collection. For this reason, they will not be explicitly considered in the following model.

3 Bayesian hierarchical VBGF modeling

Instead of the standard specification of the VBGF [5], an alternative parameterization [3] is followed to link the age and size of corals:

$$L(t) = L_\infty(1 - e^{-te^{c/L_\infty}}) \tag{1}$$

with

$$c = ln(k)L_\infty \tag{2}$$

where $L(t)$ is the individual length at age t, L_∞ is the asymptotic length and k is the rate at which growth approaches this asymptote [1]. Parameter c can be seen as the part of the length growth accountable for site-specific conditions. This parameterization allows to skip the interaction between coral asymptotic length and growth rate.

A suitable hierarchical model to link coral length and age is proposed among different alternatives [6]. For the i-th observation ($i = 1,\dots,n_j$) in the j-th site, the coral length is modeled as $L_i \sim N(\mu_i, \tau^2)$, where τ^2 is the precision, with

$$\mu_i = L_{\infty_{h,j}}(1 - e^{-t_i e^{c_j/L_{\infty_{h,j}}}}) \qquad h = 1,\dots,H \quad \text{and} \quad j = 1,\dots,J \tag{3}$$

where H is the number of species, J is the number of sites and n_{jh} is the abundance of the h-th species in the j-th site. Since relevant environmental covariates are measured with different timing and reliability with respect to the coral dataset, the second stage of the hierarchy accounts for possible differences in the asymptotic length between sites by random effects, as follows

$$L_{\infty_{h,j}} \sim TN_{(0.1,\infty)}(0, 0.1). \tag{4}$$

The effects expressed in (4) allow different environmental site-specific features together with differences between the two species. Being L_∞ strictly positive, a left truncated normal distribution with large precision has been chosen [10] as the prior distribution for this parameter, while for the c parameter in (2) a vague prior distribution has been chosen. This parameter is assumed to capture the differential growth of species between sites and is specified as

$$c_j \sim N(0.001, 1.00e^{-3}). \tag{5}$$

The model implicitly includes environmental covariates such as the solar radiation, the sea surface temperature or the marine current through the c parameter.
Finally, to complete the Bayesian specification, $\tau \sim U(0, 10)$ has been chosen. Models are implemented by means of the JAGS software [11] via the `R2jags` package of `R` [12] providing a powerful and flexible tool for analyzing grouped data, such as repeated measures data and nested designs.

4 Results

The joint posterior distribution of model parameters is obtained using 72,000 iterations with 3 chains, discarding the first 12,000 of the burn-in phase of the algorithm. Chains were checked for convergence and reasonable mixing by graphical inspection of the trace plots and by means of the Gelman-Rubin convergence diagnostics [7]. Posterior distributions of parameters are summarized in Figure 2. To ease the interpretability of results, distributions are reported distinguishing among sites. Each site panel contains graphical representations of the posterior distributions for the random effect and the asympotic length for the two species.

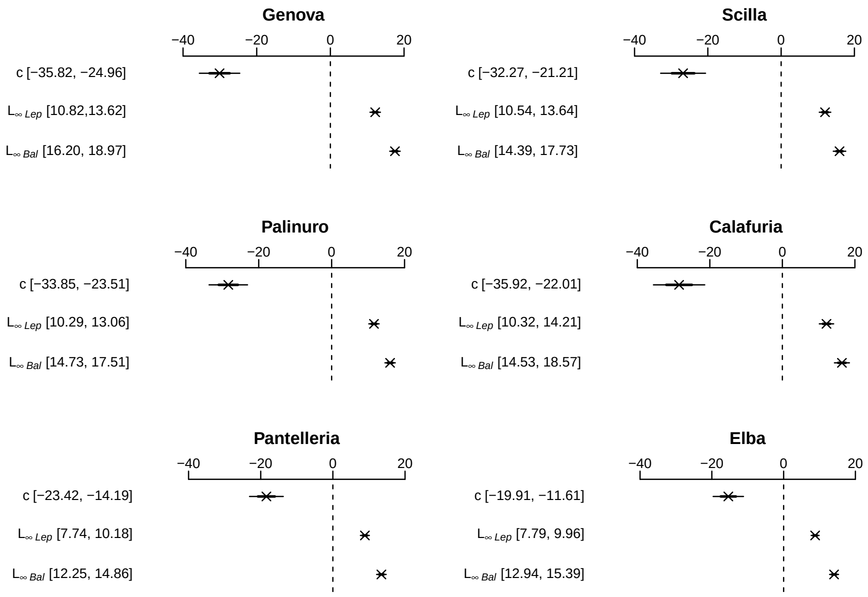


Fig. 2: Syntheses of posterior distributions of VBGF parameters. Crosses represent the means of the marginal posterior distributions; horizontal-bars represent 50% (thicker ones) and 95% (lighter ones) CIs (1 and 2 sd around the mean, respectively). For each parameter, 95% Credibility Intervals extremes are reported in brackets

The combination of species and locations is relevant in affecting L_∞ values. Results suggest a higher overall asymptotic length for *B. europea* than for *L. pruvoti*, confirming the well-known differences between the two species (Figure 1b). Posterior distributions of *c* parameters allow to distinguish among different sites, summarizing the relevant geographical effects on the growth rates of both species. The negative values of *c* are due to the fact that growth rates in (2) are less than 1 for this dataset. Sites that exhibit negative posterior c values closer to 0, tend to have lower L_∞ values with more concentrated posterior distributions. This result witnesses that very different environmental conditions influence coral growth at each site, which however translate into a direct relationship between a favorable situation and growth rate.

5 Conclusions

The examined case study provides a flexible framework that allows biologically meaningful estimates of the VBGF parameters. In particular, the use of Bayesian hierarchical models enables a better evaluation of uncertainty typical of length-at-age data and an innovative description of the relations between species growth and environmental conditions which can be represented by priors in a hierarchical model. The proposed approach demonstrates how to address ecological problems where environmental information is often limited, sparsely sampled or does not suit with collected data. By considering individual and population variation, the growth of two coral species has been estimated accurately and precisely, improving the understanding of their biological characteristics. Furthermore, this approach might be conveniently extended to multi-species models allowing tools to analyze entire communities.

Acknowledgements This work was supported by the PRIN2015 project EphaStat - *Environmental processes and human activities: capturing their interactions via statistical methods*, funded by the Italian Ministry of University and Research (MIUR).

References

1. Basso, N.G., Kehr, A.I.: Postmetamorphic growth and population structure ofthe frog Leptodactylus latinasus (anura: Leptodactylidae). Stud. Neotrop. FaunaEnviron **26**, 39–44. (1991)
2. Cadigan, N. G., Campana, Steven E.: Hierarchical model-based estimation of population growth curves for redfish (*Sebastes mentella* and *Sebastes fasciatus*) off the Eastern coast of Canada. ICES Journal of Marine Science **74** (3), 687 – 697 (2017)
3. Cafarelli, B., Calculli, C., Cocchi, D., Pignotti, E.: Hierarchical non-linear mixed-effects models for estimating growth parameters of western Mediterranean solitary coral populations. Ecological Modelling **346**, 1–9 (2017)
4. Caroselli, E., Zaccanti, F., Mattioli, G., Falini, G., Levy, O., Dubinsky, Z., Goffredo, S.: Growth and demography of the solitary scleractinian coral Leptopsammia pruvoti along a sea surface temperature gradient in the Mediterranean Sea. PLoS ONE **7**(6) , e37848 (2012)
5. Fabens, A.J.: Properties and fitting of the von Bertalanffy growth curve. Growth **29**, 265–289 (1965)
6. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, NY (2007)
7. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B. , Vehtari, A., Rubin, D. B.: Bayesian Data Analysis, 3nd edition. Chapman & Hall/CRC, Boca Raton (2013)

8. Goffredo, S., Caroselli, E., Mattioli, G., Pignotti, E., Zaccanti, F.: Relationships between growth, population structure and sea surface temperature in the temperate solitary coral Balanophyllia europaea (scleractinia, dendrophylliidae). Coral Reefs **27**, 623–632 (2008)
9. Goffredo, S., E. Caroselli, E.Pignotti, G. Mattioli, F. Zaccanti: Variation in biometry and demography of solitary corals with environmental factors in the Mediterranean Sea. Mar. Biol. , **152**, 351–361 (2007)
10. Quintero, F. O. L., Contreras-Reyes, J. E., Wiff, R., Arellano-Valle, R. B.: Flexible Bayesian analysis of the von Bertalanffy growth function with the use of a log-skew-*t* distribution. Fish. Bull., **115**(1), 13–26 (2017)
11. Plummer, M.: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria. 124 (2003)
12. Su, Y., Yajima, M.: R2jags: Using R to Run ”JAGS”. R package version 0.5-7.(2015) https://CRAN.R-project.org/package=R2jags

A Hierarchical Bayesian Spatio-Temporal Model to Estimate the Short-term Effects of Air Pollution on Human Health

Un modello bayesiano spazio-temporale gerarchico per stimare gli effetti di breve periodo dell'inquinamento atmosferico sulla salute umana

Fontanella Lara, Ippoliti Luigi and Valentini Pasquale

Abstract We introduce a hierarchical spatio-temporal regression model to study the spatial and temporal association existing between health data and air pollution. The model is developed for handling measurements belonging to the exponential family of distributions and allows the spatial and temporal components to be modelled conditionally independently via random variables for the (canonical) transformation of the measurements mean function. A temporal autoregressive convolution with spatially correlated and temporally white innovations is used to model the pollution data. This modelling strategy allows to predict pollution exposure for each district and afterwards these predictions are linked with the health outcomes through a spatial dynamic regression model.

Abstract *In questo lavoro viene introdotto un modello di regressione spazio-temporale gerarchico per studiare l'associazione spazio-temporale tra i dati sulla salute e l'inquinamento atmosferico. Il modello è stato sviluppato nell'ambito della famiglia esponenziale e consente di modellare condizionatamente le componenti spaziali e temporali in modo indipendente attraverso variabili casuali per la trasformazione (canonica) della funzione media. Per modellare i dati sull'inquinamento viene utilizzata una convoluzione temporale autoregressiva con innovazioni spazialmente correlate e temporalmente indipendenti. Questa strategia di modellizzazione consente di prevedere l'esposizione all'inquinamento per ciascun distretto e, successivamente, queste previsioni sono messe in relazione con le ospedalizzazioni attraverso un modello di regressione spaziale dinamica.*

Key words: Hierarchical model, spatio-temporal model, MCMC

Fontanella Lara
DSGS, University ”G.d'Annunzio” of Chieti-Pescara, e-mail: lara.fontanella@unich.it

Ippoliti Luigi
DeC, University ”G.d'Annunzio” of Chieti-Pescara e-mail: luigi.ippoliti@unich.it

Valentini Pasquale
DeC, University ”G.d'Annunzio” of Chieti-Pescara e-mail: pasquale.valentini@unich.it

1 Introduction

In the last 30 years, a large number of studies have provided substantial statistical evidence of the adverse health effects associated with air pollution. The statistical literature on health care research is very rich and includes a plethora of models referring to different types of study designs. Most of those studies are usually based on time series models, developed both in single and multisites frameworks (e.g., [1]). However, because air pollution concentrations vary at fine spatio-temporal scales, quantifying the impact of air pollution appeares more as an inherently spatio-temporal problem. Also, despite the availability of large data sets for multiple pollutants, only a few studies consider the joint effects of numerous air pollutants simultaneously [2]. In this paper, we thus propose a hierarchical spatio-temporal regression model, which is able to cope with different spatial resolutions in order to change the support of air pollution data (regressors) to achieve alignment with the health outcome measured at area level.
Considering the air pollution data, we opt for a modelling approach based on assuming the existence of a latent Gaussian variable which may be interpreted as a potential pollution harmful to health in the short period. Then at the process level, we use a dynamic model proposed by [3] which takes into account for spatio- temporal variation using a temporal autoregressive variable with spatially correlated innovations. It is assumed that these innovations follow a Gaussian process with an exponential covariance function. Given such model, we can interpolate the process at unobserved times and/or locations and face the change of support problem (COSP). Afterwards, assessing the effect of air pollution on human health is possible through a regression model that includes lagged exposure variables as covariates.

2 Model Specification

Assume that Y and X are two multivariate spatio-temporal processes observed at temporal instants $t = 1, 2, \ldots, T$ and generic locations, $\mathbf{s} \in \mathscr{D}_y$ and $\mathbf{u} \in \mathscr{D}_x$, respectively. Assume also that X is a predictor of Y, which thus represents the process of interest. For the two different processes, the spatial sites $\mathbf{s}$ and $\mathbf{u}$ can denote the same location but, in general, they need not be the same. Furthermore, both $\mathscr{D}_y$ and $\mathscr{D}_x$ may represent different spatial characteristics and structures. Usually, health data (Y) are collected over time in a fixed study region,$\mathscr{D}_y$, typically in the form of mortality and morbidity counts or hospital admissions, coded according to the type of disease (e.g. cardiovascular, acute respiratory, etc). While pollution concentrations (X) are measured at specific points in time and at a number of monitoring sites across a continuous region $\mathscr{D}_x$ and usually come in the form of geostatistical data.

Let n_y be the number of observed variables for Y and n_x the number of observed variables for X. The most informative case is represented by the isotopic configuration where, for each multivariate process, Y or X, all variables are measured at all their respective sites. In this case, let $\mathbf{Y}(\mathbf{s},t) = [Y_1(\mathbf{s},t), \ldots, Y_{n_y}(\mathbf{s},t)]'$

be the vector of the n_y values of Y at site $\mathbf{s}$ and time t. Equivalently, we write $\mathbf{X}(\mathbf{u},t) = [X_1(\mathbf{u},t),\ldots,X_{n_x}(\mathbf{u},t)]'$ for the vector of the n_x values of X at site $\mathbf{u}$ and time t. The opposite case is the completely heterotopic case where not all the variables can be observed at the same site $-$ this is especially true for X in our study. Without loss of generality, for the sake of simplicity, here we use the notation for the isotopic case. Accordingly, the n_y variables of Y are observed at the same sites $\mathbf{s}_i,\ i = 1,\ldots,N_y$ and the n_x variables of X are observed at sites $\mathbf{u}_r,\ r = 1,\ldots,N_x$.

Let $\tilde{n}_y = n_y N_y$ and $\tilde{n}_x = n_x N_x$. At a specific time t, by using a site ordering, the $(\tilde{n}_y \times 1)$ and $(\tilde{n}_x \times 1)$ dimensional spatial processes are denoted as $\mathbf{Y}(t) = [\mathbf{Y}(\mathbf{s}_1,t)',\ldots,\mathbf{Y}(\mathbf{s}_{N_y},t)']'$ and $\mathbf{X}(t) = [\mathbf{X}(\mathbf{u}_1,t)',\ldots,\mathbf{X}(\mathbf{u}_{N_x},t)']'$. However, the data may also be ordered by variable. In this case, we write $\mathbf{Y}(t) = [\mathbf{Y}_1(t)',\ldots,\mathbf{Y}_{n_y}(t)']'$ and $\mathbf{X}(t) = [\mathbf{X}_1(t)',\ldots,\mathbf{X}_{n_x}(t)']'$, where $\mathbf{Y}_k(t)$ is the the vector of n_y observations for variable Y_k, and $\mathbf{X}_j(t)$ is the the vector of n_x observations for variable X_j.

The model is based on the measurement equations for the conditionally independent variables,

$$Y_k(\mathbf{s},t)|\eta_{y_k}(\mathbf{s},t),\sigma^2_{y_k} \overset{ind}{\sim} F_y(\eta_{y_k}(\mathbf{s},t),\sigma^2_{y_k}), \quad k = 1,\ldots,n_y$$

$$X_j(\mathbf{u},t)|\eta_{x_j}(\mathbf{u},t),\sigma^2_{x_j} \overset{ind}{\sim} F_x(\eta_{x_j}(\mathbf{u},t),\sigma^2_{x_j}), \quad j = 1,\ldots,n_x,$$

where $\sigma^2_{y_k}$ and $\sigma^2_{x_k}$ are dispersion parameters. In general, the distributions F_y and F_x are allowed to be from any exponential family distribution. By choosing appropriate canonical link functions, the specification of the measurement equations are completed with the specification of the following linear predictors

$$g_y\left[\eta_{y_k}(\mathbf{s},t)\right] = \mu_{y_k}(\mathbf{s},t) + \phi_{y_k}(\mathbf{s},t) \tag{1}$$

$$g_x\left[\eta_{x_j}(\mathbf{u},t)\right] = \mu_{x_j}(\mathbf{u},t) + \phi_{x_j}(\mathbf{u},t) \tag{2}$$

where $\mu_{y_k}(\mathbf{s},t)$ and $\mu_{x_j}(\mathbf{u},t)$ are fixed effect terms representing the large-scale spatio-temporal variability of the processes, and $\phi_{y_k}(\mathbf{s},t)$ and $\phi_{x_j}(\mathbf{u},t)$, are random effects introduced to capture any residual spatio-temporal autocorrelation.

The random effects are modelled through the following equations

$$\phi_{x_j}(\mathbf{u},t) = \int_{\mathscr{D}_x} \kappa_{\theta_{x_j}}(\mathbf{u}-\mathbf{u}')\phi_{x_j}(\mathbf{u}',t-1)d\mathbf{u}' + \mathbf{v}_{x_j}(\mathbf{u},t) \tag{3}$$

$$\phi_{y_k}(\mathbf{s},t) = \sum_{j=1}^{n_x}\sum_{l=0}^{L} \beta_{k,j,l}\ \phi_{x_j}(\mathbf{s},t-l) + \mathbf{v}_{y_k}(\mathbf{s},t) \tag{4}$$

where $\kappa_{\theta_{x_j}}(\mathbf{u}-\mathbf{u}') = \rho_{1,x_j}\exp\Big((\mathbf{u}-\mathbf{u}')'\,\Sigma^{-1}_{x_j}\,(\mathbf{u}-\mathbf{u}')\Big)$,
$\Sigma^{-1}_{x_j} = \frac{1}{\rho^2_{2,x_j}}\begin{bmatrix}\cos(\alpha_{x_j}) & \sin(\alpha_{x_j})\\ -d_{x_j}\sin(\alpha_{x_j}) & d_{x_j}\cos(\alpha_{x_j})\end{bmatrix}$, $\alpha_{x_j}\in[0,\frac{\pi}{2}]$, $d_{x_j}>0$, $\theta_{x_j} = \left\{\rho_{1,x_j},\rho_{2,x_j},c_{x_j},\alpha_{x_j}\right\}$, $\beta_{k,j,l}$ is the distributed lag coefficient which relates the jth pollutants at lag l to the kth disease health outcome, $\mathbf{v}_{y_k}(\mathbf{s},t)$ and $\mathbf{v}_{x_j}(\mathbf{u},t)$ Gaussian innovations that are

white in time and correlated in space.
It is worth noting that $\phi_{x_j}(\mathbf{s},t-l)$, in equation (4), is the structured variation in time at space resolution $\mathbf{s}$, $\phi_{x_j}(\mathbf{s},t) = \int_{\mathbf{s}} \phi_{x_j}(\mathbf{u},t)d\mathbf{u}$. In practice one could first define a regular grid, then interpolate the non-observed grid points, and approximate the integral by a Riemann sum. Since the regular grid usually becomes very large, this is computationally expensive, other strategies to approximate the integral can be found in [3].
Model completion requires specific forms for $\mu_y(t)$ and $\mu_x(t)$. The simplest specification of the mean components assumes the form of a linear regression function to take care of the effects of confounders, i.e. $\mu_{x_l}(\mathbf{u},t) = \sum_{i=1}^{c} \sum_{l=0}^{g} \delta_{x_j,il}(\mathbf{u})\, z_i(\mathbf{u},t-l)$, and $\mu_{y_k}(\mathbf{s},t) = \sum_{i=1}^{c} \sum_{l=0}^{g} \delta_{y_k,ij}(\mathbf{s})\, z_i(\mathbf{s},t-l)$, where $z_i(\cdot,t),\ i = 1,\ldots,c$ are observed covariates or components representing seasonal and long-term trends introduced to take care of the effects of unmeasured confounders (see [1] and [4]). Note that the $z_i(\cdot,t)$ could also be smoothed versions of measured confounders represented by natural cubic splines with specified degree of freedom.
The hierarchy of our model is completed by specifying the prior distributions of all hyperparameters. Noninformative conjugate priors are assumed with the expection of the spatial correlation parameters. This model is developed within a state-space framework and full probabilistic inference for the parameters is facilitated by a Markov chain Monte Carlo (MCMC) scheme for multivariate dynamic systems.
The proposed model has an intuitive appeal and enjoys several advantages. For example, it describes the spatial-temporal variability of the disease risk and explicitly defines a non-separable spatio-temporal covariance structure of the process. Also, it allows to study how the disease risk at a specific areal unit reacts over time to exogenous impulses from the same or different areal units. Finally, several general structures that make use of different covariate information, can be easily accommodated in the different levels of the hierarchy.
Fitting our model using the MCMC algorithm is computationally intensive. However, once the statistical model is fitted and assuming that the posterior of the parameters for (3) does not change (see [3] for more details), predictions are computationally a lot cheaper.

3 Application

We illustrate our modelling approach to measure the effect of exposure variables on hospital admissions observed in Lombardia and Piemonte regions (Italy) in 2011. In particular, health data consist of counts of daily hospital admissions for cardiovascular diseases and respiratory diseases. Pollution data refer to daily-average concentration levels of CO, NO_2, PM_{10} and O_3.
To provide more insight on the way in which the disease risks spread out to surrounding districts, Figure 1 below shows maps of raw standardized morbidity ratios obtained by averaging the SMR values across time. The map on the left, shows that, on average, the highest risk areas associated with the cardiovascular diseases

correspond to districts in the Southeastern parts of Lombardia. The map for the respiratory diseases (right) also supports the idea that Lombardia is the most at risk with the highest SMR values observable at the Northern and the Southeastern parts of Milan. In general, the SMR maps show evidence of localised spatial clusters.
The MCMC algorithm was run for $35,000$ iterations. Posterior inference was based on the last $30,000$. Convergence was monitored by inspecting trace plots. Preliminary results show a positive association between air pollutants and hospital admissions. In particular, the peak response of hospital admissions for cardio-respiratory diseases after a positive shock on pollutants occurs after three days and then gradually decreases and dies out in about six days.

Acknowledgements This work is developed under the PRIN2015 supported project 'Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)' [grant number 20154X8K23] funded by MIUR (Italian Ministry of Education, University and Scientific Research).

References

1. Peng, R.D. and Dominici, F. and Louis, T.A.: Model choice in time series studies of air pollution and mortality. J. R. Stat. Soc. Ser. A. Stat. Soc., **169**, 179–203 (2006)
2. Rushworth, A. and Lee, D. and Mitchell, R.: A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. Spat. Spatiotemporal. Epidemiol. **10**, 29–38 (2014)
3. Sigrist, F., Knsch, H.R., Stahel, W.A.: A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. Ann. Appl. Stat.,**6**, 1452–1477 (2012)
4. Shaddick, G. and Zidek, J.: Spatio-Temporal Methods in Environmental Epidemiology, Chapman Hall/CRC (2015)

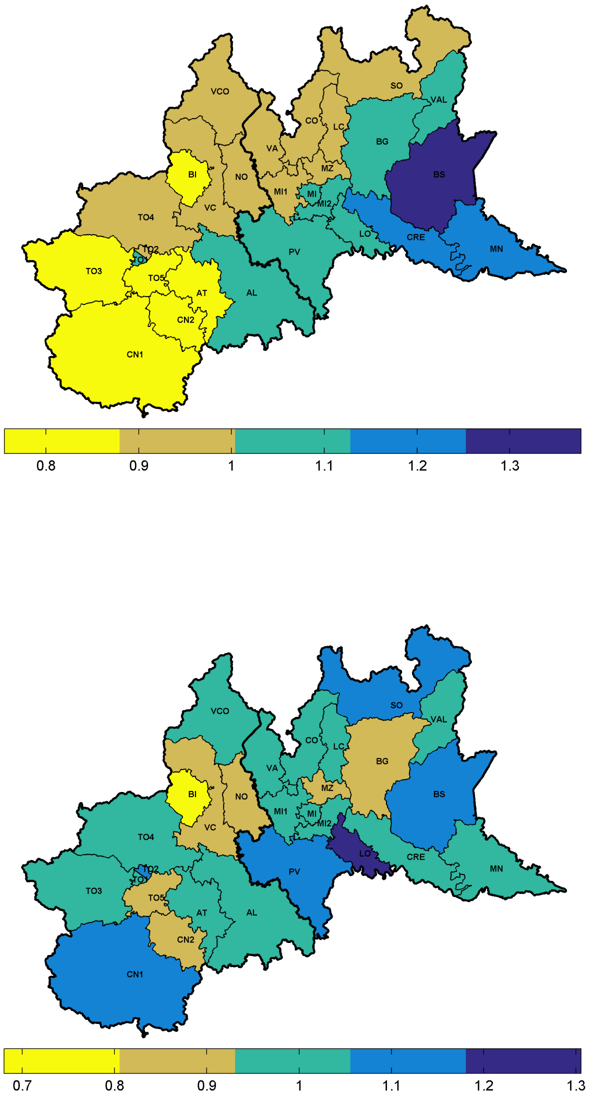


Fig. 1 *Map of the standardised morbidity ratio (SMR) for hospital admissions due to cardiovascular (left) and respiratory (right) diseases in 2011. The lable within each district represents the acronymous of the ASL.*

A multilevel hidden Markov model for space-time cylindrical data

Un modello multilivello a classi markoviane latenti per dati cilindrici spazio-temporali

Francesco Lagona and Monia Ranalli

Abstract Motivated by segmentation issues in marine studies, a novel hidden Markov model is proposed for the analysis of cylindrical space-time series, that is, bivariate space-time series of intensities and angles. The model is a multilevel mixture of cylindrical densities, where the parameters of the mixture vary at the spatial level according to a latent Markov random field, while the parameters of the hidden Markov random field evolve at the temporal level according to the states of a hidden Markov chain. Due to the numerical intractability of the likelihood function, parameters are estimated by a computationally efficient EM algorithm based on the specification of a weighted composite likelihood. The proposal is tested in a case study that involves speeds and directions of marine currents in the Gulf of Naples.

Riassunto *Motivati da problemi di classificazione in studi marini, proponiamo un nuovo modello a classi markoviane latenti per l'analisi di serie cilindriche spazio-temporali, ovvero serie spazio-temporali bivariate di intensità ed angoli. Il modello è una mistura multi-livello di densità cilindriche, con parametri che variano al livello spaziale secondo un campo markoviano latente, i cui parametri variano nel tempo secondo una catena markoviana latente. A causa dell'intrattabilità numerica della funzione di verosimiglianza, i parametri sono stimati da un algoritmo EM basato sulla definizione di una funzione di pseudo-verosimiglianza composita. Il modello è applicato all'analisi di una serie spazio-temporale contenente le velocità e le direzioni delle correnti marine del golfo di Napoli.*

Key words: Cylindrical data, hidden Markov model, EM algorithm, Composite likelihood

Francesco Lagona
Department of Political Sciences, Roma Tre University, e-mail: francesco.lagona@uniroma3.it

Monia Ranalli
Department of Political Sciences, Roma Tre University e-mail: monia.ranalli@uniroma3.it

1 Introduction

A detailed knowledge of coastal currents is crucial for a valid integrated coastal zone management. Among the different available ocean observing technologies, high-frequency radars (HFRs) have unique characteristics, that make them play a key role in coastal observatories. HFR data can be conveniently described as space-time bivariate arrays of angles and intensities that respectively indicate the directions and the speeds of sea currents across space and over time. Data with a mixed circular-linear support are often referred to as *cylindrical* data [1], because the pair of an angle and an intensity can be represented as a point on a cylinder.

The statistical analysis of cylindrical space-time series is complicated by the unconventional topology of the cylinder and by the difficulties in modeling the cross-correlations between angular and linear measurements across space and over time. Additional complications arise from the skewness and the multimodality of the marginal distributions of the data. As a result, specific methods for the analysis of space-time cylindrical data have been relatively unexplored. Proposals in this context are limited to geostastical models, where cylindrical data are assumed conditionally independent given a latent process that varies continuously across space and time [15]. Geostatistical models give good results in open sea areas, where waves and currents can move freely without obstacles. Sea motion in coastal areas provides, however, a different setting. Coastal currents are shaped and constrained by the orography of the site. As a result, coastal circulation is much more irregular than ocean-type patterns and it is inaccurately represented by traditional geostatistical models, which do not incorporate orographic information. The development of a physical model that well represents sea motion in coastal areas can be a formidable task if the orography of the site is irregular. A more practical approach relies on decomposing an observed circulation pattern into a small number of local regimes whose interpretation is easier than the global pattern.

To accomplish this goal, we propose a model that segments coastal data according to finitely many latent classes that vary across space and time and are associated with the distribution of the data under specific, space-time varying, environmental conditions. Specifically, we assume that the joint distribution of the data is well approximated by a multi-level mixture of cylindrical densities. At each time, the parameters of the mixture vary according to a latent Markov field, whose parameters evolve over time according to a latent Markov chain. The idea of using hidden Markov models to segment cylindrical data is not completely novel. Hidden Markov models have been proposed for segmenting cylindrical time series [6] and hidden Markov fields have been proposed to segment spatial cylindrical data [11]. Our proposal integrates these specifications in a space-time setting.

A potential disadvantage of the model is the intractability of the likelihood function. We address estimation issues by relying on composite likelihood (CL) methods [14, 7]. This estimation strategy, on one hand, provides feasible and fast estimation methods. On the other hand, some dependence among observations is lost, resulting in a loss of statistical efficiency. However, consistency of the CL estimators still holds under regularity conditions [9]. Under these conditions, furthermore, CLEs are asymptotically normal with covariance matrix given by the inverse of a sandwich matrix, known as Godambe information [4] rather than the usual Fisher information matrix for maximum likelihood estimators (MLEs). CL methods have been successfully applied in spatial and space-time settings [11, 3].

2 Marine currents in the Gulf of Naples

The Gulf of Naples is a semienclosed marginal basin of the central Tyrrhenian Sea. It is a coastal area characterized by striking environmental contrasts: one of the most intensely urbanized coastlines in the whole Mediterranean, with massive industrial settlements, the very polluted Sarno river mouth, a number of distributed sewage outlets, coexisting with the extremely scenic coastal landscapes of the Sorrento Peninsula, of the Islands of Capri, Procida and Ischia and with unique underwater archaeological treasures (e.g. Baiae and Gaiola). For this reason, the Gulf of Naples has been subject to intense monitoring of its meteorological and oceanographic conditions. In particular, starting in 2004 an HFR system has been installed along its coastline, consisting first of two, and from 2008 of three, transceiving antennas operating at 25 MHz, providing hourly data of the surface current field at 1-km2 horizontal resolution. Such a system has shed light on very rich, multiple-scale surface dynamics and on the mechanisms driving water renewal of individual subbasins of the gulf [8, 13, 2]. Moreover, these data have been exploited in numerical models to enhance their predictive skills through state of the art assimilation schemes [5]. The functioning principle of HFRs is based on resonant backscatter, resulting from coherent reflection of a transmitted electromagnetic wave by ocean surface waves whose wavelength is half of the transmitted electromagnetic wavelength. As a result, every station can provide only the radial component of the surface currents with respect to the antenna location. Two, at least, or even better more stations (to ensure better statistics, to minimize gaps due to physical obstacles or to electromagnetic disturbances, to lower geometric dilution of precision) are needed to combine the radial information to obtain a current vector field. A vector map (or field) decomposes the current's field into the u- and v-components (Cartesian representation) of the sea surface at each observation point of a spatial lattice, where u corresponds to the west–east and v to the south–north current component. Joint modelling of u and v is, however, typically complicated by cross-correlations that vary dramatically in different parts of the spatial domain [12]. We therefore model sea current fields by using polar coordinates. Specifically, the observed current field is represented as a cylindrical spatial series, obtained by computing for each observation site the speed $y = \sqrt{u^2 + v^2} \in [0, +\infty)$ of the current (meters per second) and the direction $x = \tan2^{-1}(u, v) \in [0, 2\pi)$ of the current (radians), where $\tan2^{-1}$ is the inverse tangent function with two arguments and x follows the geographical convention, clockwise from North (0) to East ($\pi/2$). The data that motivated this paper include current speed and direction across a grid of 300 points, observed every hour during March 2009 in the Gulf of Naples.

3 A cylindrical space-time hidden Markov model

The data that motivated this work are in the form of an $n \times T$ array of cylindrical data, say $(\mathbf{z}_{it}, i = 1 \ldots n, t = 1 \ldots T)$, where $\mathbf{z}_{it} = (x_{it}, y_{it})$ is a pair of an angle $x_{it} \in [0, 2\pi)$ and an intensity $y_{it} \in [0, +\infty)$, observed at time t and in the spatial site i. We assume that the temporal evolution of these data is driven by a multinomial process in discrete time $\boldsymbol{\xi} = (\boldsymbol{\xi}_t, t = 1 \ldots T)$, where $\boldsymbol{\xi}_t = (\xi_{t1}, \ldots, \xi_{tK})$ is a multinomial random variable with K classes. We specifically assume that such process is distributed as a Markov chain, whose distribution, say $p(\boldsymbol{\xi}; \boldsymbol{\pi})$, is known up to a vector of parameters $\boldsymbol{\pi}$ that includes the initial probabilities and the transition probabilities of the chain. Conditionally on the value assumed each time by the Markov chain, the spatial distribution of the data at time t depends on a multinomial process in discrete space $\mathbf{u}_t = (\mathbf{u}_{it}, i = 1 \ldots n)$, where $\mathbf{u}_{it} = (u_{it1}, \ldots, u_{itG})$ is a multinomial variable with G classes. We assume that such spatial process is distributed as a G-parameter Potts model, whose parameters depend on the value taken at time t by the latent Markov chain $p(\boldsymbol{\xi}; \boldsymbol{\pi})$. This model depends on $G-1$ sufficient statistics

$$n_g(\boldsymbol{u}_t) = \sum_{i=1}^{n} u_{itg}, \quad g = 1 \ldots G-1,$$

that indicate the frequencies of each latent class across the study area, and one sufficient statistic

$$n(\boldsymbol{u}_t) = \sum_{i=1}^{n} \sum_{j>i: j \in N(i)} \sum_{g=1}^{G_{t-1}} u_{itg} u_{jtg},$$

which indicates the frequency of neighboring sites which share the same class (for each i, $N(i)$ indicates the sets of neighboring sites of i). Precisely, we assume that the joint distribution of a sample $\mathbf{u}_t$, conditionally on $\boldsymbol{\xi}_t$, is known up to an array of class-specific parameters $\boldsymbol{\alpha} = (\alpha_{gk}, g = 1 \ldots G-1, k = 1 \ldots K)$ and a vector of auto-correlation parameters $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_k)$, and given by

$$p(\boldsymbol{u}_t \mid \boldsymbol{\xi}_t; \boldsymbol{\alpha}, \boldsymbol{\rho}) = \frac{\exp\left(\sum_{g=1}^{G_{t-1}} n_g(\boldsymbol{u}_t)\alpha_{gt} + n(\boldsymbol{u}_t)\rho_t\right)}{W(\boldsymbol{\alpha}, \boldsymbol{\rho})}, \tag{1}$$

where

$$\alpha_{gt} = \sum_{k=1}^{K} \xi_{tk} \alpha_{gk}$$

and

$$\rho_t = \sum_{k=1}^{K} \xi_{tk} \rho_k.$$

Our proposal is completed by assuming that, conditionally on the values taken by the Markov chain and the Potts model, the observed cylindrical data are independently distributed according to cylindrical densities, known up to a vector of parameters that depends on the latent class taken by the latent Markov random field at time t in site i. Precisely, we assume that

$$f(\mathbf{z} \mid \boldsymbol{\xi}, \mathbf{u}) = \prod_{i=1}^{n} \prod_{t=1}^{T} f(\mathbf{z}_{it}; \boldsymbol{\theta}_{itg}),$$

where

$$\boldsymbol{\theta}_{itg} = \sum_{g=1}^{G} u_{itg} \boldsymbol{\theta}_g,$$

and $\boldsymbol{\theta}_g$ is the gth entry of a vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G)$. Under this setting, we follow [1] and exploit the following parametric cylindrical distribution, namely

$$f(\mathbf{z}; \boldsymbol{\theta}) = \frac{\alpha \beta^{\alpha}}{2\pi \cosh(\kappa)} (1 + \lambda \sin(x - \mu)) y^{\alpha - 1} \exp(-(\beta y)^{\alpha} (1 - \tanh(\kappa) \cos(x - \mu))), \tag{2}$$

known up to five parameters $\boldsymbol{\theta} = (\alpha, \beta, \kappa, \lambda, \mu)$, where $\alpha > 0$ is a shape parameter, $\beta > 0$ is a scale parameter, $\mu \in [0, 2\pi)$ is a circular location parameter, $\kappa > 0$ is a circular concentration parameter, while $\lambda \in [-1, 1]$ is a circular skewness parameter.

The joint distribution of the observed and the latent variables is therefore given by

$$f(\mathbf{z}, \mathbf{u}, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}) = f(\mathbf{z} \mid \boldsymbol{u}; \boldsymbol{\theta}) p(\mathbf{u}; \boldsymbol{\rho}, \boldsymbol{\alpha}) p(\boldsymbol{\xi}; \boldsymbol{\pi}). \tag{3}$$

By integrating this distribution with respect to the unobserved variables, we obtain the likelihood function of the unknown parameters

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}; \mathbf{z}) = \sum_{\boldsymbol{\xi}} \sum_{\boldsymbol{u}} f(\mathbf{z}, \boldsymbol{u}, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}). \tag{4}$$

The maximization of the corresponding complete log-likelihood through an EM algorithm is unfeasible. As a result, we propose to estimate the parameters by maximizing a surrogate function, namely a composite log-likelihood function. Our proposal is based on the specification of a cover $\mathbb{A}$ of the set $S = \{1 \dots n\}$ of the observation sites, i.e. a family of (not necessarily disjoint) subsets $A \subseteq S$ such that $\cup_{A \in \mathbb{A}} = S$. For each subset A, we respectively define $\mathbf{z}_A = (\mathbf{z}_{it}, i \in A, t = 1 \dots T)$, $\mathbf{u}_A = (\mathbf{u}_{it}, i \in A, t = 1 \dots T)$, and

$$L^A(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}; \mathbf{z}_A) = \sum_{\boldsymbol{\xi}} \sum_{\boldsymbol{u}_A} f(\mathbf{z}_A, \boldsymbol{u}_A, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}) \tag{5}$$

as the contribution of the data in A to the composite likelihood (CL), where CL=$\prod_{A \in \mathbb{A}} L^A$. This composite likelihood function involves summations over all the possible values that $\boldsymbol{u}_A$ can take. As a result, the numerical tractability of these steps dramatically decreases with the cardinality of the largest subset of the cover $\mathbb{A}$. On the one side, this would suggest to choose a cover with many small subsets. On the other side, a cover that includes a few large subsets is expected to provide a CL function that is a better approximation of the likelihood function. Because summations over $\boldsymbol{u}_A$ become cumbersome for $\mid A \mid \geq 3$, a natural strategy is a cover that includes subsets with 2 elements. When $\mathbb{A}$ include all the subsets of two elements, then composite likelihood reduces to the pairwise likelihood function [14]. In a spatial setting, a pairwise likelihood can be further simplified by discarding all the pairs (i, j) that are not in the neighborhood structure $N(i), i = 1 \ldots n$. This choice provides a computationally efficient EM algorithm, without sacrificing the good distributional properties that are expected by a CL estimator.

Acknowledgements The authors were supported by the 2015 PRIN project 'Environmental processes and human activities: capturing their interactions via statistical methods', funded by the Italian Ministry of Education, University and Research.

References

1. Abe, T. and C. Ley (2017). A tractable, parsimonious and flexible model for cylindrical data, with applications. Econometrics and Statistics 4, 91-104.
2. Cianelli, D., Falco, P., Iermano, I., Mozzillo, P., Uttieri, M., Buonocore, B., Zambardino, G. and Zambianchi, E. (2015) Inshore/offshore water exchange in the Gulf of Naples. Journal of Marine Systems, 145, 37-52.
3. Eidsvik, J., B. A. Shaby, B. J. Reich, M. Wheeler, and J. Niemi (2014). Estimation and prediction in spatial models with block composite likelihoods. Journal of Computational and Graphical Statistics 23(2), 295-315.
4. Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics 31(4), 1208-1211.
5. Iermano, I., Moore, A. and Zambianchi, E. (2016) Impacts of a 4-dimensional variational data assimilation in a coastal ocean model of southern Tyrrhenian Sea. Journal of Marine Systems, 154, 157-171.
6. Lagona, F., M. Picone, and A. Maruotti (2015). A hidden markov model for the analysis of cylindrical time series. Environmetrics 26, 534–544.
7. Lindsay, B. (1988) Composite likelihood methods. Contemporary Mathematics, 80, 221-239.
8. Menna, M., Mercatini, A., Uttieri, M., Buonocore, B. and Zambianchi, E. (2007). Wintertime transport processes in the Gulf of Naples investigated by HF radar measurements of surface currents. Nuovo Cimento C, 30, 605-622.
9. Molenberghs, G. and G. Verbeke (2005). Models for discrete longitudinal data. Springer Series in Statistics. Springer Science+Business Media, Incorporated New York.
10. Okabayashi, S., L. Johnson, and C. Geyer (2011). Extending pseudo-likelihood for Potts models. Statistica Sinica 21(1), 331-347.
11. Ranalli, M., F. Lagona, M. Picone, and E. Zambianchi (2018). Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. Journal of the Royal Statistical Society C 67(3), 575-598.
12. Reich, B. and Fuentes, M. (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. Annals of Applied Statistics, 1, 249-264.
13. Uttieri, M., Cianelli, D., Nardelli, B. B., Buonocore, B., Falco, P., Colella, S. and Zambianchi, E. (2011) Multiplatform observation of the surface circulation in the Gulf of Naples (southern Tyrrhenian sea). Ocean Dynamics, 61, 779-796.
14. Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. Statistica Sinica 21(1), 1–41.
15. Wang, F., A. Gelfand, and G. Jona Lasinio (2015). Joint spatio-temporal analysis of a linear and a directional variable: space-time modeling of wave heights and wave directions in the Adriatic sea. Statistica Sinica 25, 25–39.

Estimation of entropy measures for categorical variables with spatial correlation
Stima di misure di entropia per variabili categoriche con correlazione spaziale

Linda Altieri, Giulia Roli
Department of Statistical Sciences, University of Bologna, via Belle Arti, 41, 40126, Bologna, Italy.

Abstract

Entropy is a measure of heterogeneity widely used in applied sciences, often when data are collected over space. Recently, a number of approaches has been proposed to include spatial information in entropy. The aim of entropy is to synthesize the observed data in a single, interpretable number. In other studies the objective is, instead, to use data for entropy estimation; several proposals can be found in the literature, which basically are corrections of the estimator based on substituting the involved probabilities with proportions. In this case, independence is assumed and spatial correlation is not considered. We propose a path for spatial entropy estimation: instead of intervening on the global entropy estimator, we focus on improving the estimation of its components, i.e. the probabilities, in order to account for spatial effects. Once probabilities are suitably evaluated, estimating entropy is straightforward since it is a deterministic function of the distribution. Following a Bayesian approach, we derive the posterior probabilities of a binomial distribution for categorical variables, accounting for spatial correlation. A posterior distribution for entropy can be obtained, which may be synthesized as wished and displayed as an entropy surface for the area under study.

Riassunto

L'entropia è una misura di eterogeneità ampiamente utilizzata nelle scienze applicate, dove spesso i dati sono georeferenziati. Recentemente, sono stati proposti svariati approcci per includere informazione spaziale nell'entropia, il cui scopo è sintetizzare l'osservato in un numero interpretabile. In altri studi, invece, l'obiettivo è utilizzare i dati per stimare l'entropia di un processo: esistono diverse proposte in letteratura, che sono correzioni dello stimatore basato sulla sostituzione delle probabilità con proporzioni. In questo caso, si assume indipendenza e non è considerata la correlazione spaziale. Proponiamo un nuovo percorso per la stima dell'entropia: invece di intervenire sullo stimatore globale, miglioriamo la stima delle sue componenti, le probabilità, per tenere conto di effetti spaziali. Una volta che le probabilità sono accuratamente valutate, si può stimare direttamente l'entropia, essendo questa una funzione deterministica della distribuzione di probabilità. Con un approccio Bayesiano, deriviamo la distribuzione a posteriori per variabili categoriche spazialmente correlate. Si ottiene quindi una distribuzione a posteriori per l'entropia, che può essere sintetizzata a piacere e diventa un superficie di entropia per l'area di interesse.

Keywords: Spatial entropy, entropy estimation, CAR models, categorical variables

1 Introduction

Shannon's entropy is a successful measure in many fields, as it is able to synthesize several concepts in a single number: entropy, information heterogeneity, surprise, contagion. The entropy of a categorical variable X with $I < \infty$ outcomes is

$$H(X) = \sum_{i=1}^{I} \log(p(x_i)) \log\left(\frac{1}{p(x_i)}\right), \tag{1}$$

where $p(x_i)$ is the probability associated to the i-th outcome (Cover and Thomas, 2006). The flexibility of such index and its ability to describe any kind of data, including categorical variables, motivate its diffusion across applied fields such as geography, ecology, biology and landscape studies. Often, such disciplines deal with spatial data, and the inclusion of spatial information in entropy measures has been the target of intensive research (see, e.g., Batty, 1976, Leibovici, 2009, Leibovici et al., 2014). In several case studies, the interest lies in describing and synthesizing data. This is usually not a simple task: large amounts of data require advanced computational tools, qualitative variables have limited possibilities, and, when data are georeferenced, spatial correlation should be accounted for. When it comes to measuring the entropy of spatial data, we suggest an approach proposed in Altieri et al. (2018), which allows to decompose entropy into a term quantifying the spatial information, and a second term quantifying the residual heterogeneity.

In other cases, though, the aim lies in estimating the entropy of a phenomenon, i.e. in making inference rather than description. Under this perspective, a stochastic process is assumed to generate the data according to an unknown probability function and, consequently, an unknown entropy. One realization of the process is observed and employed to estimate such entropy. The standard approach relies on the so-called 'plug-in' estimator, presented in Paninski (2003), which substitutes probabilities with observed relative frequencies in the computation of entropy:

$$\widehat{H}_p(X) = \sum_{i=1}^{I} \log(\widehat{p}(x_i)) \log\left(\frac{1}{\widehat{p}(x_i)}\right), \tag{2}$$

where $\widehat{p}(x_i) = n_i/n$ is the relative amount of observations of category i over n data. It is the non-parametric as well as the maximum likelihood estimator (Paninski, 2003), and performs well when $I < \infty$ is known (Antos and Kontoyiannis, 2001). For unknown or infinite I, estimator (2) is known to be biased; the most popular proposals at this regard consist of corrections of the plug-in estimator: see, for example, the Miller-Madow (Miller, 1955) and the jack-knifed corrections (Efron and Stein, 1981). Recently, Zhang (2012) proposed a non-parametric solution with faster decaying bias and upper limit for the variance when $I = \infty$. Under a Bayesian framework, the most widely known proposal is the NSB estimator (Nemenman et al., 2002), improved by Archer et al. (2014) as regards the prior distribution. Other approaches, linked to machine learning methods, directly estimate entropy relying on the availability of huge amounts of data (Hausser and Strimmer, 2009). In all these works, independence among realizations is assumed.

Two main limits concern entropy estimation. Firstly, the above mentioned proposals only focus on correcting or improving the performance of (2). Secondly, no study is available about estimating entropy for variables presenting spatial association: the assumption of independence is never relaxed, while spatial entropy studies do not consider inference.

In this paper, we take a perspective to entropy estimation, which moves the focus from the index itself to its components. Entropy, as defined in equation (1), is a deterministic function of the probability mass function (pmf) of the variable of interest. Therefore, once the pmf is properly estimated, the subsequent steps are straightforward. In the case of categorical variables following, e.g., a multinomial distribution, the crucial point is to estimate the distribution parameters. A Bayesian approach allows to derive the pmf of such distribution and can be extended to account for spatial correlation among categories. After obtaining a posterior distribution for the parameters, this is used to compute the posterior distribution of entropy as a transformation. Thus, a point estimator of entropy can be, e.g., the mean of the posterior distribution of the transformation; credibility intervals and other syntheses may be obtained via the standard tools of Bayesian inference. This approach can be used for non-spatial settings as well; in the spatial context, coherently with standard procedures for variables linked to areal and point data, the estimation output is a smooth spatial surface for the entropy over the area under study.

The paper is organized as follows. Section 2 summarizes the methodology for Bayesian spatial regression and shows how to obtain the posterior distribution and the Bayesian estimator of entropy. Then, Section 3 assesses the performance of the proposed method on simulated data for different spatial configurations. Lastly, Section 4 discusses the main results.

2 Bayesian spatial entropy estimation

For simplicity of presentation, we focus on the binary case. Let X be a binary response variable with $x_1 = 1$ and $x_2 = 2$; consider a series of n realizations indexed by $u = 1, \ldots, n$, each carrying an outcome $x_u \in \{1, 2\}$. This may be thought of as a n-variate variable, or alternatively as a sequence of variables $X_1, \ldots X_n$, which are independent, given the distribution parameters and any effects modelling them. For a generic X_u, the simplest model is:

$$X_u \sim Ber(p_u) \tag{3}$$

$$logit(p_u) = z_u^{'}\beta \tag{4}$$

in absence of random effects, where z_u are the covariates associated to the u-th unit.

To the aim of including spatial correlation, consider n realizations from a binary variable over the two-dimensional space, where u identifies a unit via its spatial coordinates. Let us consider the case of realizations over a regular lattice of size $n = n_1 \times n_2$, where u identifies each cell centroid. The sequence $X_1, \ldots, X_n$ is now no longer independent, but spatially correlated. In order to define the extent of such correlation for grid data, the notion of neighbourhood must be introduced, linked to the assumption that occurrences at certain locations are influenced by what happens at surrounding locations, i.e. their neighbours. The simplest way of representing a neighbourhood system is via an adjacency matrix: for n spatial units, $A = \{a_{uu'}\}_{u,u'=1,\ldots,n}$ is a square $n \times n$ matrix such that $a_{uu'} = 1$ when unit u and unit u' are neighbours, and $a_{uu'} = 0$ otherwise; in other words, $a_{uu'} = 1$ if $u' \in \mathcal{N}(u)$, the neighbourhood of area u, and diagonal elements are all zero by default.

In the remainder of the paper, the word 'adjacent' is used accordingly to mean 'neighbouring', even when this does not correspond to a topological contact. The most common neighbourhood systems for grid data are the '4 nearest neighbours', i.e. a neighbourhood formed by the 4 pixels sharing a border along the cardinal directions, and the '12 nearest neighbours', i.e. two consequent pixels along each cardinal direction plus the four ones along the diagonals.

Auto-models provide a way of including spatial correlation, by explaining a response via the response values of its neighbours. They are thus developed by combining a logistic regression model with autocorrelation effects, and are initially developed for the analysis of plant competition experiments and then extended to spatial data in general. Besag (1974) proposed to model spatial dependence among random variables directly (rather than hierarchically) and conditionally (rather than jointly). The autologistic model for spatial data with binary responses emphasises that the explanatory variables are the surrounding array variables themselves; a joint Markov random field is imposed for the binary data. A recent variant of this model, substituting (4) with (5), is proposed by Caragea and Kaiser (2009):

$$logit(p_u) = z(u)'\beta + \sum_{u' \in \mathcal{N}(u)} \eta_{u'}(X_{u'} - \mu_{u'}) \tag{5}$$

where η parametrizes dependence on the neighbourhood and, in the simplest case, $z(u)'\beta = \beta_0$ only includes an intercept. Parameter $\mu_u = \exp(z(u)'\beta)/(1+\exp(z(u)'\beta))$ represents the expected probability of success in the situation of spatial independence.

An analogous rewriting in the form of a CAR model (Cressie, 1993), in absence of covariate information, is

$$\begin{aligned} &logit(p_u) = \beta_0 + \phi_u \\ &\phi \sim MVN_n(0,\Sigma) \\ &\Sigma = [\tau(D - \rho A)]^{-1} \end{aligned} \tag{6}$$

where $\phi = (\phi_1, \dots, \phi_n)'$ is a spatial effect with a structured covariance matrix Σ, which depends on a precision parameter τ and a dependence parameter $\rho \in [-1,1]$ quantifying the strength and type of the correlation between neighbouring units. The symbol A denotes the adjacency matrix reflecting the neighbourhood structure, and D is a diagonal matrix, where each element contains the row sums of A.

The estimation of the parameters for Bayesian spatial logit regression models may proceed via MCMC methods or the INLA approach. We exploit the latter (Rue et al., 2009) and obtain a posterior distribution for the parameters of the probability of success for each grid cell. A synthesis, such as the posterior mean, is chosen in order to obtain an estimate for p_u over each cell. Such estimate is used for the computation of a local estimated entropy value for each pixel:

$$\widehat{H}(X)_u = \hat{p}_u \log\left(\frac{1}{\hat{p}_u}\right) + (1-\hat{p}_u)\log\left(\frac{1}{1-\hat{p}_u}\right). \tag{7}$$

This way, an entropy surface is obtained for estimating the process entropy, whose smoothness may be tuned by the neighbourhood choice, or by the introduction of splines for the spatial effect. Any other surface can be obtained following the same approach for different aims, e.g. for plotting the entropy standard error or the desired credibility interval extremes.

3 Simulation study

To the aim of assessing the performance of the proposed entropy estimator, we generate binary data on a 40×40 grid under two spatial configurations: clustered and random. Figure 1 shows an example of the generated datasets. The underlying model is (6), with a 12 nearest neighbour structure for A, $\tau = 0.1$ and $\rho = \{0.999, 0.001\}$ for the two scenarios, respectively. For each scenario, 200 datasets are generated with varying values for β_0 so that the expectation of p_u in a situation of independence varies between 0.1 and 0.9; values for β_0 differ across replicates but are constant across scenarios, so that the proportion of pixels of each type is comparable.

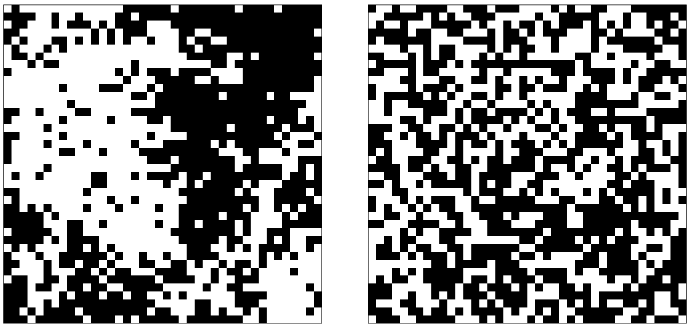


Figure 1: Clustered scenario (left) and random scenario (right) - example with $\beta_0 = 0.57$.

Results show that fitting the model over the generated data leads to good estimates for the p_us. For all replicates on both scenarios, the estimated parameters are very close to the true ones, which are always included within the 95% credibility intervals. The proposed approach is able to produce good estimates for the probabilities of success, ensuring the goodness of the estimates for spatial entropy, which is a function of such probabilities.

Obtaining the entropy surface proceeds as follows. First, the posterior distribution for each p_u is synthesized with its posterior mean. This way, for each scenario and replicate we obtain a single number for $\hat{p}_u$ on every cell. Then, an entropy value is computed over each cell following (7) and a smooth spatial function is produced. An example is shown in Figure 2, where values range from 0 (dark areas in the figure) to $\log(2)$ (white areas). The clustered situation (left panel) shows a smoothly varying surface. By comparing the left panels of Figure 1 and 2, one can see that the entropy surface takes low values in areas where pixels are of the same type: white pixels in the top-left part in Figure 1, and black pixels in the top-right part of Figure 1, correspond to the darker areas of Figure 2 where entropy values are low. In the areas where white and black pixels mix, the entropy surface tends to higher values (whiter areas in Figure 2). The random configuration (right panel of Figure 2) has a constant entropy close to the maximum $\log(2)$; this is expected, as no spatial correlation influences the entropy surface in this scenario. Therefore, such spatial functions properly estimates the entropy of the underlying spatial process.

Thanks to the availablity of the marginal posterior distribution of all parameters, any other useful synthesis is straightforward to compute. An example is shown in Figure 3, where the standard error of the estimate over each cell is plotted. Again, it is possible to appreciate a smooth surface for the clustered scenario, while the value is constant for the random one.

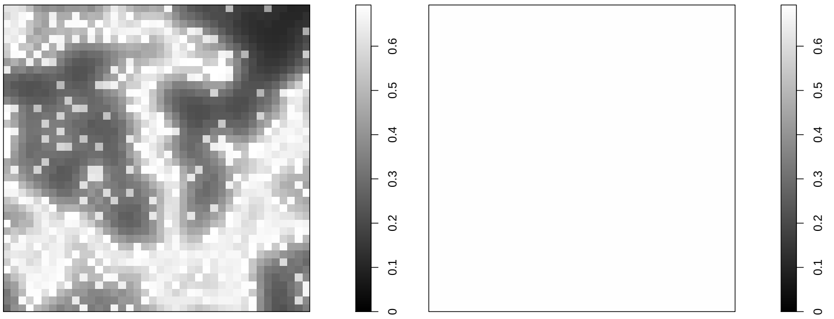


Figure 2: Example of estimated entropy surface for the two scenarios.

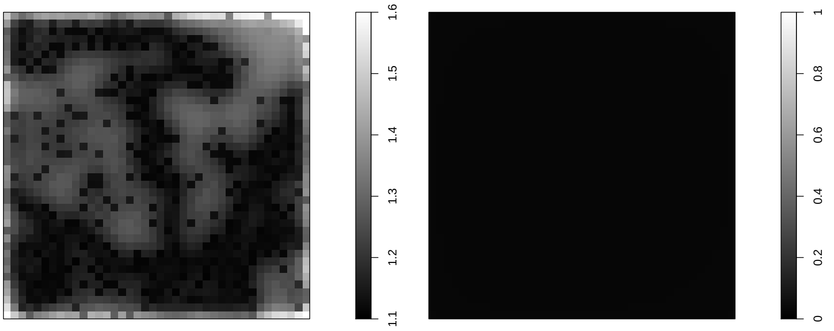


Figure 3: Example of the standard error of the estimate for the two scenarios.

4 Concluding remarks

In this paper, we describe an approach to entropy estimation which starts from rigorous posterior evaluation of its components, i.e. the probabilities. This way, we frame entropy within the theory of Bayesian models for spatial data, thus assembling the available results in this field.

Results from the simulation study enforce the validity of the approach in providing good estimates for the distribution parameters and, consequently, for entropy. The flexibility of the Bayesian paradigm allows to synthesize the posterior distribution of entropy as wished, in order to answer different potential questions.

Our procedure ensures realistic results, since, when the behaviour of a spatial process is under study, the basic hypothesis is that it is not constant but smoothly varying over space. In the same spirit, an appropriate spatial entropy measure is not a single number, rather it has to be allowed to vary over space as a smooth function.

The choice of the INLA approach allows to obtain results in a very reasonable time (minutes) for models including covariates and for very fine grids too, provided the model does not get extremely complicated in terms of random effects.

Acknowledgements
This work is developed under the PRIN2015 supported project 'Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)' [grant number 20154X8K23] funded by MIUR (Italian Ministry of Education, University and Scientific Research).
We thank an anonimous referee for the useful comments.

References

Altieri, L., D. Cocchi, and G. Roli (2018). A new approach to spatial entropy measures. *Environmental and Ecological Statistics 25*, 95–110.

Antos, A. and I. Kontoyiannis (2001). Convergence properties of functional estimates for discrete distributions. *Random structures and algorithms 19*, 163–193.

Archer, E., I. M. Park, and J. W. Pillow (2014). Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research 15*, 2833–2868.

Batty, M. (1976). Entropy in spatial aggregation. *Geographical Analysis 8*, 1–21.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Caragea, P. and M. Kaiser (2009). Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics 14*, 281–300.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory. Second Edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Cressie, N. A. C. (1993). *Statistics for spatial data (rev. ed.).* New York, Wiley.

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *Annals of statistics 9*, 586–596.

Hausser, J. and K. Strimmer (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research 10*, 1469–1484.

Leibovici, D. G. (2009). *Defining spatial entropy from multivariate distributions of co-occurrences*. Berlin, Springer: In K. S. Hornsby et al. (eds.): COSIT 2009, Lecture Notes in Computer Science 5756, pp 392-404.

Leibovici, D. G., C. Claramunt, D. LeGuyader, and D. Brosset (2014). Local and global spatio-temporal entropy indices based on distance ratios and co-occurrences distributions. *International Journal of Geographical Information Science 28*(5), 1061–1084.

Miller, G. (1955). *Note on the bias of information estimates*. Glencoe, IL free press: In H. Quastler (ed.) Information Theory in psychology II-B, pp. 95-100.

Nemenman, I., F. Shafee, and W. Bialek (2002). *Entropy and inference, revisited*. Cambridge, MA: MIT Press: In T. G. Dietterich, S. Becker, and Z. Ghahramani (eds.) Advances in neural information processing, 14.

Paninski, L. (2003). Estimation of entropy and mutual information. *Journal of Neural Computation 15*, 1191–1253.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B 71*, 319–392.

Zhang, Z. (2012). Entropy estimation in Turing's perspective. *Journal of Neural Computation 24*, 1368–1389.

Innovations in Census and in Social Surveys

A micro-based approach to ensure consistency among administrative sources and to improve population statistics

Un approccio basato sui microdati per garantire la coerenza tra fonti amministrative e per migliorare le statistiche demografiche

Gianni Corsetti, Sabrina Prati, Valeria Tomeo, Enrico Tucci

Abstract According to Istat modernisation programme the Registers Integrated System (RIS) is the core of the data production process. The RIS enables to enlarge the level of the analysis and the quality of the information collected linking at micro level the economic and social phenomena. In this view, the Base Register on Individuals and Households (BRIH) that identifies the usually-resident population will be enriched by the information collected by the other Base Registers (enterprises, addresses, activities) and by the thematic ones (Labour, Educational level, etc.). BRIH will be the common target for Continuous Census and Demographic statistics and for Social Statistics. To improve the accuracy of the population outputs in terms of timeliness, coverage and consistency it has been developed an information system of micro demographic accounting based on a longitudinal statistical register of individuals (Anvis). Anvis let us to continuously monitor population changes by linking the population stock with outflows and inflows. The new system allows us to adopt a multistate demographic approach for longitudinal models and for indicators based on change-of-state probabilities. Moreover, it let us to bring together all the events involving a given individual especially for the analysis of migration and migration trajectories.

Abstract *Secondo il programma di modernizzazione dell'Istat, il Sistema Integrato dei Registri (SIR) è il nucleo del processo di produzione dei dati. Il SIR consente di ampliare il livello delle analisi e la qualità delle informazioni raccolte integrando a*

[1] Gianni Corsetti, Istat; email: giacorsetti@istat.it
Sabrina Prati, Istat; email: prati@istat.it
Valeria Tomeo, Istat; email: tomeo@istat.it
Enrico Tucci, Istat; email: tucci@istat.it

livello micro i fenomeni economici e sociali. In questa prospettiva, il Registro Base su Individui e Famiglie (RBI), che identifica la popolazione abitualmente dimorante, sarà arricchito dalle informazioni raccolte dagli altri registri di base (imprese, indirizzi, attività) e da quelli tematici (manodopera, livello di istruzione, eccetera.). RBI sarà il punto di riferimento comune sia per il censimento permanente che per le statistiche demografiche e sociali. Per migliorare la qualità delle stime sulla popolazione in termini di tempestività, copertura e coerenza è stato sviluppato un sistema informativo di contabilità micro-demografica basato su un registro statistico longitudinale di individui (Anvis). Anvis ci permette di monitorare continuamente i cambiamenti della popolazione collegando lo stock di popolazione con gli eventi demografici. Attraverso Anvis sarà possibile, inoltre, realizzare analisi longitudinali seguendo l'approccio della demografia multistato e costruire indicatori basati sulle probabilità di cambiamento di stato. Infine, il sistema Anvis riunirà gli eventi riferiti ad un singolo individuo in modo da consentire studi sulle migrazioni e sulle traiettorie migratorie.

Key words: administrative data, population statistics, data integration, longitudinal database.

1 Introduction

This paper presents an overview of some results from the studies which Istat is carrying out in the context of the modernisation of Population and Social statistics. The object is to make the best use of the available administrative sources within the population statistics. New technologies and administrative changes allow to enhance the exploitation and the quality of data collected mainly due to the possibility to integrate different data sources. Istat has been developed a project of engineering and implementing an information system of micro demographic accounting based on a statistical population register of individuals Anvis (ANagrafic VIrtual Statistical database). Anvis makes it possible to continuously monitor population changes by linking the population stock with outflows and inflows. Fundamental requirements in Anvis information system are data integration and longitudinal consistency. For this reason, Istat needs to check and correct the system to improve data quality in terms of timeliness, coverage and consistency with other administrative sources. Italy supports the adoption of a Regulation on Demographic Statistics. We are aiming at improving the features of our systems according to the harmonization principle incorporated in the Regulation. ANVIS ensure the coherence and the consistence among the information requested by the Regulation on demographic statistics (in progress) and the information collected pursuant Regulation (EC) No 862/2007(of the European Parliament and of the Council on Community statistics) on migration and international protection and Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses. At the same time Anvis ensure the consistency of the Regulation to our National Legal requirements.

The new system will allows to achieve two research targets:

- A new longitudinal microdata system (MIDEA) for longitudinal models and techniques and demographic indicators based on change-of-state probabilities;
- An integrated system of administrative signals for measuring emigration flow.

2 The new longitudinal microdata system (Mi.DE.A-ANVIS)

The MIcro Demografic Accounting (Mi.DE.A) is a very large-scale linkage study which is going to be build up at Istat by using data of last Population Census and those of current administrative and statistical sources. These include all individual data available from the compulsory registration System (POPREG), such as Vital Events data (births, deaths, migration in or out Italy). Other data will be included in future steps, i.e marriages separations and divorces, or Health data collected by the National Health Service (Cedap, hospital admissions), or Household sample data.

There are some key differences among the MiDEA project and other longitudinal studies conducted in Europe as the Scottish Longitudinal Study (SLS) or as the England and Wales Longitudinal Study (LS). First of all the size of population involved into the Study. MiDEA is designed to include all the individual records captured in the last Population Census, around 60 million of records. Information for these people have been linked with vital events in order to follow the changing through time. In addition new records will be included or withdrawn each year using migration flows.

The ANagrafic VIrtual Statistical database (AN.VIS.) is one of the most relevant outcome of the project. It includes one records for each individual registered as resident according to census and flows data. An important aspect of the ANVISs individual data is the information on other household members (those who live in the same household as resulting at the time of last census or according to the variations collected by population registers).

The db [ANVIS] is a part of the Base Register on Individuals and Households (BRIH) in which individual Information can be merged with information of those living in the same household. Household data provide valuable additional information. In fact is the universe of populations and sub-populations for census and samples social surveys. This opens new opportunities and methodological challenges for the design of sample surveys. In perspective for each individual demographic information will be integrated with other socio-economic also in longitudinal perspective.

The starting point of the MiDEA are the Unit Records of Population Census at census time. The first step was to up-date census data until 31th December 2012 using Population register (POPREG) flows report entries and exits (Births, Deaths, Immigration and Emigration, Administrative adjustments); this process have been repeated for each year in order to up-date data at 31th December 2016 (latest year available at moment).

At the end of each year we match Anvis and POPREG unit records in order to carry out corrections due to underreporting of changes on personal data. However, due to mistakes by the individuals or administrations concerned, errors occur especially related coverage issues of the demographic flows.

At the same time MiDEA-ANVIS ensure the consistency of the Regulation to our National Legal requirements.

The new longitudinal microdata system can improve the accuracy of currently released demographic indicators, especially the ones that are called incidence indicators because they require the calculation of the population at risk.

In the macro approach population at risk is calculated as an average between initial and final population, making uniform distribution hypothesis in the time of reference.

Longitudinal microdata system can link population stock and demographic events in the continuous time, so that it is possible a closer calculation of the population at risk in terms of person-time.

Person-time is an estimate of the actual time-at-risk – in years, months, or days – that all participants contributed to a population in a given reference period. A subject is eligible to contribute person-time only so long as that person does not yet have left population due to emigration or death (Figure 1).

Since ANVIS contains full information for birthdate and dates referred to any migratory events or death, we can use exact dates to calculate person-time.

Figure 1: Exposure to risk of the individuals in a reference year by type of demographic event

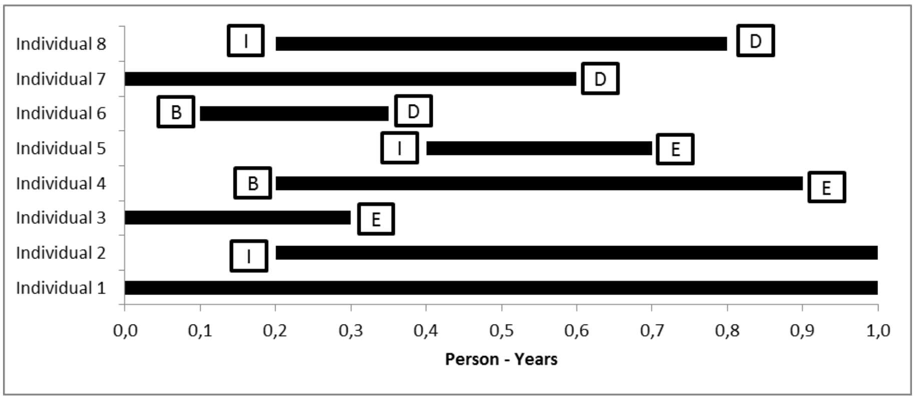
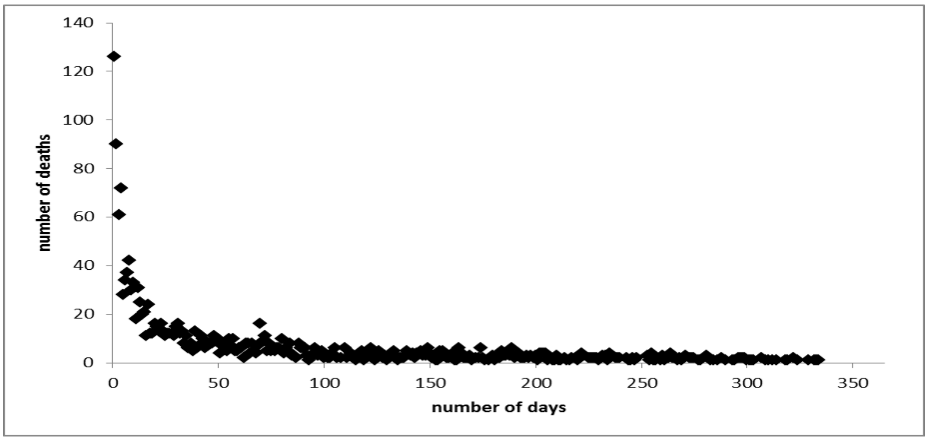


Table 1 shows age-specific death rates per thousand person-year by sex referred to 2016. They are compared with age-specific death rates per thousand inhabitant currently release by Istat and estimated with aggregated data (http://dati.istat.it/?lang=en&SubSessionId=1ca4016d-d5e1-4d13-a15c-a7ec98464094&themetreeid=-200).

Table 1: Age-specific death rates by sex and type of estimation – Year 2016

Age class	Male			Female		
	average population	person-years	relative difference	average population	average population	relative difference
	(A)	(B)		(A)	(B)	
0-4	0.7	0.8	14.3	0.6	0.7	16.7
5-9	0.1	0.1	-	0.1	0.1	-
10-14	0.1	0.1	-	0.1	0.1	-
15-19	0.3	0.3	-	0.1	0.1	-
20-24	0.4	0.4	-	0.2	0.2	-
25-29	0.4	0.4	-	0.2	0.2	-
30-34	0.5	0.5	-	0.2	0.2	-
35-39	0.7	0.7	-	0.4	0.4	-
40-44	1.1	1.1	-	0.7	0.7	-
45-49	1.8	1.7	5.6	1.1	1.1	-
50-54	3.0	3.0	-	1.8	1.8	-
55-59	5.0	5.0	-	2.8	2.8	-
60-64	8.0	8.0	-	4.3	4.2	2.3
65-69	13.2	13.1	0.8	7.1	7.0	1.4
70-74	21.4	21.3	0.5	11.7	11.7	-
75-79	35.7	35.4	0.8	20.8	20.6	1.0
80-84	65.9	65.4	0.8	42.8	42.5	0.7
85-89	123.7	122.9	0.6	87.3	86.7	0.7
90-94	218.9	216.3	1.2	168.5	166.5	1.2
95+	351	345.2	1.7	295.9	292.6	1.1
Total	10.0	10.0	-	10.2	10.2	-

Figure 2: Distribution of deaths in the first year of life by number of days between birth and death – Year 2016



Results are very close but differences increase in the first age class, where the uniform distribution hypothesis in the reference period is highly unlikely, because of the super mortality of the newborns in the very early period after birth, as shown in Figure 2.

This is a first attempt to fully exploit the potential of the new system under construction. Nevertheless we can underline the importance of longitudinal microdata in pursuing the goal of more consistently estimates.

3 Building an integrated system of administrative signals for measuring emigration flow

The necessity for improvements in migration statistics is not an issue that is confined to a single country, with an increased international policy focus on the socio-economic impacts of demographic change. As a dominant driver of population growth, robust estimation of immigration and emigration is key to the production of consistent national and regional population projections for EU countries (Lanzieri, 2007) and the implementation of EU Regulation 862/2007 has provided a statutory basis for greater harmonization of international migration statistics in Europe (Boden and Rees, 2008). However, there is a certain asymmetry between data availability on immigration and emigration, for two main reasons: departures tend to be less well recorded than arrivals and it is difficult to count persons leaving the country from a statistical point of view because of their absence (United Nations, 2010).

In Italy, a total of 410 thousand people emigrated abroad from the Census date (9th October 2011) to 1th January 2016, while around 450 thousand people were deregistered ex officio from the population registers. The latter figure is included in the demographic balance as «deregistration for other reasons» (based on a macro approach) but it is excluded from the migration flows statistics (micro events). Nevertheless, it is likely that a deregistration refers to someone who left the country without informing the administrative accountable office of their departure. Among these 450 thousand, there were almost 300 thousand of deregistration that could be assimilated to emigrations since they are related to individuals that, after the date of cancellation, do not reappear at least for 12 months. In order to classify these deregistration as migrations flows, the date of occurrence of the emigration has to be estimated.

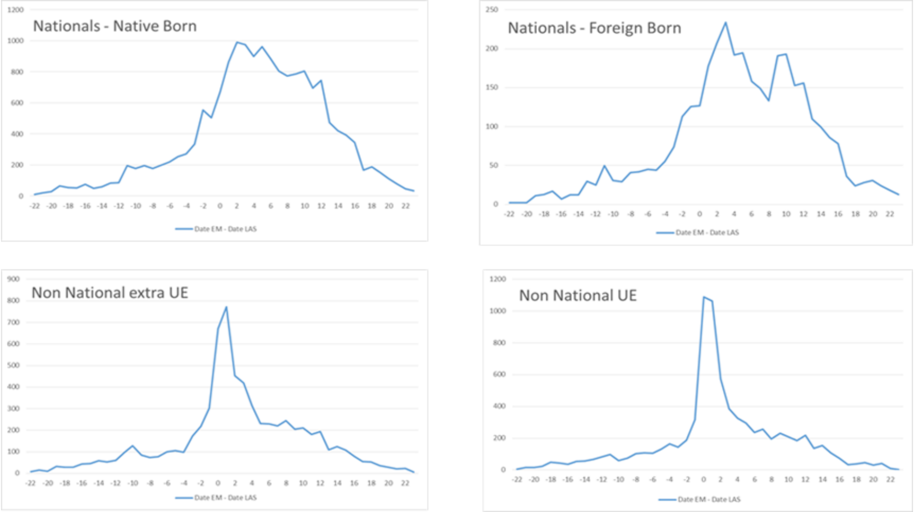
To this aim, we have used the information coming from other administrative sources (Labor and Education registers, Tax Returns register, Earnings, Retired, and Non-Pension Benefits registers, Permits to Stay archive) to derive a monthly Administrative Signal of presence in Italy.

As already mentioned, there have been 409,758 emigrations to abroad (EM) from the Population Census (9th October 2011) to 1th January 2016. Every individual emigration flow has been associated with the respective Last Administrative Signal (LAS) on the territory.

The underlying idea is that there is a relationship between the Date of emigration (Date EM) and the date of the Latest Administrative Signal (Date LAS). In particular, we have made the hypothesis that the difference between the mentioned dates is distributed normally.

Figure 3 shows that for non-nationals, the difference is strongly concentrated around 0 and 1 month, meaning that the majority of foreign citizens leaves the country very soon after they lose the signal of presence in Italy. The variability of the distribution is different for Nationals, meaning that, even without a signal of study or work, they can afford to stay longer in the country before emigrating abroad. In order to consider a deregistration ex officio as an emigration to abroad, the date of occurrence of the emigration (Date EM*) has to be imputed since the date of deregistration ex officio (Date DER) of a person has probably been postponed respect to the date of occurrence of the emigration (Date EM). To this aim, also the individual Date of deregistration ex officio has been associated with the respective Date of latest Administrative Signal on the territory (Date LAS).

Figure 3: Observed differences (in month) between Date of emigration and Latest Administrative Signal by citizenship (CTZ) and country of birth (CTB)



Observed Diff EMLAS distribution (by CTZ and CTB) has been used to estimate the date of occurrence of the emigration (Date EM*) for those deregistration carried out ex officio. In summary, the procedure was as follows:

1. Estimation of normal distribution parameters from observed data (Diff EMLAS = Date EM – Date LAS).
2. Computation of emigration data (Date EM* = Random Normal (Diff EMLAS) + Date LAS) for individuals deregistered ex officio by simulating Diff EMLAS with the normal distributions estimated in the previous step.

Figure 4: Observed differences (in month) between Date of deregistration and Latest Administrative Signal by citizenship (CTZ) and country of birth (CTB)

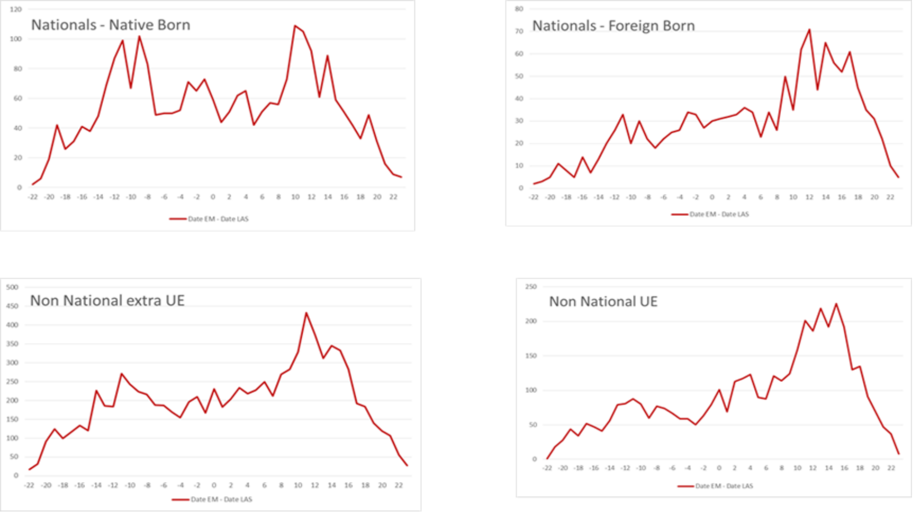
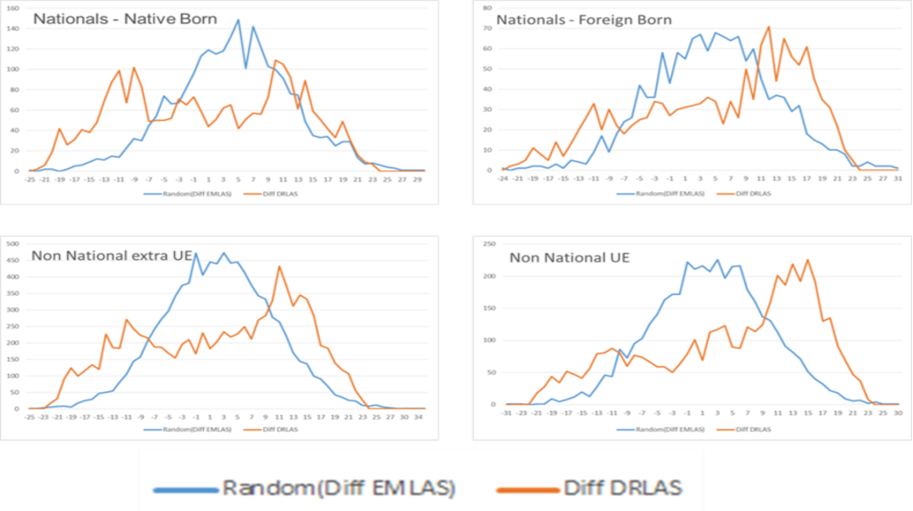


Figure 5: Estimated random normal differences (Diff EMLAS*) and observed ones (Diff DERLAS) by citizenship and country of birth



According to a first estimation of the emigration flows, the number of emigrants, almost doubled the number published by the official statistics. The latest figure for 2015, along with the number of immigrants during the same year, produces a negative net migration of -4 thousand while official statistics reported a positive net migration equal to +133 thousand units (Table 2).

An analysis by citizenship reveals that the Non Nationals report the largest impact on the emigration figures (from 45 thousand to 112 thousand). As a result the number of

Non National leaving the country is almost equal the number of Nationals. If we take into consideration that more than 20% of emigrants with Italian citizenship is born abroad, we can conclude that the majority of the people that leave the country has a foreign origin.

Table 2: Immigration, emigration flows and net migration observed and estimated (observed and estimated) by citizenship – Year 2015

Citizenship	***Immigration***	***Emigration***	***Net MIG***	***Emigration****	***Net MIG****
		Observed		***Estimated***	
Nationals	30,052	102,259	-72,207	147,300	-117,248
Non nationals	250,026	44,696	205,330	137,450	112,576
Total	280,078	146,955	133,123	284,750	-4,672

4 Conclusion

There is a lack of longitudinal database in Italy. An integrated period and cohort approach is a pillar of the MiDEA-ANVIS project. In order to produce information useful to the interpretation and the explanation of demographic phenomena, a deep change of strategies in organising basic statistics data is required, overcoming the traditional period approach at favour of cohort one.

Migration is the most difficult component of population change to evaluate, as if there is a comprehensive system which registers migration in our Country, either moves to or from the rest of the world, or moves within the Italian Municipalities.

The traditional concepts of resident or present population are less and less reliable to assess people usually or actually living in a Municipality, due to the increasing mobility of the population and the international migration (an international migrant is defined by the United Nations (UN) as someone who changes country of residence for 12 months or more).

Measuring migration movements of populations in different countries is a challenging task.

Administrative sources are an extremely powerful tool and a potential source of information for emigration statistics. The Italian case provided empirical evidence of the issues to be confronted and the challenges to the use of a data integration for improving the quantity and quality of data on emigration. A coherent and consistent database that contains detailed, up-to-date and accurate information allows to study migration through a longitudinal approach (Return Migration, Circular Migration).

References

1. Verhoef, Rolf, and Dirk J. van de Kaa. "Population registers and population statistics." Population Index (1987): 633-642.
2. Poulain, Michel, Anne Herm, and Roger Depledge. "Central population registers as a source of demographic statistics in Europe." Population 68.2 (2013): 183-212.http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999.
3. Myrskyla, Pekka. "New statistics made possible by the use of registers." Statistical Journal of the United Nations Economic Commission for Europe 16.2, 3 (1999): 165-180.

Demographic changes, research questions and data needs: issues about migrations

Cambiamenti demografici, domande di ricerca e necessità informative: considerazioni in tema di migrazioni

Salvatore Strozza and Giuseppe Gabrielli

Abstract After a synthetic introductory analysis on the trends of legal and resident population in Italy according to different sources, such as census, post-census estimations and Municipal Population Registers, the paper aims at focusing on the migratory component of population dynamics. In particular, it underlines the recent changes occurred to foreign immigration (for example, the importance of family reunifications and the growth of asylum seekers) to foreign-origin presence (from first to second generation, and from foreigners to new Italians), as well as to Italian emigration. This background framework will outline new information and research needs to which Italian public statistics system could or should handle considering the gained experience in other European countries (from cross-sectional to longitudinal surveys).

Abstract *Dopo aver proposto un sintetico esame introduttivo sulla determinazione della popolazione legale e residente attraverso i censimenti, la ricostruzione post-censuaria e i registri d'anagrafe, l'attenzione viene focalizzata sul bilancio demografico e sulla sua componente migratoria. In particolare vengono evidenziati i cambiamenti in atto con riguardo all'immigrazione e alla presenza straniera in Italia, ma anche all'emigrazione italiana all'estero. Si tratta del quadro di sfondo per richiamare le necessità informative e di ricerca che dovrebbero o potrebbero trovare risposta nel sistema delle rilevazioni italiane, anche tenendo conto dell'esperienza maturata a livello europeo.*

Keywords: migration, population, data sources, census, Italian Statistical System

[1] Salvatore Strozza, Università di Napoli Federico II, strozza@unina.it
Giuseppe Gabrielli, Università di Napoli Federico II, giuseppe.gabrielli@unina.it

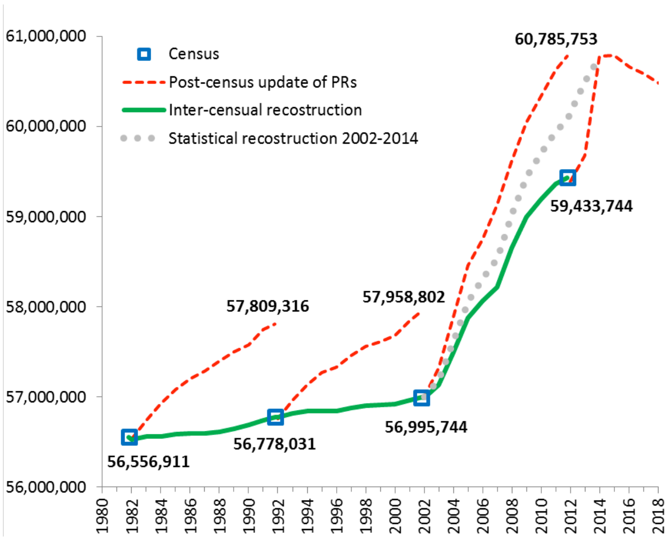
1 Introduction

The contribute has undoubtedly an ambitious title, maybe it is too ambitious compared to what it intends to analyse. It not refers to the general picture of demographic changes in recent decades and to all issues of current and future demographic interest, but it focuses only on the determination over time of the resident population, on the collection and estimation of migration flows and stocks and on some current and future challenges related to migration processes. These challenges may found an answer through information that will be acquired with the new population census strategy [14, 12, 6]. The aim is to recall the main information and research needs that the Italian Statistical System should or could provide on migration issues.

2 Determination over time of the resident population

It is well known that the census resident population is the so-called "de iure" population or “legal” population. Until now, the update of resident population, coming from Municipal Population Registers (MPRs), started from the amount of legal population has been recorded through the last census (Figure 1). Also the re-count of population (the so-called “inter-censual reconstruction”) was provided in previous inter-census period starting from (using as constraints) the legal population of the last two census.

Figure 1: Total resident population by different data collection or estimates, 1980-2018.



Source: own elaborations on Istat data.

The update population of MPRs is always higher than census resident population (1,031,000 people in 1991, 963,000 in 2001 and 1,352,000 in 2011). Synthetically, the difference depends on the over-estimation of resident people by MPRs (due to no deregistration of the people leaving the municipality) and under-count of census resident population (due to the well-known coverage problems of the total survey). A relatively recent novelty is that the ISTAT [11] proposed also a statistical reconstruction estimating the regional population in the period 2002-2014.

The population estimates are the architrave not only of demographic statistics but also of socio-economic statistics. The reduction in the timing of the release of the 2011 census results and of the revision and update times of the registers (resulted in a considerable recovery of people re-registered in 2012-2013) produced a large growth of the resident population. The statistical reconstruction of the population was therefore determined in order to eliminate the perturbations due to administrative issues, being used for the estimates coming from the main sample surveys (starting from the Labour Force Survey) and for the definition of the economic aggregates (starting from National Accounting). The main novelty consists in the correction of the 2001 census resident population based on the results of the Post Enumeration Survey (PES).

In general, it is clear that the decennial census involved a revision of population statistics every 10 years. In other words, the "reconstructed" population replaced the "updated" population about 10 years later. Even no large disparities occurred between the two different estimates (less than 2%), it is likely that 'permanent' Census of Population and Housing should overcome the re-count of population every decennial cycle.

In future, it will be interesting to understand how the "legal population" will be defined. For example: will the usual inhabitants be determined not considering the legal status?

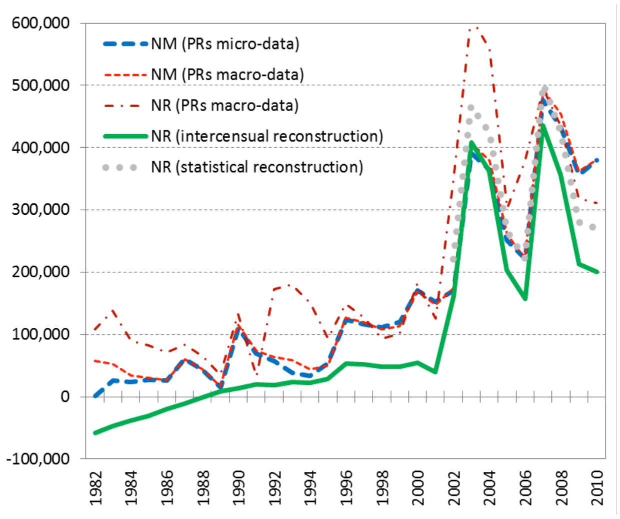
3 Determination over time of the net migration

Differences between the "updated" resident population and the "reconstructed" inter-censual population concerns also the (count or evaluation of) migratory component. Figure 2 shows net migration (NM) and net residue (NR) in the period 1982-2010 according to different data collection or estimates. In particular, net residue comprises all population variations except the natural change (births minus deaths).

The data of inter-censual reconstruction shows lower net migration than the other sources. Net migration turns from negative to positive during 80s, increases during 90s and widespread in the new Millenium. This is the story of the last 30 years but with different levels of the Italian migratory balance with the rest of the world.

During 80s net migration is positive (more than 500,000 people) according to direct estimation coming from the up-data population of MPRs but is negative (for almost 200,000 people) in the inter-censual reconstruction. During 90s, both sources of data show positive values, but the difference between them is of 600,000 of individuals. In the last observed decade the same difference enlarges to almost 800,000, but less important in relative terms.

Figure 2: Net Migration (NM) and Net residue (NR) according to different data collection or estimates, 1982-2010.



Notes: NR = NM plus other variations (excluding natural change); PRs = Population Registers
Source: own elaborations on Istat data.

Migrations become the main component of the demographic dynamics in Italy [7, 19]. In the last decade, the natural increase rate was -0.3 people per year every 1,000 inhabitants, while the net migration was in the range 4.5-5.8 immigrants per year every 1,000 inhabitants, with a significant gap of 1.4 points for every 1,000 inhabitants greater than in the previous two decennial periods (Table 1).

Table 1: Net Migration (NM) and Net Residue (NR) according to different data collection or estimates, 1981-1991, 1991-2001, and 2001-2011 inter-censual periods.

	Phenomenon (source of data)	Absolute values (in thousand)			Rates (per 1,000 inhabit.)		
		1981-91	1991-01	2001-11	1981-91	1991-01	2001-11
A.	NM (PRs micro-data)	404	978	3,350	0.7	1.7	5.8
B.	NM (PRs macro-data)	508	995	3,392	0.9	1.7	5.8
C.	NR (PRs macro-data)	833	1,358	3,953	1.5	2.4	6.8
D.	NR (inter-censual reconstruction)	-198	395	2,601	-0.3	0.7	4.5
E.	NR (statistical reconstruction)	…	…	3,243	…	…	5.6
F.	Difference between B and D	706	600	791	1.2	1.1	1.4

Source: own elaborations on Istat data.

Some scholars have talked about: “Italian migratory crossroads” [17]. Italy has not only become an important immigration country, but it remains a country of emigration of its citizens in the last years (more than during 80s and 90s). Data on registrations and de-registrations from and for/to abroad well show this picture. However, there are statistical uncertainties on the emigration size and sometimes on the sign of the net migration of Italians with foreign countries [18].

The direct data collection of MPRs counts 425,000 registration from abroad and almost 491,000 deregistration for abroad of Italian citizens in the period 2002-2011. Thus the net emigration of Italians is negative but of the small size (-65,000). As the emigration from Italy is generally under-estimated (for different reasons) and postponed (in respect to the effective date of the transfer), it was reasonable to expect a more consistent net emigration coming from indirect inter-censual estimation. Conversely, last two censuses lead to estimate a positive net migration of Italian citizens: arrivals from abroad exceed for 70,000 departures for abroad (Table 2). In particular, this outcome is evident considering Southern regions and Islands. The net immigration is even larger in the inter-censual reconstruction of population (almost 500,000). On the opposite, the net migration is negative of more than 200,000 people according to the Post Enumeration Surveys (PES) [5, 13].

Table 2: Registered and estimated net migration for Italian citizens in the 2001-2011 period, using different evaluations of the population residing at the two censuses (or around this data).

Sources of data	Resident population		Period 2001-2011		Net migration
	2001	2011	Natural increase	Acquisition citizenship	
Population Registers (PRs)*	…	…	…	…	-65
The two censuses (inter-censual recostruction)	55,661	55,406	-709	385	69
The two censuses, the second re-evaluated (statistical recostruction)*	55,653	55,812	-726	393	492
The two censuses, both re-evaluated according to the Post Enumeration Surveys	56,340	55,772	-709	385	-245

Note: (*) Population at the end of the year and/or period 2002-2011.
Source: own elaborations on Istat data. See also [18].

All these simple applications show in synthesis the existing problems to measure migration from and for abroad still in the past decade. In recent years this problem has returned to the center of attention. Enrico Pugliese in his last book [17] underlines that "Having a clear vision of the numerical consistency of the Italian emigrations also allows us to have an idea of its social significance" (p. 23). Thus … "It is not a useless exercise to try to understand if 110,000 people go away or a double or triple number, as some experts believe" (p. 25). Paradoxically, "arrivals (in the destination countries) exceed departures (from Italy)" (p. 27). This is particularly evident in Germany or in the Great Britain, as the author shown. It could depend on the source of data too. Anyway, the difference between the statistics of origin countries and the ones of destination countries exists since long time.

According to ISTAT data of cancellations for abroad, Italians departed toward countries of EU-15, Swiss and Norway were 83 thousands in 2016, equal to 73% of total cancellations. Thus, it is sufficient to look at these destination countries to get a not-exhaustive but general situation. Using Eurostat data, information are missing in 4 out of 16 considered countries: Germany, Greece, Ireland and Portugal. Unfortunately, data about Germany are missing. Looking at the 12 available destination countries in the period 2012-2016, we observe, according to Eurostat, almost 400,000 arrivals (registrations) of Italians in the 12 European countries. At the same time, Istat data register more than 230 departures (de-registrations from Italian MPRs) of Italians directed towards the same 12 European countries. The two sources present an hypothetical under-estimation of 69%. This is just an exercise to underline an existing counting problem of migration from and to abroad.

Can the New Population Census Strategy solve this discrepancy in the next future? Can the linkage between different administrative sources identify possible unregistered (undocumented) migrations and also unregistered return migrations? Can longitudinal data provide further information about the determinants of Italian emigration as the characteristics and the socio-economic conditions of the individuals and their origin families?

4 Foreign presence in Italy: resident and non-resident population

After twenty years of zero population growth (80s and 90s), resident population in Italy increases thanks to immigration. The number of foreign residents has been around 5 million in the last years. Today the information framework is much richer than in the past [16, 20, 10, 4] but there are still several gray areas and questions that cannot be answered. New issues and research questions emerge over time. The Statistical System, starting from the permanent census, should consider it.

ISMU estimation [1] counts almost 6 million of foreign population (resident, not resident and irregular people). They represent the 10% of total population. Irregular migrants represent in the last years a marginal quota (less than 10%) but a significant size (about 500,000) of the phenomena. It is unknown how many people, arrived in Italy from the Mediterranean Sea, enlarged the number of the irregular component or went elsewhere. In other words, now day it is not possible to determine the inclusion paths of (irregular) sea arrivals that can become asylum seekers and refugees in Italy.

Will the ‘permanent’ census overcome this limit? Will the permanent census contribute to estimate them as well as the regular non-resident population? Will it count naturalization of immigrant and of transnational couple descendants?

The National Statistical System has produced more and more reliable and exhaustive statistical information on migrations in the 80s and 90s. Significant improvements have been made to ensure the availability of unreliable information on the foreign population [16, 20, 10, 4], disaggregated on a territorial scale (at least at the provincial level). However, at the end of the last century, there were still numerous limitations and gaps such as [20]: the discrepancy between flows and stock data; the difficulty to define net migrations in the population equation; the lack of many information between two successive census due to the absence of reliable data on immigrants and/or foreigners in the National sample surveys (e.g. education, work, income, housing).

Census of 2001 represents the first and most important and reliable source on foreign (or immigrant) people, not only for the territorial detail (up to sub-municipal level), but above all for the importance of the acquired information [3, 2]. In adding, starting from the new Millenium, labor force (sample) survey became representative also for the foreign component of the population residing in Italy, ad-hoc sample surveys on households with at least one foreign component have been designed and implemented [4], the role of administrative data collections has been re-defined (in particular, residence permits) and record-linkage techniques have been performed.

Now day census data and survey data provide a wide and heterogeneous range of information on people with migration background. However, new requests of information raised and some lacks of information are still present about, in particular: characteristics and settlement process of undocumented arrivals; naturalization and integration processes of 1st and 2nd generations [15, 9, 8]; returning and emigration (that is generally under-estimated).

5 Future challenges related to migration processes

In conclusion, it is interesting to underline (in a schematic way) partial and not exhaustive future challenges related to migration processes. The Municipal Population Registers and Households (Surveys) have always been two reference elements for official statistics. However, this paradigm will be less and less relevant to analyze a more complex and dynamic society. This is particularly evident for migrant families/individuals. It will be increasingly necessary to identify new indicators of habitual residence and of different (non-cohabitant) family forms.

The data sources are increasingly differentiated and articulated (e.g. social data). In the future it will be necessary to link information that are different for definitions, concepts and methods, coming from a plurality of subjects.

It is increasingly necessary in the scientific research to be able to observe individuals according to a longitudinal approach to define the causal relationships of processes. Examples are: 1) the life trajectories of sea arrivals that become asylum seekers and refugees; 2) mobility and return migrations; 3) demographic behavior of migrants and their descendants; 4) processes of integration and naturalization.

The use of big and complex databases and of micro-data (as in the case of ‘permanent’ census), for the analysis of phenomena, has become indispensable. On the other hand, the “Cambridge Analytica” scandal has shown the increasing risks in data acquisition processes and the need of a correct use of information. The right balance between confidentiality of information and access to micro-data for research will be an important challenge. These issue is even more relevant when considering small groups of individuals, for example, by citizenship or by geographical context.

Centralization of data collection guarantees better monitoring and uniformity of statistical information. This is the case, for example, of the next National Register of Resident Population (ANPR). At the same time, given the persistent geographical differences in Italy, it will be increasingly necessary to guarantee analytical information at local level in order to identify local policies: in manner of fact the integration of immigrants regards the specific contexts in which they live.

References

1. Blangiardo G.C.: Gli aspetti statistici. In: ISMU, Ventitreesimo Rapporto sulle migrazioni 2017, Franco Angeli, Milano (2018).
2. Bonifazi, C., Gallo, G., Strozza, S., Zincato, D.: Popolazioni straniere e immigrate: definizioni, categorie e caratteristiche, Studi Emigrazione, n. 171, 2008, pp. 519-548 (2008)
3. Bonifazi, C., Strozza, S.: Conceptual Framework and Data Collection in International Migration. In: Caselli, G., Vallin, J., Wunsch, G. (eds.), Demography: Analysis and Synthesis. A Treatise in Population, pp. 537-554, Volume IV, Elsevier Inc., USA (2006)
4. Bonifazi, C., Strozza, S., Vitiello, M.: Measuring integration in a reluctant immigration country: the case of Italy. In: Bijl, R., Verweij, A. (eds.), Measuring and monitoring immigrant integration in Europe, pp. 183-199, The Netherlands Institute for Social Research / SCP, The Hague (2012)
5. Brancato, G., D'Orazio, M., Fortini, M.: La copertura del censimento e l'errore di risposta. In: Fortini, M., Gallo, G., Paluzzi, E., Reale, A., Silvestrini, A. (a cura di), La progettazione dei censimenti generali 2010-2011. Criticità di processo e di prodotto nel 14° Censimento generale della popolazione e delle abitazioni: aspetti rilevanti per la progettazione del 15° Censimento, pp. 107-123, ISTAT, Roma (2009).
6. Corsetti, G., Prati, S., Tomeo, V., Tucci, E.: A micro-based approach to ensure consistency among administrative data sources and to improve population statistics, 49th Scientific Meeting of the Italian Statistical Society, Palermo (2018)
7. De Rose, A., Strozza, S. (eds.): Rapporto sulla popolazione. L'Italia nella crisi economica, il Mulino, Bologna (2015)
8. Di Bartolomeo, A., Gabrielli, G., Strozza, S.: The labour market insertion of immigrants into Italy, Spain and the United Kingdom: similarities and differences and the Southern European model of migration. In: Ambrosetti, E., Strangio, D., Wihtol de Wenden, C. (eds.), Migration in the Mediterranean. Socio-economic perspectives, pp. 57-84, Routledge, London and New York (2016)
9. Donadio, P., Gabrielli, G., Massari, M. (eds.): Uno come te. Europei e nuovi europei nei percorsi di integrazione, FrancoAngeli, Milano (2014)
10. ISTAT (ed.): La presenza straniera in Italia: l'accertamento e l'analisi, Roma (2008)
11. ISTAT, Ricostruzione statistica delle serie regionali di popolazione del periodo 1/1/2002-1/1/2014, Nota informativa (2015).
12. Mastroluca, S., Verrascina, M.: Towards more timely census statistics: the new Italian multiannual dissemination programme, 49th Scientific Meeting of the Italian Statistical Society, Palermo (2018)
13. Mazziotta, M.: L'indagine di copertura (PES) del 15° Censimento generale della popolazione e delle abitazioni. I risultati definitivi, Seminario su "La misurazione della qualità del 15° Censimento generale della popolazione e delle abitazioni: i risultati dell'indagine di copertura (PES)", ISTAT, Roma, 27 giugno (2014).
14. Mazziotta, M., Falorsi, S.: The New Population Census Strategy: from Tradition to Innovation, 49th Scientific Meeting of the Italian Statistical Society, Palermo (2018)
15. Ministero dell'Interno, ISTAT (eds.): Integration. Knowing, Measuring, Evaluating, Rome (2013)
16. Natale, M., Strozza, S.: Gli immigrati stranieri in Italia. Quanti sono, chi sono, come vivono?, Cacucci Editore, Bari (1997)
17. Pugliese, E.: Quelli che se ne vanno. La nuova emigrazione italiana, Il Mulino, Bologna (2018)
18. Strozza, S.: L'emigrazione netta italiana: apparenza o realtà?, Neodemos, 23 luglio (2014)
19. Strozza, S., De Santis, G. (eds.): Rapporto sulla popolazione. Le molte facce della presenza straniera in Italia, il Mulino, Bologna (2017)
20. Strozza, S., Natale, M., Todisco, E., Ballacci, F.: La rilevazione delle migrazioni internazionali e la predisposizione di un sistema informativo sugli stranieri, Rapporto di ricerca n. 02.11, Commissione per la Garanzia dell'Informazione Statistica (CGIS), Presidenza del Consiglio dei Ministri, ottobre (2002)

Towards more timely census statistics: the new Italian multiannual dissemination programme

Verso una produzione censuaria più tempestiva: il nuovo piano di diffusione pluriennale italiano

Simona Mastroluca and Mariangela Verrascina

Abstract Italy is moving towards a new census strategy, integrating information arising from registers and recurring sample surveys. One of the main targets of such approach is the timeliness of the data; the delay in release of census data reduces their value and usefulness for many stakeholders. Therefore, a new Italian dissemination programme has been carried out to ensure a subset of census statistics every year; the contents can evolve overtime as user needs change and the availability of administrative data improves. The annual data dissemination will be supplemented by a larger decennial publication, which will occur for the reference year 2021, the same defined at European level. Meanwhile, Eurostat is working to define a strategy for future censuses founding on a more frequent data supply, especially on population and migration topics. The purpose is a multiannual data collection, essentially relied on administrative data sources starting from the mid-2020s maintaining the traditional decennial data collection that will be done in 2031. It reveals evident similarities with the Italian case.

Abstract *La nuova strategia censuaria italiana si basa sull'integrazione di informazioni provenienti da registri e da indagini campionarie periodiche. Uno dei principali obiettivi è quello di assicurare una maggiore tempestività dei risultati: il ritardo nel rilascio dei dati censuari riduce il loro valore e l'utilizzabilità degli stessi. A tal fine è stato predisposto un nuovo piano di diffusione che prevede, oltre alla tradizionale diffusione decennale riferita al 2021 (lo stesso definito a livello europeo), la produzione annuale di un ridotto set di incroci rimodulabili nel tempo in funzione delle esigenze degli utenti e dei dati amministrativi che via via si rendono disponibili. Contemporaneamente Eurostat sta lavorando alla progettazione dei censimenti futuri che dovranno garantire una fornitura più frequente dei dati a partire dal 2025, soprattutto su temi relativi a popolazione e*

[1] Simona Mastroluca, Istat, mastrolu@istat.it
Mariangela Verrascina, Istat, verrasci@istat.it

migrazioni, mantenendo comunque la consueta e più ampia diffusione decennale del 2031. Emerge chiaramente che la strategia europea per i censimenti post-2021 presenta numerose affinità con quella già adottata nel nostro Paese.

Key words: permanent census, data dissemination, European strategy, timeliness

1 The new Italian census strategy

The 2011 Italian general population and housing census has been the last one carried out in a traditional way even if many innovations as the mail out of forms and the self-completed questionnaires collected by a multimode system were planned to achieve a stable and enduring balance between census costs and benefits. Considering the huge amount of data from administrative sources, the reduced budget allocated for the enumerations and a growing need for more frequent census-like data, Italy designed a new census strategy. The systems that would allow annual data production are: the register-based systems, combined approaches based on administrative sources supported by surveys and the rolling census. The expected user need for more frequent information, especially on population topics, is reflected also in Eurostat's plans to develop a redesign of European census statistics after the next census round. A key aspect will be to ensure a complementarity of data to meet key user needs, while avoiding duplication of data collection. The new Italian population and housing Census, that will no longer be decennial but permanent, is based on the Census and Social Surveys Integrated System (CSSIS), a complex statistical process integrating the information arising from registers and surveys. It foresees a two phases Master Sample design consisting of a set of balanced and coordinated sampling surveys. The first phase of Master Sample is based on two different component samples, namely A (Areal) and L (List). The first is designed to satisfy the needs of estimating under-coverage and over-coverage rates of the Population Register at national and local level for different sub-population profiles like sex, age, citizenship. The component L - based on a list sample - has the purpose of thematic integration that is estimating the hypercubes which cannot be obtained using the replaceable information coming from Registers. Census topics have been classified as totally, partially or not replaceable. The first group includes variables for which the administrative sources provide the correspondent proxy information; they are considered complete because they are available for all units in the thematic registers and accurate, having a good level of coverage and quality. Partially replaceable variables are considered complete and accurate only for a subset of the target population; for the others, the topics are unknown or cannot be considered accurate. This is, for example, the case of the educational attainment: the Italian Ministry of Education does not provide information on Post-secondary non-tertiary education or degrees obtained abroad. In the last group (not replaceable variables) there are topics not yet available from administrative sources. Therefore, the Italian permanent census is a combined census using both the information produced by the

Statistical Registers, for replaceable variables, and the CSSIS for the remaining topics. For Population and Housing Census purposes, at the moment three Statistical Registers are under construction: the Population Register, containing the census topics sex, age, place of birth, legal marital status and citizenship, the Employed Register with information on Economic characteristics and the Register of Buildings, including data on buildings and dwellings. Seven years after the last traditional Italian Census, the first survey of the new census project is scheduled for October. The A and the L components involve almost 1400 thousand households and 2800 municipalities with different population size. The Area survey is a totally paperless door-to-door enumeration with CAPI (computer-assisted interview) technique. Questionnaire will be completed by enumerators supported by electronic devices (tablets). The List survey is totally paperless too. Households self-complete web questionnaire (CAWI). First step is the mailing of invitation letters with credential to login the web questionnaires; eventually, all not respondent households up to a specific date will be interviewed by enumerators to get all the forms filled (CAPI). The questionnaire is the same for both the components. It includes not only partially or not replaceable variables, but all the necessary topics to produce hypercubes with the aim of using collected information also to test quality and coverage of data already available in Registers. One of the main targets of the new census strategy is the timeliness of the information; the decennial data dissemination would be supplemented by an annual dissemination programme in order to face the stakeholders' needs.

2 Census data dissemination between tradition and innovation

2.1 **2011 data dissemination programmes**

The 2011 European Census of Population and Housing was the first census exercise based on a common legislation that covered in detail all aspects of census outputs. In particular, an output harmonisation approach was taken: countries were free to produce the required data using any appropriate methods and data sources according to national availability and preference. However, the European data needed to fully meet the standards as defined in the legislation – such as the definitions, breakdowns, cross-tabulations and quality reporting to be applied. The Regulation (EC) 763/2008 defined the variables to be included necessarily in the survey plan and the time required to make the definitive census data and the related metadata available to Eurostat; allowed NSIs flexibility in the choice of data sources that could be used to fulfil the requirements of the data collection. Member States provided the Commission (Eurostat) with final, validated and aggregated data and with metadata, as required by the Regulation, within 27 months of the end of the reference year. The Regulation Annex listed the census topics that should be included and sets the geographical levels for which data on these topics should be available. With a view to standardizing the census output, the Regulation no. 519/2010, concerning the dissemination programme of data and metadata, provided the list of cross-tabulations which, to various territorial details (from national to municipal) and at different classification levels, had to be validated and made available. The list of mandatory core topics present in the hypercubes was that reported in the Framework Regulation, while classifications referred to the Regulation no. 1201/2009 on breakdowns and technical specifications. ISTAT prepared and made available to Eurostat 60 hypercubes (5 at national level, 36 at NUTS2 level, 10 at NUTS3, 5 at LAU2 and 4 at place of work), corresponding to 175 principal marginal distributions; the data dissemination programme focused mainly on regional detail. The Regulation on hypercubes also defined the textual metadata that must be supplied for the census topics. Data were produced in the form of detailed multidimensional cross-tabulations (hypercubes) and accessible via the online Census Hub system, allowing users the flexibility to define tables according to their needs. The dissemination programme defined by Eurostat did not correspond to the Italian programme; in fact, the hypercubes listed in Regulation no. 519/2010 differed from the cross-tabulations characterizing the Italian dissemination in terms of information, breakdowns, definitions, classification details and territorial level. For the Italian dissemination it was necessary to satisfy the information needs of the national users and to guarantee continuity to the time series. Consequently, for the 2011 census, two different dissemination programmes were produced, one defined on the basis of the European Regulations, the other with tables crossings all the variables contemplated by the Italian survey plan, which, in addition to ensure continuity with the past, allow to disseminate information on phenomena of national interest (for example, mobility at municipal level) or for the first time investigated during the census in Italy (for example, the place of birth of parents). The dissemination took place only via the web, through the corporate datawarehouse (I.Stat); there are more than 300 multidimensional tables that users can export. Data were then disseminated for enumeration areas, thematic maps compared with the previous censuses and data in open format (Linked Open Data).

2.2 **EU Programme of the statistical data for the reference year 2021**

A European Task Force composed of 18 EU Member States and 3 Candidate Countries worked for the preparation of the new legislation. The 2021 EU census data collection is based on the existing Framework Regulation (763/2008) that provided the legal basis for the 2011 census, while new implementing regulations have been developed to define in detail the 2021 census: definitions of topics, classifications, programme of cross-tabulations, metadata and quality reports; even if the list of topics and their geographical breakdown cannot be changed. The Regulation no. 2017/543 concerns the technical specifications of the topics and of their breakdowns to be applied for the 2021 data collection. Compared to 2011, a number of changes for 2021 have been introduced to simplify wherever possible definitions and breakdowns and to remove optional categories. Definitions and breakdowns have also been updated to reflect new data needs, changes in society, or to take into account changes in other related statistical exercises. The Regulation no. 2017/712 is on the reference year, the table programme and textual metadata to be applied for the 2021 EU census data collection. The table programme for 2021 retains the richness of the census as a data source and covers the priority issues, while reducing as far as possible the burden on the NSIs. It contains significant reductions in terms of the number of tables, as well as in the dimensionality of the tables. Thus, the simplification of the data (resulting from the 2011 experience and outputs) has implied a reduction in the number and complexity of the cross-tabulations, the removal of cross-tabulations and/or dimensions that were little used in 2011 and a focus on variables that are priorities for users (in particular, data on migration-related variables). The 2011 data transmission programme included 175 principal marginal distributions, while the programme for 2021 is based on 119 smaller datasets. As was the case for 2011, the 2021 data will be disseminated primarily via the Census Hub. The 2021 EU census data collection can be seen as consisting of two areas: the 'main' data collection, as described above, and an additional set of data geo-referenced to a standard 1km² grid. Geographically detailed data from the census are essential for Regional and cohesion policy to plan, develop and evaluate policies. The collection of population data at the 1km² grid level is an area where there is significant user demand and which is undergoing rapid development in many NSIs. The potential value of merging geospatial data with official statistics is to provide better social and environmental information. The 2020 round of censuses would be an important opportunity for the integration of statistical and geospatial data. Only 13 numbers would be collected for each grid square, then disseminated via the Census Hub. This is a new development for the 2021 census, distinct from the main EU census data collection programme. This data collection is not based on Regulation no. 763/2008 but, instead, will use a 'temporary direct statistical action' based on the Regulation on European Statistics.

2.3 **The Italian dissemination programme for the Permanent Census**

Istat is working on the preparation of different dissemination programmes. The proposals start from a critical evaluation of the 2011 Italian programme (complexity of the tables, redundancies, outdated phenomena, requests from stakeholders, etc.), from the analysis of the use of the information available on I.Stat and on other dedicated sites and the results of the survey on information content, tools and dissemination strategies adopted in some foreign countries. The extended dissemination programme, in line with the EU, will have as reference year the 2021. With a view to simplification, the new ten-year programme has been redefined, ie cross-tabulations at regional level have been eliminated (preferring tables with a lower territorial detail), particularly complex cross-tabulations have been simplified but, at the same time, new cross-tabulations have been introduced to study social emerging phenomena. The draft decennial programme proposes around 250 tables, structured on various thematic areas and with different classification details and territorial levels (over 70 municipal tables) that, between tradition and innovation, guarantee the updating of the time series, a fine territorial detail, a high classification detail and new focus (seniors, children, foreigners, inactive people, NEET). Considering then the information available through the Registers and the annual survey (MS) on over one million households, a minimum set of tables at municipal level and possibly sub-municipal for municipalities over a certain demographic threshold has been developed (about 30, reconfigurable over time) to be disseminated from 2019 on an annual basis. The planned municipal cubes concern the areas “Population, Households and Family Nuclei”: population structure by sex, age and citizenship, migrations, educational attainment, current activity status, commuting, type of family nucleus and type of private household. The hypothesis of a uniform base dissemination for all the municipalities and of an additional dissemination depending on the population size is under study; the choice is strictly linked to the sample design that Istat intends to adopt for the Master Sample. From a further analysis carried out on the 2021 national dissemination programme, a new set of tables to be released to the provincial (and big cities) details has been identified, with a periodicity to be defined. The additional tables have variables not included in the annual proposal or more detailed classifications than those foreseen at municipal level. The Permanent Census dissemination can be read from the point of view of the Statistical data (annual, multiannual and decennial dissemination programmes) and of view of the Output presentation (tools to be adopted for the census data dissemination). The starting point is the preparation of predefined tables so the dissemination is made up of tables defined a priori. Then a dissemination differentiated by population size of municipalities should be introduced. The dissemination could be completed with data representation (maps and cartography improved than that used for 2011 census output). All these kinds of output need the use of new technologies and new dissemination tools, that means new approaches to the production and presentation of census data. For example, the following tools could be used: Stories (short analysis designed specifically for web publishing); Summaries (concise articles that illustrate only the main results of an analysis); Infographics (graphical representations of the main results of the analysis with some contextual information or annotations); Video podcasts (video to explain the main results of the analysis); Interactive content (maps, graphs and images that the user can manipulate to highlight areas of interest). Finally, customised analyses could be permitted, by creating a system that allows customised queries in the case users do not find the information they are looking for in the predefined tables or preparing a datawharehouse that allows the cross-tabulation of variables and classifications according to user needs.

3 Strategy for the post-2021 European census

In 2016 representatives of France, Germany, the Netherlands and the UK on the Task Force on future EU censuses drafted jointly with Eurostat a vision for the post-2021 census. This vision was a result of the consultations with key users that showed a growing need for more frequent census-like data. The vision is made possible by the fact that the vast majority of European countries have made, or are planning or considering, major changes to the data sources and methods used for the census; most widely, the increased use of data from administrative sources and a move away from a traditional census enumeration. A fundamental goal for all European countries should be the development of a post-2021 strategy that provides relevant census results for data users. Based on the 2011 data provision, greater relevance for the census can be achieved if countries agree to more frequent and timelier census publications. A good balance will need to be found between the level of detail and accuracy in the publications on the one hand and timeliness of the publications on the other hand. Moreover, it is important to develop a strategy that is appropriate for both countries with and countries without field enumerations. Although many countries have moved or are moving from traditional censuses towards combined and register-based censuses, it is clear that different approaches towards census methodology will remain within Europe. The post-2021 census can be defined in the form of two separate collections. The first of these would be an annual data dissemination based essentially on administrative data sources, to be introduced in the mid-2020s. This annual data dissemination could include occasional modules on particular topics where data were collected at an intermediate frequency (such as every 5 years). The annual data dissemination would be supplemented by a larger decennial data dissemination in the form of the recurring decennial census first to be done in 2031. The initial need identified is to have basic demographic and migration data available on an annual basis; further topics could be added to the developing annual data dissemination. Given this uncertainty on how the annual data dissemination will develop over time, and the need to look ahead for a period of up to 10-15 years in the future, the framework for the post-2021 development will be implemented in a flexible manner. The key features of the new EU census strategy are a census round every ten years supported by annual updates on relevant topics, annual updates starting with a core set of demographic and migration topics with limited cross-tabulations, flexibility regarding the topics of the annual updates and the timetable for increasing the amount of data available, timeliness of the publication of annual census data, data provision of annual updates on a detailed regional level and of annual population counts for a geographical grid (1 km² or smaller if possible), integration of annual census updates into the annual work programmes of the national and European Statistical systems. This implies a convergence and merging with the annual demographic statistics, as well as coherence with other European social statistics developments. The introduction of 1km² grid data already for the round of 2021 (above mentioned) is an important development. There is a growing user need for grid-based data because grid squares are more stable over time than, for example, municipality boundaries. As the production of geo-referenced data is time-consuming and cost-demanding, and the level of experience with geo-referenced census data differs greatly among European countries, a good starting point for annual updates is to begin with annual population counts by 1 km² grids. If that approach succeeds, the annual updates could gradually be extended. Of course future availability of administrative data is not so clear since no one knows what data will be accessible by the mid-2020s. Dissemination within 12 months is a difficult goal to reach for countries with large field data collections, even when (part of) the forms are filled in via internet. For those countries it might be an option to provide provisional results within 12 months and then, later, according to their own dissemination plan, definitive results. For countries with large field data collections it would probably be possible to make use of estimations to reach the 'within 12 months criterion'. The estimations should be precise enough for the European level of analyses made through the Census Hub data. Users said they prefer timely provisional results rather than outdated definitive information. As far as the content of annual census results is concerned, it is important to produce census data on a more detailed regional level, but to achieve good timeliness and policy relevance, the annual census results would have to focus on demographic and related migration variables. The need for some basic explanatory variables must also be considered, although these may prove to be more complex to produce. This implies that the annual updates only concern the Population Census and not the Housing Census and the tables should not be as high-dimensional as the current decennial hypercubes.

4 Summarizing

Thanks to the new Italian census strategy a subset of census statistics will be disseminated every year starting from 2019. A core set of demographic, socio-economic and migration topics with limited cross-tabulations and less detailed disaggregations will be promptly made available to the stakeholders. In 2016 the "Task Force on future EU censuses of population and housing", composed by Eurostat and 18 countries (including Italy), drafted a vision for the post-2021 census providing a framework for the development of census-type population statistics over a period of around 10-15 years, covering the more frequent (annual) statistics to be introduced and expanded after the completion of the 2021 census programme, as well as consideration of the likely needs for a full census programme in 2031. A fundamental goal for all European countries is the development of a post-2021 strategy that provides relevant and timely census results for data users. It means that Italy is realizing in advance what will be applied at European level only from the mid-2020s: it's a big challenge for our country that has to face many innovations in a very short time in order to produce high quality territorial annual census data.

References

1. Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses**,** *Official Journal L 218 , 13/08/2008.*
2. Commission Regulation (EU) No 519/2010 of 16 June 2010 adopting the programme of the statistical data and of the metadata for population and housing censuses.
3. Commission Regulation (EC) No 1201/2009 of 30 November 2009 regards the technical specifications of the topics and of their breakdowns.
4. Commission Implementing Regulation (EU) 2017/543 of 22 March 2017 regards the technical specifications of the topics and of their breakdowns
5. Commission Regulation (EU) 2017/712 of 20 April 2017 establishing the reference year and the programme of the statistical data and metadata for population and housing censuses

6. Documents in progress of the European Commission (Eurostat, Directorate F: Social Statistics) presented in Luxembourg during the meetings related to Population and Housing Censuses in 2016, 2017 and 2018.
7. Dardanelli S., Sasso A., Verrascina M., La diffusione dei dati del censimento della popolazione e delle abitazioni del 2011 alla luce di alcune novità introdotte a livello nazionale e internazionale, XXXI Conferenza Italiana di Scienze Regionali, Aosta, 20-22 Settembre 2010.
8. Falorsi S., “Census and Social Surveys Integrated System”, 2016, internal document.
9. Mastroluca S. (ed.) I contenuti informativi della rilevazione, la validazione e diffusione dei dati, Istat 2017.
10. Verrascina M., La diffusione dei dati censuari della popolazione in alcuni stati esteri: contenuti e strategie Istat Working Paper (soon to be published).

Living Conditions and Consumption Expenditure in Time of Crises

Household consumption expenditure and material deprivation in Italy during last economic crises

Spesa delle famiglie e deprivazione materiale in Italia durante le ultime crisi economiche

Ilaria Arigoni and Isabella Siciliani

Abstract

The severe economic crises affecting Italy during the last decades have had serious effects on the living conditions of residing households. Household disposable income decline was associated with a general decrease of household consumption expenditure, mainly in the period 2011-2013, and a rearrangement of its structure, with a slight increase in the expenditure share on food and non-alcoholic beverages (*foodshare*) whose increase is a well-known symptom of stronger budget constraints. However, economic hardships affected households differently, producing inequality increase and an enlargement of the share of population suffering from material deprivation.

This paper attempts to shed light on extent and the characteristics of the impact on household living conditions of the economic crises in Italy, and tries to describe the most affected subgroups of population. Results are based on the 2007-2013 data of the two main socioeconomic surveys conducted by Istat: the Household Budget Survey (HBS) and the Income and Living conditions Survey (IT-SILC).

The general framework is deepened by modelling separately the share of expenditure on food and non-alcoholic beverages and the material deprivation on a set of pooled data (in order to disentangle the time effect) and by single year.

1 Introduction

Italian economy was first affected by economic crisis in 2008, year of the initial decline of the GDP (-1.1% at constant prices), and then in 2009, when GDP suffered a heavy fall (-5.5%), followed later by other significant drops in 2012 and 2013 (-2.8% and -1.7%, respectively).

This long period of economic crisis, characterizing the last decade, has had serious effects on household living conditions. Average and median household disposable income, at constant prices, has declined, while the general reduction of available economic resources has implied a similar decay of household savings and household consumption expenditure. The latter has been rearranged in its structure, with the share of expenditure on food and non-alcoholic beverages (*foodshare*) gaining a slight increase during the crisis period, perhaps as a symptom of stronger budget constraints.

The main objective of this paper is to shed light on the extent and the characteristics of the impact of the most important economic crises of this century on the living conditions of households residing in Italy, trying to find out the features of the most affected subgroups of population. Analyses embrace the period from 2007, formally considered the pre-crisis year, until 2013, the last year of the GDP decrease. Results are based on data from the two main socioeconomic surveys conducted by Istat: the Household Budget Survey (HBS) and the Income and Living conditions Survey (IT-SILC).

[1] Ilaria Arigoni, Istat; email: ilaria.arigoni@istat.it

Isabella Siciliani, Istat; email: isabella.siciliani@istat.it

Views expressed in this article represent the opinion of the authors and do not necessarily reflect the official position of the affiliation institution.

After a general portrait of the main changes occurred in household incomes and household consumption expenditures, the paper first focuses on trends of foodshare and severe material deprivation and then on household subgroups most affected by economic crises, taking into account both the foodshare and the breadth of severe deprivation. It is worthwhile highlighting that the severe material deprivation concerns households suffering stronger economic hardships.

The identification of the main characteristics of the most vulnerable subgroups of population is carried out through linear regression models. First, by modelling the foodshare on HBS pooled data 2007-2013, a clearer picture of the factors influencing its variation (predictors) is obtained; more, controlling for household socioeconomic characteristics, the time effect during the crisis period can be isolated. Then, running separate models for each year of the above mentioned period, changes in the relationship between the foodshare and its main predictors may be investigated. In the same way, pooling the IT-SILC data 2007-2013, the household deprivation share is modelled on a set of variables (including time effects) in order to describe its main predictors and figure out, *ceteris paribus*, the impact of the time dimension. Afterwards, separate models are performed for each year, to point out possible changes in the relationship between the breadth of severe deprivation and its predictors.

2 Data and Methodology

As already mentioned above, this paper is based on data from the Household Budget Survey (HBS) and from the Income and Living conditions Survey (IT-SILC, included in the system of European Statistics on Income and Living Conditions - EU-SILC), both carried out by Istat.

The Italian HBS focuses on consumption expenditure behaviours of households residing in Italy. It analyses the evolution of level and composition of household consumption expenditure according to their main social, economic and territorial characteristics. The main focus of the HBS is therefore represented by all expenditures incurred by resident households to purchase goods and services exclusively devoted to household consumption (self-consumptions, imputed rentals and presents are included); every other expenditure for a different purpose is excluded from the data collection (e. g., payments of fees and business expenditures). The Italian HBS represents the informative base for the official estimates of relative and absolute poverty in Italy.

The EU-SILC, set up with the Regulation of the European Parliament no. 1177/2003 and first launched in 2004, is the reference source for comparative statistics on income distribution and social inclusion in the European Union. It provides both cross-sectional and longitudinal annual data on income, poverty, social exclusion and other living conditions.

As far the key indicators analysed in this paper, the foodshare is the ratio of household expenditure on food and non-alcoholic beverages to total household consumption expenditure; according to Engel's Law, "this share in the budget declines as income or total outlay increases". The assertion, made by Engel, is that "foodshare is a good indicator of welfare" (Engel, 1895). More recently, Deaton wrote that "since food is seen as the first necessity, the share of food in total expenditure can be regarded as an (inverse) indicator of welfare. It is also a very convenient indicator, since its definition as a dimensionless ratio renders it comparable over time periods and between geographical locations, at least if the relative price of food does not vary too much. However, the real interest in the food share is that it may be capable of acting as a better indicator of welfare than measures based on income or expenditure alone" (Deaton, 1981). This indicator has been largely used in studies on developing economies; however, considering the severity of the economic crises of the last decade and their tangible effects on the living conditions of households residing in Italy (at macro indicator level), in this paper it is assumed that in the period 2007-2013 the foodshare behaved as a welfare indicator although in the context of an advanced economy such as the Italian one. The foodshare indicator is expressed in percentage values.

The rate of population at risk of poverty or social exclusion (AROPE) is among the main indicators used to measure the status of economic hardship of people living in private households at EU level[2], based on EU-SILC data. It includes the share of population at risk of poverty (with an equivalised income below 60% of the median equivalised income for the population as a whole) or severely material deprived (suffering from at least four out of nine items of deprivation) or living in households with low work intensity. After first economic crisis, in Italy AROPE increased primarily for the worsening of severe deprivation, one of the two indicators considered in this work. The choice of such an indicator depends on its nature of non-relative indicator, that aims at measuring economic and financial difficulties in absolute terms, independently by changes affecting the whole distribution. That is why it is particularly appropriated for across

[2] In fact, for this indicator a precise target has also been chosen within the 2020 Europe strategy (established in 2010) to fight against poverty and social exclusion: that is to lift from this condition at least 20 million people, 2,2 million of which in Italy (see, European Commission 2010 and European Commission 2011).

time analyses. However, the severe deprivation indicator is a synthetic measure covering nine different symptoms of deprivation; once a person overcomes the cut-off threshold of four items, no further distinction is possible. Actually, a severely deprived person could suffer potentially from 4 to 9 symptoms of deprivation; therefore, in order to capture the breadth of severe material deprivation, the Alkire-Foster approach (Alkire and Foster, 2011) has been applied, based on the dual cut-off identification method. Given n individuals and d dimensions, for each dimension j a cut-off z_j is identified to establish if a person i is deprived in j-dimension, and then a cut-off k is used to determine who is multidimensional poor (in the specific case, severely deprived over nine dimensions, using k=4). Once identified the severely deprived, a censored deprivation matrix $\boldsymbol{g_0(k)}$ of n rows and d columns is constructed with the value of deprivation on each dimension only for the severely deprived, while for non-severely deprived individuals all the d values are set to zeros. The measure M_0 (adjusted headcount ratio) is the mean of the elements of the censored deprivation matrix. It can be interpreted as the total deprivations experienced by the severely deprived, divided by the maximum number of deprivations that could possibly be experienced by all people, that is nd:

$$M_o = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} g_0(k)_{ij}}{nd}$$

One of the properties of the above indicator is the decomposability across population groups and across dimensions, not feasible with the standard severe deprivation indicator. This feature allows to measure the contributions of the several symptoms of deprivation to the overall indicator across time.

In order to outline which subgroups of population suffered most from economic crises, linear regression models have been performed, first on pooled data 2007-2013 and then for each year in the same period; dependent variables are foodshare and the breadth of deprivation, considered as proxy measures of economic hardships. Models, that have run separately on HBS and IT-SILC data, have been based as much as possible on the same set of covariates[3]: place of residence (geographical area and type of municipality); household size and type; number of in-work members; age, level of education attained and activity status of the Reference Person (HBS) or Bread Winner (IT-SILC); house tenure status; quintiles of equivalent household total expenditure (HBS) or of income (IT-SILC). Models based on pooled data have shed light on time effects, while models by single year have enlightened the existence of changes in parameter estimates across time and their strength. Households are units of analysis, being all the variables considered common to all household members.

As far the model based on IT-SILC data, a couple of clarifications are needed. The breadth of deprivation is measured as the percentage share of deprivation, ranging from a minimum of zero, if a household is not deprived at all, to a maximum of 100, if a household is affected by all symptoms of deprivation. In analytical terms, the dependent variable, the household deprivation share (HDS), is the sum of the elements of the censored deprivation matrix divided by the maximum number of dimensions (multiplied by 100) for each row of the matrix[4]:

$$y_i = \frac{\sum_{j=1}^{d} g_0(k)_j}{d} * 100$$

As far the pooled data model, due to the longitudinal structure of the EUSILC survey, in order to deal with repeated measures on the same set of households and the presence of a non-negligible level of intra-class correlation (ICC=0.44) within the same households across time, a linear mixed model[5] with random effect on the intercept has been applied:

$$\left.\begin{aligned} \mathrm{Y}_{it} &= \beta_{0i} + \beta_1 X_{1it} + \cdots + \beta_H X_{Hit} + \lambda_1 \delta_{it} + \cdots + \lambda_{T-1}\delta_{iT-1} + \varepsilon_{it} \\ \beta_{0i} &= \gamma_{00} + u_{oi} \\ \beta_h &= \gamma_{0h} \quad \forall\, h \end{aligned}\right\} \Rightarrow$$

$$y_{it} = \gamma_{00} + \gamma_{01} X_{1it} + \cdots + \gamma_{0H} X_{Hit} + \lambda_1 \delta_{it} + \cdots + \lambda_{T-1}\delta_{iT-1} + u_{0i} + \varepsilon_{it}$$

$\text{with } \mathrm{i} = 1,2, \ldots \mathrm{n} \text{ households}, \; \mathrm{t} = 1,2, \ldots \mathrm{T} \text{ time} - \text{units}, \; \varepsilon_{it} = \text{error term for household } i \text{ and time } t,$
$u_{0i} = \text{ random effect for household } i.$

3 General Framework

In the observed period (2007-2013), two economic crises were registered: the first significant drop in GDP was recorded in 2009 (-5.5% compared to 2008), the second was shown in 2012-2013 (respectively, -2.8% and -1.7% compared to the previous year value).

Household income (at constant prices) in 2009 had, on average, an increase and a trend inversion compared to the previous two years; its first decrease started in fact in 2010, with one-year lag respect to the GDP. That reduction went on until 2013, showing an impoverishment of the average household economic conditions: the overall decrease, since

[3] The choice of the covariates has been based on information common to the two surveys.
[4] The mean value of this dependent variable across all population units provides the adjusted headcount ratio M_0.
[5] The MIXED Procedure included in the SAS package has been used.

2007, was -12.3% (Figure 1). This trend is not reasonably due to a decrease in household size, since for the same period the equivalised income showed the same trend, although with an overall reduction less intense (-10.8%).

During the whole period 2007-2013, household consumption expenditure (at constant prices) had a continuous decline, even sharper than the one observed for household income in the same time interval (for consumptions, the overall decrease since 2007 was -16.7%) and double than the decrease of GDP after 2011 (-8.6% versus -4.4%). Still compared to the GDP, between 2009 and 2011 household consumption expenditure did not confirm the same slight recovery: its lagging behind the economic growth was probably related to the fact that, in light of international uncertainty, households were cautious in their consumption expenditures.

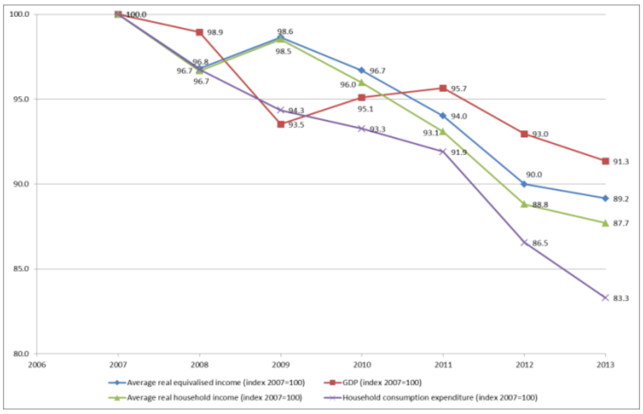


Figure 1 - GDP, household and equivalised income at costant prices, household consumption expenditure at costant prices

After first economic crisis in 2008, in Italy foodshare started increasing slightly since 2010 onwards, to reach the value of 22.4% in 2013 (Figure 2). This trend was quite differentiated within the national borders: in the North there was no dynamic at all; in the Centre, it started later but the foodshare increased from 20.4% in 2011 to 21.5% in 2013, registering the highest percentage increase from 2011 to 2013, considering the three geographical areas; in the South, where the situation is structurally worse than in the rest of the country, the foodshare started increasing since 2010 but more softly than in the Centre.

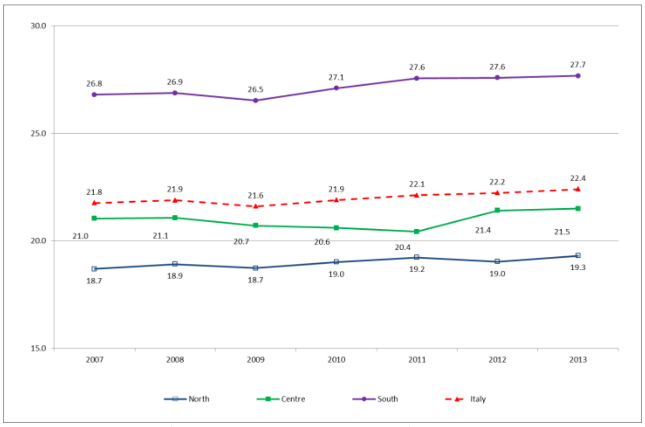


Figure 2 – Share of household expenditure on food and non-alcoholic beverages at national level and by geographical area *(percentage values)*

Focusing on the economic most disadvantaged groups, the share of population AROPE, equal to 26% in 2007, after a slight improve in the following three years, reached in 2011 the value of 28.1% and a peak of 29.9% in 2012 (Figure). Investigating the different components of the AROPE, it is clear that this big raise is due to the growth of severely deprived: the population share suffering from severe deprivation passed from 7.4% in 2010 to 11.1% in 2011, followed by a successive increase to 14.4% in 2012. Until 2010 there was not a significant increase in material deprivation, thanks to the strengthening of workers' income support measures, such as unemployment benefits and salary integration allowances, and thanks to household strategies, set up to tackle the progressive erosion of their purchasing power (drawing assets, saving less or borrowing). With the continuation of the crisis, however, in 2011 there was a strong deterioration of the situation, with an increase in the material deprivation rate, and in 2012 household economic difficulties further widened (Istat, 2014). After this year the severely deprived decreased, but never accounting less than 12%.

Even if in 2011-2012 the severe deprivation widened, it seemed to have become less hard in terms of symptoms of deprivation: in fact, the average number of deprivation items, equal to 4.57 in 2010, dropped to 4.42 in 2011 and 4.36 in

2012 (Figure), and the adjusted headcount ratio (following the Alkire-Foster methodology[6]), that takes into account not only the deprived but also the width of deprivation among the deprived, recorded a smoother increasing trend than the severe deprivation headcount ratio. The latter approach allows also measuring the contribution of the different dimensions to the overall index of severe deprivation, in order to know which one of them influences more the synthetic index, and whether its relevance has changed across time. Looking at the different dimensions (or items) of deprivation, the main contributions before and after 2011 were given by the unaffordability of an annual week holiday away from home (22%) and by the inability to face unexpected financial expenses corresponding to the monthly national at-risk-of-poverty threshold (22%). A little less than a fifth was the contribution of the inability to keep home adequately warm, while the relevance of the unaffordability of an appropriate protein meal every second day increased from 14% in 2010 to 17% in 2011 and 19% in 2012; on the contrary, being in arrears (with utility bills, rent, mortgage or other debts) weakened its significance from 16% in 2010 to 14% in 2011, and even 11% in 2012[7].

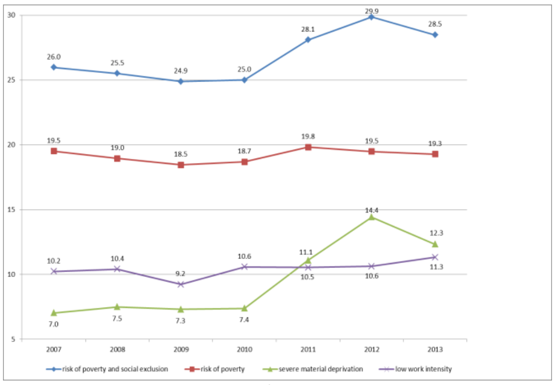


Figure 3 - Population at risk of poverty and social exclusion, at risk of poverty, severely deprived and in low work intensity households *(values per 100 individuals)*

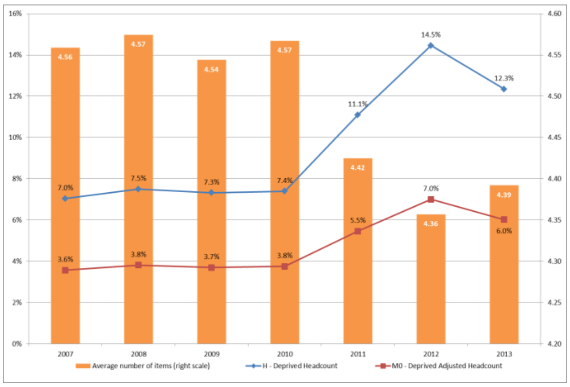


Figure 4 - Population severely deprived, adjusted headcount ratio of severely deprived *(values per 100 individuals)*, average number of items of deprivation

4 Results

As reported above, analyses presented in this paper rely on a twofold strategies to measure household economic hardships during the last economic crises. On one side, it has been hypothesized an expansion of the expenditure share on food and non-alcoholic beverages (FOSH) for stronger budget constraints, on the other a strengthening of material deprivation has been theorized, measured in terms of household deprivation share (HDS). FOSH and HDS dependent variables were regressed separately on a set of common dependent variables describing household place of residence and socioeconomic conditions[8].

[6] For a wider application of the Alkire-Foster methodology to the analysis of multidimensional poverty in the EU Countries. see Alkire and Apablaza, 2017.

[7] Taking into account the headcount ratios, from 2010 until 2012 the items of deprivation showing the biggest increases were: unaffordability of an annual week holiday away from home (40.5% in 2010, +5.94 p. p. in 2011 and +4.41 p.p. in 2012); unaffordability of an appropriate protein meal (7% in 2010, +5.72 p.p. in 2011 and +4.38 in 2012); inability to keep home adequately warm (11.6% in 2010, +6.22 p.p. in 2011 and +3.45 p.p. in 2012); inability to face unexpected financial expenses (33.8% in 2010, +4.43 p.p. in 2011 and +3.91 p.p. in 2012). The increase of the severe deprivation started in 2013 (also due to a more favourable inflation dynamics compared to 2012) concerned mostly the affordability of a protein meal every second day (-3.15 p.p.), the ability to keep home adequately warm (-2.45 p.p.) and the ability to face unexpected financial expenses (-1.91 p.p.) (Istat, 2014).

[8] Namely, covariates are: the natural logarithm of household size (LNHSIZE); the age of the Reference Person (AGERP); the education level attained by the Reference Person, distinguished in low (EDURPLOW), medium (EDURPMEDIUM) and high (EDURPHIGH); the activity status of the Reference Person, grouped in self-employed (ACTRPSELF), employee (ACTRPDEP), retired (ACTRPRET) and other than the mentioned conditions (ACTRPOTHER); the number of household members working (NHHWORKER); the household type, classified in single person less than 65 years (HHTYPE_SPLESS65), single person 65 years and over (HHTYPE_SPLEAST65), couple without children with Reference Person less than 65 years (HHTYPE_CNOCHILDLESS65), couple without children with Reference Person 65 years and over (HHTYPE_CNOCHILDLEAST65), couple with one child (HHTYPE_C1CHILD), couple with two children (HHTYPE_C2CHILDREN), couple with three or more children (HHTYPE_C3CHILDREN), single parent (HHTYPE_SPARENT), other household typologies than the mentioned ((HHTYPE_OTHER); the accommodation tenure status, if rented (HOUSE_RENTED) or owned (HOUSE_OWNED); the equivalent total expenditure/income quintile, from QUINTILE1 (the lowest) to QUINTILE5 (the highest). More, the place of household residence has been introduced in the models, taking into account: the geographical area (North, Centre, South) and the size of municipality (Metropolitan area - Centre (BIGMC), Metropolitan area suburbs and municipalities with 50,001 inhabitants and over (MEDIUMM), other municipalities until 50,000 inhabitants (SMALLM)).

Please note that: as IT-SILC reference person, the main income recipient or Bread Winner (BW) has been used; in order to allow for the remaining curvature in the relationship between the dependent variables and AGERP, the square of AGERP has been also introduced (AGERPSQUARED).

By modelling FOSH and HDS on pooled data 2007-2013, controlling for household socioeconomic characteristics, parameter estimates show a time effect during the crisis period on both dependent variables.

Table 1 - Estimates of the OLS regression model coefficients of household foodshare.

Variable	Pooled data		Y2007		Y2008		Y2009		Y2010		Y2011		Y2012		Y2013	
Intercept	5.876	**	7.276	**	6.723	**	4.000	**	6.731	**	5.098	**	5.926	**	7.538	**
Y2007	***0.000***		-	-	-	-	-	-	-	-	-	-	-	-	-	-
Y2008	0.192	*	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Y2009	-0.081		-	-	-	-	-	-	-	-	-	-	-	-	-	-
Y2010	0.201	*	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Y2011	0.428	**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Y2012	0.496	**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Y2013	0.728	**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
North	-1.545	**	-1.816	**	-1.669	**	-1.650	**	-1.292	**	-0.862	**	-1.846	**	-1.631	**
South	3.826	**	3.474	**	3.587	**	3.384	**	4.039	**	4.607	**	3.834	**	3.866	**
Centre	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
bigMC	0.480	**	-0.213		0.583	**	-0.119		0.095		0.706	**	1.028	**	1.316	**
smallM	0.619	**	0.077		0.836	**	0.422	**	0.374	**	0.959	**	0.869	**	0.788	**
mediumM	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
lnhsize	1.262	**	1.615	**	1.100	*	2.244	**	1.062	*	1.205	*	1.156	*	0.458	
ageRP	0.168	**	0.115	**	0.175	**	0.196	**	0.183	**	0.196	**	0.155	**	0.135	**
ageRPsquared	-0.001	**	-0.001	*	-0.001	**	-0.001	**	-0.001	**	-0.002	**	-0.001	**	-0.001	**
eduRPlow	4.738	**	4.696	**	4.757	**	5.181	**	4.512	**	5.299	**	4.053	**	4.553	**
eduRPmedium	2.524	**	2.566	**	2.396	**	3.196	**	2.031	**	2.649	**	2.142	**	2.650	**
eduRPhigh	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
actRPdep	0.524	**	0.425	*	0.316		0.185		0.699	**	0.506	*	0.919	**	0.758	**
actRPret	0.406	**	0.434		0.004		0.172		0.537		0.358		0.606	*	0.805	**
actRPother	0.471	**	0.038		0.281		0.621	*	0.760	*	0.001		0.832	**	0.926	**
actRPself	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
nhhworker	-1.005	**	-1.113	**	-1.157	**	-0.843	**	-1.249	**	-1.104	**	-0.732	**	-0.855	**
hhtype_spless65	1.230	**	1.206		0.231		2.098	**	0.816		1.282		2.450	**	0.586	
hhtype_spleast65	1.917	**	1.494	*	0.926		3.056	**	1.072		2.137	**	2.871	**	1.947	*
hhtype_cnochildless65	0.725	**	0.985	*	0.040		1.188	**	0.352		0.722		1.241	**	0.681	
hhtype_cnochildleast65	1.488	**	1.161	**	0.936	*	2.366	**	0.944	*	1.367	**	2.201	**	1.517	**
hhtype_c1child	0.371	**	0.421		0.344		0.669	**	-0.092		0.583	*	0.790	**	-0.204	
hhtype_c3children	0.427	**	0.365		0.263		0.434		0.139		0.203		1.239	**	0.469	
hhtype_sparent	0.331	*	0.437		-0.410		0.197		0.327		0.186		1.297	**	0.328	
hhtype_other	0.879	**	0.555		0.909	**	0.695	*	0.928	**	1.051	**	1.385	**	0.701	*
hhtype_c2children	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
house_rented	3.209	**	3.266	**	3.645	**	3.227	**	3.161	**	3.504	**	2.963	**	2.655	**
house_owned	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
quintile1	8.930	**	9.251	**	8.744	**	8.553	**	8.539	**	8.839	**	9.036	**	9.692	**
quintile2	6.885	**	7.023	**	6.649	**	6.718	**	6.688	**	6.669	**	7.200	**	7.391	**
quintile3	5.310	**	5.067	**	5.037	**	4.971	**	5.385	**	5.163	**	5.409	**	6.353	**
quintile4	3.506	**	3.454	**	3.385	**	3.464	**	3.707	**	3.474	**	3.428	**	3.717	**
quintile5	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
N	159845		24400		23423		23005		22246		23158		22933		20680	
adj-R-squared	0.2814		0.291		0.2836		0.2717		0.2748		0.2916		0.2699		0.2949	
* significant at p=0.05																
** significant at p=0.01																

For FOSH variable (Table 1, pooled data), the increase of time effect starts in 2010 (when the estimate is 0.201 higher than in the reference year 2007) to reach the value of 0.728 in 2013.

For households belonging to lower equivalent total expenditure quintiles, residing in the South, with low educated Reference Person and living in a rented accommodation, FOSH variable raises the most. Across time, the relative wider foodshare of households residing in the South rises slightly, especially in 2010 and 2011 (respectively 4.039 and 4.607 compared to 3.474 of year 2007). Since 2010, just after the first economic crisis, the distance between the better-off households of the upper quintile (fifth) and the ones in the first quintile starts increasing monotonically, and in 2013 the parameter for the lowest quintile (first) reaches the value of 9.692.

It is noteworthy mentioning that, while in the pre-crisis time no significant difference has been observed for FOSH variable between households whose reference person is self-employed and households with reference person who is an employee, a retired or another inactive, in the post-crisis phase these latter households show a feebly higher FOSH, in particular for employees from 2010 onwards, while for the retired the rearrangement of the foodshare takes place later, since 2012.

Others factors have contributed, although to a minor extent, to the increase of foodshare among households residing in Italy. As far the level of education attained by the Reference Person, parameter estimates for low and medium educated Reference Person (always significantly different from the reference category of high educated Reference Person, and always positive) reached their peaks in 2009 (respectively, 5.181 and 3.196) and 2011 (5.299 and 2.649), although for medium educated Reference Person the value observed in 2013 is the same as 2011. Looking at the household type, elderly people, single or in a couple, compared to a reference couple with 2 children, just after the first economic crisis, in 2009, worsened their conditions (parameter estimates were, respectively, 3.056 and 2.366), while during the second phase of the crisis, in 2012, they actually showed a feebly higher FOSH, which nevertheless reached values lower than 2009 (2.871 and 2.201). To be mentioned also the behaviour of other household typologies than single person and couples with or without children: since 2009 onwards their conditions slightly deteriorated, although with a lower magnitude (the peak was in 2013, when the parameter estimate was 1.385).

For HDS variable (Table 1, pooled data), again a time effect is observed; in this case, until 2010 time parameters are not significantly different from the reference year 2007, while they show a meaningful increase since 2011 (when HDS is 1.943 percentage points higher than that in 2007) to reach the highest value (3.333) in year 2012.

Table 2 Estimates of the Mixed and OLS regression model coefficients of household deprivation share

Variable	Pooled data		Y2007		Y2008		Y2009		Y2010		Y2011		Y2012		Y2013	
Intercept	-2.389	**	-3.629	**	-6.416	**	-3.610	**	-2.787	*	-0.674		-1.568		-0.892	
Y2007	***0.000***		-	-	-	-	-	-	-	-	-	-	-	-	-	
Y2008	0.082		-	-	-	-	-	-	-	-	-	-	-	-	-	
Y2009	-0.167		-	-	-	-	-	-	-	-	-	-	-	-	-	
Y2010	-0.110		-	-	-	-	-	-	-	-	-	-	-	-	-	
Y2011	1.943	**	-	-	-	-	-	-	-	-	-	-	-	-	-	
Y2012	3.333	**	-	-	-	-	-	-	-	-	-	-	-	-	-	
Y2013	2.507	**	-	-	-	-	-	-	-	-	-	-	-	-	-	
North	-0.538	**	-0.642	**	-0.479	*	-0.121		-0.555	*	-0.474		-0.897	**	0.657	*
South	3.875	**	2.962	**	3.565	**	2.503	**	2.830	**	4.308	**	5.052	**	6.036	**
Centre	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
bigMC	0.524	**	0.942	**	0.958	**	1.511	**	1.283	**	-0.745	*	-0.115		0.447	
smallM	-0.505	**	-0.922	**	-1.007	**	-0.126		-0.687	**	-0.270		-0.007		-0.393	
mediumM	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
lnhsize	0.492		1.843	**	3.453	**	0.925		0.104		0.932		1.230		-2.085	**
ageRP[a]	0.046	**	-0.010		-0.004		0.083	*	0.118	**	0.008		0.009		0.070	
ageRPsquared[a]	-0.001	**	0.000		0.000		-0.001	**	-0.002	**	-0.001	*	-0.001		-0.001	**
eduRPlow[a]	3.244	**	2.887	**	2.989	**	3.624	**	3.354	**	5.456	**	5.450	**	3.823	**
eduRPmedium[a]	1.338	**	1.319	**	0.929	**	0.953	**	1.068	**	1.807	**	2.676	**	1.815	**
eduRPhigh[a]	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
actRPdep[a]	0.449	**	0.479		0.872	**	0.420		1.019	**	0.802	*	1.126	**	1.395	**
actRPret[a]	1.059	**	1.768	**	1.859	**	1.118	**	1.572	**	0.965		1.419	**	1.782	**
actRPother[a]	4.105	**	5.058	**	5.470	**	4.042	**	4.818	**	5.061	**	6.327	**	6.519	**
actRPself[a]	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
nhhworker	-0.625	**	-0.030		0.027		-0.553	**	-0.624	**	-0.796	**	-1.044	**	-0.323	
hhtype_spless65	1.129	**	3.892	**	6.157	**	1.766	*	0.076		2.604	**	2.076	*	-1.819	
hhtype_spleast65	0.292		1.569		2.927	**	0.026		-0.534		2.377	*	1.362		-2.544	*
hhtype_cnochildless65	-0.137		0.893		2.099	**	0.087		0.000		0.401		0.054		-1.682	**
hhtype_cnochildleast65	-0.805	**	-0.044		0.884		-0.629		-1.127		0.510		-1.365		-0.964	
hhtype_c1child	0.209		0.717	*	1.140	**	0.873	*	-0.425		0.514		0.420		-0.375	
hhtype_c3children	1.281	**	0.657		0.735		1.603	**	1.413	*	2.170	**	2.200	**	2.474	**
hhtype_sparent	1.734	**	3.185	**	3.985	**	3.050	**	1.048	*	1.841	**	1.759	**	0.691	
hhtype_other	1.559	**	2.144	**	1.975	**	2.151	**	1.287	**	2.056	**	1.723	**	1.217	*
hhtype_c2children	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
house_rented	5.152	**	5.275	**	4.865	**	5.207	**	5.267	**	6.427	**	6.178	**	6.445	**
house_owned	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
quintile1	4.938	**	5.484	**	6.006	**	5.205	**	5.420	**	5.530	**	7.525	**	7.983	**
quintile2	2.163	**	1.614	**	2.065	**	1.523	**	1.303	**	1.753	**	3.323	**	3.451	**
quintile3	0.737	**	0.389		0.651	*	0.214		0.208		0.073		1.043	**	1.237	**
quintile4	0.115		0.109		0.339		-0.370		-0.135		-0.610		-0.028		0.128	
quintile5	***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***		***0.000***	
N	138972		20982		20928		20492		19147		19399		19578		18486	
adj-R-squared	-		0.1265		0.1344		0.1220		0.1289		0.1416		0.1622		0.1658	
Chi-squared Likelihood Ratio (null model with only the fixed effects)	26552.43															
Pr>Chi-Squared	<0.0001															

* significant at p=0.05
** significant at p=0.01
[a] RP, Reference Person=Bread Winner (BW)

Among the factors associated with a higher HDS, the most important are: residing in the South, living in a household belonging to lower income quintiles, with a low-educated Bread Winner or where the Bread Winner is inactive (other than retired), living in a rented accommodation.

The disadvantage of residing in the South increased monotonically from 2009 onwards: in 2009, a household living in this geographical area had a HDS 2.503 percentage points higher than a household living in the Centre, and this estimate reaches the value of 6.036 in the year 2013. Also the magnitude of low and medium educated Bread Winner effects (always positive compared to high educated BW) increased during the crisis period, especially in 2011 and 2012 (when for low educated Bread Winner the estimate was 5.45, compared to the pre-crisis value of 2.887). Deprivation for low income households increased as well: in 2007, households in the first income quintile had a HDS 5.484 percentage points higher than the ones in the fifth quintile, while in 2013 this estimate was equal to 7.983. When the Bread Winner is self-employed (reference category) households suffer less from deprivation: after the economic crises manifested their effects, the relative disadvantage of inactive (other than retired) Bread Winner widened (6.327 in 2012 and 6.519 in 2013), while the conditions of households whose Bread Winner is an employee or a retired worsened less.

Others factors that have contributed to exacerbate HDS, but to a minor extent than the above mentioned, are firstly: being a single non-elderly person, a single parent or a couple with 3 or more children. However, in these cases the relative magnitude of the effect has changed across the analysed period. In the pre-crisis time, compared to a reference couple with 2 children, being a single person less than 65 years (6.157 in 2008) or a single parent (3.985 in 2008) worsened the HDS more than in the post-crisis time (respectively, 2.604 and 1.841 in 2011). The opposite happened for household with 3 or more children: from a non-significant difference from couple with 2 children in 2008, in 2011 the parameter estimate reached the value of 2.17; from that year onwards the situation for couples with 3 or more children has continued to deteriorate monotonically, up to the estimated effect of 2.474 in 2013. Until 2010, living in metropolis was associated to a slightly major extent of the HDS than in medium towns, while afterwards a minor or not significantly different association has been observed. As far the household size, in the pre-crisis time a positive effect was observed, which surprisingly changed sign in the last year of the considered period. Although further analyses are definitely required, one possible explanation is that in bigger households other than couples with 3 or more children (that have on the contrary a positive estimate in 2013), additional household members are elderly people who are likely to be retired. In this sense, they can contribute to improve household financial resources thanks to the guarantee represented by their pension income (that, for the lower amounts, was linked to the price indexes in the considered period), leading to a lower degree of material deprivation.

5 Concluding remarks

The economic crises of the last decade have had considerable effects on household living conditions in Italy, with a sharp decline, equal to -12.3%, of the average household income, and an even higher decline (-16.7%) for the average household consumption expenditure.

Stronger budget constraints have induced a moderate raise of foodshare and an enlargement of the share of population suffering from severe material deprivation. In particular, main evidences show that households residing in the South, belonging to lower income or expenditure quintiles, having a reference person with a low level of education, living in a rented dwelling and in large households (namely, other household typologies than single person and couples with or without children and couples with three or more children) are the ones hit hardest by the economic crises of the last decades; unfortunately, these population subgroups were already the most disadvantaged in the pre-crisis time.

References

1. Alkire S., Foster J.: Counting and multidimensional poverty measurement, Journal of Public Economics Vol. 95, pp. 476-487 (2011).
2. Alkire S., Apablaza M.: Multidimensional poverty in Europe 2006-2012: illustrating a methodology. In: Atkinson A., Guio A.C., Marlier E. (eds), Monitoring inclusion in Europe, pp. 225-240. Eurostat (2017).
3. Deaton A.: Three Essays on a Sri Lanka Household Survey, Living Standard Measurement Study Working Paper No. 11. Washington, DC: The World Bank (1981).
4. Engel, E.: Die Productions-und Consumtionsverhaltnisse des Königreichs Sachsen, in Zeitschrift des Statistischen Biireaus des Koniglich Sachsischen Ministeriums des Innern, No. 8 and 9, pp. 1-54 (1857). It was reprinted as an appendix to Die Lebenskosten Belgischer Arbeiter Familien friiher und jetzt. Bulletin de I'Institut International de Statistique, No. 9. Dresden (1895).
5. European Commission, The European Platform against Poverty and Social Exclusion: A European framework for social and territorial cohesion, COM 758 final. Brussels (2010).
6. European Commission, The social dimension of the Europe 2020 Strategy A report of the Social Protection Committee. Publications Office of the European Union, Luxembourg (2011).
7. Istat, La situazione economica delle famiglie. In: Capitolo 4 - Tendenze demografiche e trasformazioni sociali: nuove sfide per il sistema di welfare, Rapporto Annuale 2014 – La situazione del Paese, pp. 173-180 (2014).

Network Data Analysis and Mining

Support provided by elderly Italian people: a multilevel analysis

L'aiuto offerto dagli anziani italiani: un'analisi multilivello

Elvira Pelle, Giulia Rivellini and Susanna Zaccarin

Abstract The characteristics of social networks determine the availability of social support, that is the aid individuals gain from their network members. Despite the literature usually investigate the importance of having support, the role of the support provided to alters had not yet received the same attention. In particular, the support elderly provide to their networks members shows an active participation in the social life, that is one dimension of the active ageing. Using data from 2009 edition of "Famiglia e Soggetti Sociali (FSS)" survey carried out by the Italian National Statistical Institute, we propose a Bayesian multilevel model to highlight the determinants of observing a provided support tie by elders to family or non-family members.

Abstract *Le caratteristiche della rete sociale in cui un individuo è inserito determinano la disponibilità di supporto ricevuto dai membri della rete. Sebbene la letteratura si concentri sull'importanza di ricevere aiuto, minore attenzione è dedicata all'analisi dell'aiuto fornito. Per gli anziani, per esempio, l'aiuto dato ai membri della loro rete è segno di un'attiva partecipazione alla vita sociale, che rappresenta una dimensione dell'invecchiamento attivo. Usando i dati 2009 dell'indagine Istat su "Famiglia e Soggetti Sociali" (FSS), si propone un'analisi multilivello per evidenziare le caratteristiche della popolazione anziana (persone di età 65 e oltre) e dei loro alter che influenzano la probabilità di dare aiuto a familiari e non familiari.*

Key words: ego-centered support network, provided and received support, Bayesian multilevel model

Elvira Pelle
Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy, e-mail: elvira.pelle@deams.units.it

Giulia Rivellini
Department of Statistical Science, Catholic University of Milan, Milan, Italy, e-mail: giulia.rivellini@unicatt.it

Susanna Zaccarin
Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy, e-mail: susanna.zaccarin@deams.units.it

1 Introduction

The characteristics of social networks and their composition determine the availability of social support, which, in turn, is defined as the aid that individuals gain from their network members [15, 5]. Whatever the type of support (emotional, informational and instrumental), a growing number of studies have documented the positive influence of social support and social network on various health outcomes and wellbeing (among others, see [17]). Social interactions have the potential to protect individuals at risk (e.g., encouraging them to develop adjustment to face the difficulties) and promote positive personal and social development; as consequence the exposure to various types of stress diminishes [14, 8] while there is an increasing of the ability of coping with it. Despite the literature usually investigate the importance of having support, especially that received from the personal network, the role of the support provided by ego to alters had not received the same attention. Focusing on elderly people, the support they provide to their networks members can be view as a sign of an active participation in the social life; the latter is one dimension of the active ageing, a multidisciplinary concept that identify both experience and capability of being autonomous in the economic, political and social life [4].

In a familistic country such Italy informal intergenerational transfers are the most important pillar of the national welfare system, replacing formal sources of support [9]. Moreover, recent analyses on the individual potential support ego (PSE) networks in Italy [2] provides evidence to the existence of sources of potential support that extend beyond the family circle.

The aim of this contribution is to study the types of support provided by the elderly to other people, stressing how features of both ego (the elderly) and alters in their network (siblings, children, grandchildren, other relatives, neighbors, friends) affect the form of aid given. Using data from Family and Social Subjects (FSS) survey carried out in 2009 by the Italian National Statistical Institute, in this study we propose a Bayesian multilevel model to highlight the determinants of observing a provided support tie by elders to family or non-family members, controlling also for the kind of PSE-network in which the elders are embedded.

The remain of the article is organised as follow: in Section 2 we present the characteristics of elderly Italian people along with the characteristics of their potential support network. In Section 3 we describe the multilevel approach used to investigate the provided support given by elderly. Section 4 ends the paper with a discussion and some concluding remarks.

2 Individual (ego) characteristics, network typologies and support exchanged

In this work we exploit data drawn from “Family and Social Subjects” (FSS) survey carried out by the Italian National Statistical Institute in 2009. Since 1998, FSS is part of the Multipurpose Survey Program on Italian households and represents the primary statistical source providing information on contacts and provided and received support by Italian individuals. 2009 FSS edition involved 43850 individuals living in about 24000 family units. We focus on Italian individuals aged 65 years and more (n=9202, 21% of the total). Looking at the gender, 57% are females, while 43% are males. Most of elders are in couple (married or unmarried) without cohabiting children (45.9%), while 15.3% are in couple with cohabiting children; 27.8% live as single (without other members), 6.3% are single-parents and 4.8% have other family typologies. More than half of the elders (51.3%) are aged 65-74ys, while the oldest-old (85+) are the 11.6% of the total. With respect to the place of residence, 42.8% of elders live in the North of Italy, 38% in the South or Islands, and the remaining 19.2% in the Center. Looking at the health, 76.7% declare to have good/satisfactory health conditions.

Following the methodological approach proposed in [1], we use data from the FSS on the presence and the contacts with siblings, children, grandchildren, other relatives, neighbors and friends, to define the potential support ego-centered (PSE)-network in which Italian elders were embedded. In particular, PSE-network is defined as "the set of not-cohabiting people (along with their role relations) who can be a possible source of support to the respondent" [1, p. 6]. The average number of alters is quite small (2.65). Some differences in the network size emerge by age classes: 48% among the oldest-old (80+) have a number of alters between 1 and 2, while 47% of the group of 65-69ys old have 2-3 alters.

We grouped the 6 identified potential alters in the ego-centered network in 3 main networks' typologies: "Immediate family" (siblings and children), "Extended family" (grandchildren and relatives) and "No family" (neighbours and friends). The most widespread network typology among the elderly is the *Comprehensive* one composed by siblings, children, grandchildren and other relatives plus friends and neighbors (30.7% of individuals aged 65+), followed by more oriented *family* network typologies with alters only from Immediate and Extended family (28%) or only children and siblings (9.9%).

FSS supplies also data on several types of provided and received support by respondents, with detailed information on the nature and the characteristics of the recipient. Looking at the support received by elderly, only 16.6% declared to have received support from a non-cohabiting people; on the other hand, about 26% of the elders declared to provide support to non-cohabiting people (55% of which are females and 45% are males) inside or outside their family circle. The provided support to the non-familiar circle represents about the 30%, with some relevant differences with respect to the place of residence (51.3% for elders living in the North of Italy, 30.6% for those in the South or Islands, while only 18.1% for those living in the Center). We highlight some differences according to the most important type of support provided to a family or non-family member: for the first category the most provided types of support are the care of children (51.6%) and monetary help (12.7%), while for the latter are keep company (23.3%) and economic help (17.4%).

3 Multilevel analysis

We propose a multilevel approach in order to analyse the network characteristics of the group of the 2386 elders who declare to provide support [10, 3]. In particular, we adopt a Bayesian multilevel analysis [6] to provide a new insight into the determinants of observing a provided support tie (our dependent variable) by elders to family or non-family members. We use a Bayesian approach mainly for two reasons: first, given the particular structure of our data, it allows a great flexibility in the estimation of multilevel models; second, it offers some advantages in terms of computational ease, as models can be easily estimated using the package “rstanarm” of the language Stan available in R.

We specify a 3-level logistic regression model for the presence of a support tie to a family member as opposed to a non-family member, where level 3 is represented by the Italian regions to account for geographical variation in providing support to family or non-family alters (denoted by k subscript), level 2 is represented by ego (denoted by j subscript) and level 1 (denoted by i subscript) by the alter. The model can be summarised as follow:

$$logit(\pi_{ijk}) = \beta_{jk} + \mathbf{x}'_{ijk}\beta \tag{1}$$

where π_{ijk} is the probability that the observed support tie between alter i and ego j within region k is to family members; β_{jk} is the intercept varying by level-two unit j and level-three unit k and $\mathbf{x}'_{ijk}\beta$ are models fixed effects, which may be characteristics of the ego, alters and the dyad ego-alter. In particular, gender, age, family typology, education, health conditions and PSE-network typology (comprehensive, family and other) are considered as egos attributes.

It is well known that homophily is an important explanatory factor for the configuration of personal networks [12, 11, 13]. To gain insight into the determinants of giving support, we also test two hypotheses: first the homophily by gender (the elders are easier willing to provide support to individuals of the same gender). Second, the type of personal network in which the elder is embedded can determine the homophily by generation: an intergenerational PSE-network facilitates the so-called intergenerational transfers; on the other hand, an intragenerational PSE-network eases intragenerational transfers. Thus, to take into account homophily, from information on alters we construct the two variables “same generation” and “same gender” (comparing the birth generation and gender of both ego and alter, where available).

To compare models, we use the leave-one-out information criterion (looic), that uses the log-likelihood evaluated at the posterior simulations of the parameter values [7, 18]. Note that the lower the value of looic, the higher the fit of the model.

First of all, we compare the single alter-level null model and the 3-level null model. The improvement in terms of looic (from looic=3831.3 to looic=3679.3) indicates that the multilevel approach we propose is suitable to investigate our data structure.

We consider 3 models: model 1 with only characteristics of ego; model 2 with characteristics of ego and homophily terms for generation and gender and model 3 including in model 2 the type of support provided. As usual in Bayesian analysis, we monitored the Markov Chain convergence through the Gelman-Rubin statistic $\hat{R}$ [6]: chains convergence has been reached for all the estimated models, since the value of such statistic is below the recommended value of 1.1. Table 1 summarises the results.

Table 1 Results of Bayesian multilevel models: posterior quantiles at 50%; 2.5% and 97.5% in brackets

	Model 1		Model 2		Model 3	
	(looic: 3473.3)		(looic: 2858.7)		(looic: 2344.8)	
	median		median		median	
(Intercept)	0.2	(-0.7;1.1)	-0.3	(-1.2;0.6)	1.5	(0.7;2.5)
Gender (cat ref: Male)						
Female	0.0	(-0.3;0.3)	-0.2	(-0.5;0.1)	-0.4	(-0.7;-0.1)
Living arrangement (cat ref: Other)						
Couple with cohabiting children	0.5	(-0.2;1.3)	0.4	(-0.4;1.1)	0.3	(-0.4;1.1)
Couple without cohabiting children	1.0	(0.3;1.7)	1.0	(0.3;1.8)	0.8	(0.1;1.5)
Single-parent	-0.1	(-1.0;0.8)	-0.3	(-1.2;0.6)	-0.1	(-0.9;0.8)
Single	0.0	(-0.7;0.8)	0.1	(-0.6;0.8)	0.2	(-0.5;0.9)
Age (cat ref: 65-69)						
70-74	-0.3	(-0.6;0.1)	-0.1	(-0.4;0.2)	-0.1	(-0.5;0.2)
75-79	-0.4	(-0.8;0.0)	-0.4	(-0.8;0.0)	-0.2	(-0.6;0.1)
80-84	-1.2	(-1.7;-0.7)	-1.0	(-1.5;-0.5)	-0.6	(-1.1;-0.1)
85+	-1.1	(-1.8;-0.4)	-1.1	(-1.9;-0.4)	-0.7	(-1.4;0.0)
Health conditions (cat ref: Good)						
Bad	0.1	(-0.3;0.5)	0.0	(-0.4;0.4)	0.2	(-0.2;0.6)
Satisfactory	0.1	(-0.2;0.4)	0.0	(-0.3;0.3)	0.0	(-0.3;0.3)
PSE Network (cat ref: Comprehensive)						
Family	0.4	(0.1;0.7)	0.4	(0.1;0.7)	0.3	(0.1;0.7)
Other	-1.6	(-2.0;-1.2)	-1.7	(-2.1;-1.3)	-1.0	(-1.3;-0.6)
Education	0.1	(0.1;0.2)	0.1	(0.0;0.2)	0.1	(0.1;0.2)
Homophily of ego-alter						
Same generation			6.4	(4.6;9.0)	6.8	(5.1;9.2)
Same gender			7.0	(5.4;9.6)	7.1	(5.4;9.5)
Type of support (cat ref: kid care)						
Companionship					-3.7	(-4.4;-3.1)
Material/Other					-3.1	(-3.7;-2.6)
Economic					-2.5	(-3.1;-2.0)

Model 1 provides a good improvement in the fit with respect to the null model (looic= 3473.3). Elders who live in couple without cohabiting children are more likely to provide support to a family member if compared with other living arrangements, while this probability decreases as the age of ego increases. With respect to egos which can count on a *comprehensive* PSE-network, egos embedded in a *family* network typology tend to have a greater probability of a support tie to a family member, while this is less likely for ego embedded in other kinds of network. The perceived health conditions as well as the gender of the ego do not have an impact on the probability of a support tie to a family member.

According to the looic measure, adding homophily (model 2) terms to model 1 results in a remarkable improvement in the model fit (looic=2858.7, compared with the previous value of 3473.3). Parameter estimates for both generation and gender homophily have an impact on the probability of a family support tie; in particular, it is more likely to observe a support tie to a family member when ego and alter belong to the same generation as well as when they are of the same gender.

Considering also the type of support provided by ego (model 3), the model fit is still improved, with a looic=2344.8. According to the model estimates, with respect to provide support for kid care, it is less likely that other types of support (such as companionship, economic and other material support) are provided to a family member.

4 Conclusions

We proposed a Bayesian multilevel analysis of support ties provided by elders to family or non-family members. Some differences in the probabilities of ties to family alters compared with non-family alters can be noted: older elderly are less likely to provide support tie to their family circle as well as among the types of support the kid care is the more likely to be provided to a family member. Homophily between ego and alters appears to be an important explanatory factor in providing support, in particular with respect to gender and generation as revealed by our results. This can be interpreted as an evidence of a positive disposal elders have (or, more in general, people have) to provide aid to their family members also on the basis of same life experiences and attitudes.

References

1. Amati V., Rivellini G., Zaccarin S.: Potential and effective support networks of young Italian adults. Soc Indic Res, **122**, 807-831 (2015)
2. Amati V., Meggiolaro S., Rivellini G., Zaccarin S.: Relational Resources of Individuals Living in Couple: Evidence from an Italian Survey. Soc Indic Res, **134**, 547-590 (2017)
3. Bilecen B., Cardona A.: Do transnational brokers always win? A multilevel analysis of social support, Social Networks, **53**, 90–100 (2018)
4. Boudiny K.: Active ageing: from empty rethoric to effective policy tool. Ageing & Society, **33**, 1077–1098 (2013)
5. Dykstra, P. A., Bühler, C., Fokkema, T., Petrič, G., Platinovšek, R., Kogovšek, T., et al.: Social network indices in the Generations and Gender Survey: An appraisal. Demographic Research, **34**, 995-1036 (2016)
6. Gelman, A., Hill J.: Data analysis using regression and multilevel/hierarchical models. Cambridge university press (2006).
7. Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for Bayesian models. Statistics and computing, **24(6)**, 997–1016 (2014)

8. Halpern, D. Social capital. Cambridge: Polity Press (2005)
9. He W, Goodkind D, Kowal P.: An Aging World: 2015. International Population Reports. P95/16-1. Washington, DC:U.S. Census Bureau (2016)
10. de Miguel Luken V., Tranmer M.: Personal Support Networks of Immigrants to Spain: a Multilevel Analysis. Social Networks, **32**, 253–262 (2010)
11. Louch, H.: Personal network integration: transitivity and homophily in strongties relations. Social Networks, **22**, 45-64 (2000)
12. Marsden, P.V.: Homogeneity in confiding relations. Social Networks, **10**, 57-76 (1988)
13. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Annual Review of Sociology, **27**, 415-444 (2001)
14. Myers, D. G.: The funds, friends, and faith of happy people. American Psychologist, **55**, 56–67 (2000)
15. Song, L., Son, J., and Lin, N.: Social support. In: J. Scott & P. J. Carrington (eds.) The Sage Handbook of Social Network Analysis, 116–128. London: Sage Publication (2011)
16. Stan Modeling Language. Users Guide and Reference Manual. Stan Version 2.17.0 (2017) http://mc-stan.org/users/documentation/
17. Taylor, S.E.: Social Support. In: H.S. Friedman & R. Cohen Silver (eds). Foundations of Health Psychology, 145-171. Oxford: Oxford University Press (2007)
18. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, **27**, 1413– 1432 (2017)

Data mining and analysis of comorbidity networks from practitioner prescriptions

Il trattamento e l'analisi delle reti di comorbilitá delle prescrizioni dei medici generici

Giancarlo Ragozini, Giuseppe Giordano, Sergio Pagano, Mario De Santis, Pierpaolo Cavallo

Abstract For the present paper, the administrative databases of general practitioners were mined for healthcare systems analysis. A sample of 14,958 patients along with their 1,728,736 prescriptions were considered over a time span of eleven years. From this database, we derived a set of comorbidity networks by considering pathologies as nodes and the co-occurrences of two pathologies to the same prescription as links. The aim is to mine the complexity of this information by using network analysis techniques. Specifically, Islands algorithm method is well-suited to extract the most relevant and connected parts of large and dense networks, as in the present case. Main comorbidity patterns are discussed, along with future lines of research.

Abstract *In questo lavoro si mostra come un database amministrativo di un insieme di medici di base può essere considerato per un'analisi del sistema sanitario. Il database contiene un campione di 14,958 pazienti con le loro 1,728,736 prescrizioni mediche registrate in un arco temporale di 11 anni. Utilizzando i codici delle patologie contenute nelle prescrizioni, è possibile costruire la rete di comorbidità. Tale rete, densa di legami e con un numero elevato di nod, richiede metodologie di analisi appropriate. L'uso dell'algoritmo delle Islands è proposto per estrarre le principali strutture di associazione fra coppie di patologie.*

Key words: Community detection, Islands, Large network, Comorbidity pattern

Giancarlo Ragozini
Department of Political Science, University of Naples Federico II e-mail: giragoz@unina.it

Giuseppe Giordano
Department of Economics and Statistics, University of Salerno e-mail: ggiordan@unisa.it

Sergio Pagano
Department of Physics E.R. Caianiello, University of Salerno e-mail: spagano@unisa.it

Mario De Santis
Cooperativa Medi service Salerno e-mail: mariodesantis@osservatoriosanitario.it

Pierpaolo Cavallo
Department of Physics E.R. Caianiello, University of Salerno e-mail: pcavallo@unisa.it

1 Introduction

Nowadays, the presence of patients affected by many different diseases at the same time is becoming a major health and societal issue. In the United States, for instance, 80% of the health budget is spent on patients with four or more diseases (8). In clinical literature, this phenomenon is known as comorbidity.

The idea of comorbidity has been around since 1921 (9) used in a positive or negative sense, with the terms ”syntropy” and ”dystropy” (10). The former is the mutual disposition, or the appearance of two or more diseases in the same individual, while the latter indicates those pathologies that are rarely found in the same patient at the same time. In addition, the literature differentiates two core concepts, comorbidity and multimorbidity: the latter is defined as the coexistence of two or more long-term conditions in an individual not biologically or functionally linked (8), while the former concept refers to a coexistence of conditions that are linked, either biologically or functionally (12).

Simply considering this taxonomy, the intrinsically complex nature of comorbidity can be easily understood. On this basis, the burden of morbidity and comorbidity is considered to be influenced by a number of factors –namely, health-related, socioeconomic, cultural, environmental, and behavioral characteristics– but there is a lack of agreement (4) on how to understand the complex interdependent relationships between diseases due to: 1) a large number of variables (many of which are latent), 2) a lack of accuracy in measurements, and 3) technological limitations in generating data.

In this paper, we propose an indirect approach for a large-scale study of comorbidity patterns based on the administrative databases of prescription data from general practitioners (GPs), without the necessity of a complex clinical study. This methodology could be easily replicated. Access to prescription data from GPs is relatively simple, as these data are used for administrative purposes by the national health system. Given this kind of data, a morbidity state is associated with a patient and such a state is considered both over time for the same subject and within different categorized subjects. A recent literature review (13) has shown that comorbidities can be studied from GP databases based on both diagnoses, using the International Classification of Diseases –ICD– codes, and medications, using pharmacy data.

The paper is organized as follows. Section 2 presents the data and how they are used to define the comorbidity networks, along with an overview of the methodology we used. In Section 3 we report the first results, while the final section is devoted to a brief discussion with the future lines of research.

2 Data and methodology

2.1 Data

The Electronic Health Recordings (EHR) of the prescriptions made by a group of ten GPs belonging to the Cooperative Medi Service and operating in a town in Southern Italy were analyzed. A total number of 14,958 patients, covering a time interval of eleven years from 2002 to 2013, was considered. The total number of analyzed prescriptions was 1,728,736. The data were provided in anonymous form, for both patients and GPs, in accordance with Italian law on privacy and the guidelines of the Declaration of Helsinki. This retrospective observational study involved data mining and analysis of comorbidity networks from practitioner prescriptions, with data analyzed in aggregate form, and the relevant Ethics Committee granted approval (Comitato Etico Campania Sud, document number 59, released on 2016-06-08).

After a patients visit to a GP, there is generally a prescription containing a series of items of various types: drugs, laboratory tests, imaging tests, etc. For each patient visit, the GP administrative prescription data provided the following information:

- patient ID: a unique random number assigned to the patient;
- demographic data: age and sex;
- prescription date;
- prescription type: drug, laboratory test, imaging, specialist referral, hospitalization;
- prescription code: a specific code for each prescription type;
- associated ICD diagnostic code: the pathology connected to the specific prescription.

In accordance with the Italian National Health System rules, each item present in a GP prescription has an associated possible disease, encoded using the International Classification of Diseases, Ninth Revision, Clinical Modification –ICD-9-CM.[1] The ICD is the standard diagnostic tool for epidemiology, health management, and clinical purposes and is maintained by the World Health Organization. This includes the analysis of the general health situation of population groups. It is used to monitor the incidence and prevalence of diseases and other health problems, proving a picture of the general health situation of countries and populations. Each code has the general form xxx.yy, where xxx is the general disease and yy is a specific occurrence. For example, 250 is the code for ”Diabetes”, and 250.91 is the code for ”Diabetes Type 1 (juvenile) with unspecified complications, not stated as uncontrolled”.

[1] For details see http://www.salute.gov.it/portale/home.html.

2.2 Network data definition

To analyze the comorbidity patterns, we defined a set of networks. In the first step, a two-mode network was derived by considering the ICD9CM diagnostic codes and the prescriptions as the two disjoint sets of nodes. They are linked if corresponding codes appear in prescriptions made to the same patient on the same day. The sex and age of patients, and type and time of prescriptions can be considered as attributes of a given prescription.

Formally, the two-mode network can be represented as a bipartite graph $\mathscr{B}$ consisting of the two sets of relationally connected nodes and can be represented by a triple $\mathscr{B}(\mathscr{V}_1, \mathscr{V}_2, \mathscr{L})$, with $\mathscr{V}_1$ denoting the set of ICD-9-CM codes, $\mathscr{V}_2$ the set of prescriptions, and $\mathscr{L} \subseteq \mathscr{V}_1 \times \mathscr{V}_2$ the set of ties.

Being interested in the association of pathologies, we derived the one-mode network of the ICD-9-CM codes by projecting the two-mode network. The corresponding graph will be represented by $\mathscr{G}(\mathscr{V}_1, \mathscr{E}, \mathscr{W})$, with $\mathscr{V}_1$ the set of ICD-9-CM codes, $\mathscr{E} \subseteq \mathscr{V}_1 \times \mathscr{V}_1$ the set of edges, and $\mathscr{W}$ the set of weights, $w : \mathscr{E} \to \mathscr{N}$, $w(v_{1i}, v_{1j}) =$ the number of times that two ICD-9-CM codes appear in the same prescription.

In addition, different networks were generated by selecting a subset of the original dataset according to the patients sex, age intervals (0-15, 15-30, 30-45, 45-60, 60-75, 75-105), and type of prescription.

2.3 The methodology

The one-mode comorbidity network $\mathscr{G}$ is a large and dense network, and it is sometimes difficult to extract the main relevant comorbidity patterns. In order to identify the most relevant and connected parts of the network, we used the Islands approach (3), which is an algorithm useful for finding important parts in large networks with respect to a given property of nodes or lines (edges). By representing a given or computed value of nodes/lines as a height of nodes/lines and by immersing the network into water up to selected level, the islands are derived (2). Given a threshold value t, it is possible to obtain a cut of the network $\mathscr{G}(t)$. By varying this level t, different islands are identified. For our purposes, we decided to use the line island approach, which looks for a connected subnetwork with several nodes in a specified interval, such that the edges inside the island have a higher weight than the edges connecting nodes in the island with their neighbors. Formally, a set of nodes $\mathscr{I} \subseteq \mathscr{V}_1$ is a regular line island in the network $\mathscr{G}$ if:

$$\max_{(v_{1i}, v_{1j}) \in \mathscr{E}, v_{1i} \notin \mathscr{I}, v_{1j} \in \mathscr{I}} w(v_{1i}, v_{1j}) \leq \min_{(v_{1i}, v_{1j}) \in \mathscr{T}} w(v_{1i}, v_{1j}),$$

where $\mathscr{T}$ is the spanning tree over $\mathscr{I}$.

In the following, we first have deleted all nodes with a degree lower than two (representing rare diseases) and then we have computed the islands with a minimum size of 2 and maximum size of 20.

3 First results

By using the island algorithm implemented in Pajek software (6) on the full one-mode network including all epidemiological codes and all patients, 28 islands are obtained (Figure 1): one large group consisting of 20 ICD-9-CM codes and 27 islands of size 2-4 nodes. All the islands are included in one large component. As for comorbidity, the largest island includes cardiovascular disease, hypertension, osteoarthritis and osteoporosis, diabetes, prostatic hypertrophy, atherosclerosis, carotid artery stenosis, kidney disease, cystitis, and thyroiditis. In the smaller islands, we found the check up and treatment for pregnancy and infertility, as well as links among different kinds of cancer. It seems that there is a strong comorbidity structure that includes a set of diseases already linked in the clinical studies (e.g. cardiovascular disease and hypertension, diabetes, and so on). Others are probably related to the patient's age (diseases of the elderly such as arthritis, osteoporosis, prostatic hypertrophy) or to the patient's gender (e.g., pregnancy and infertility).

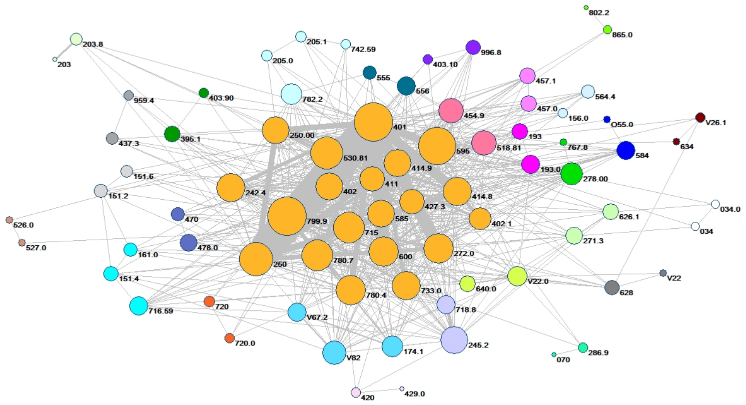
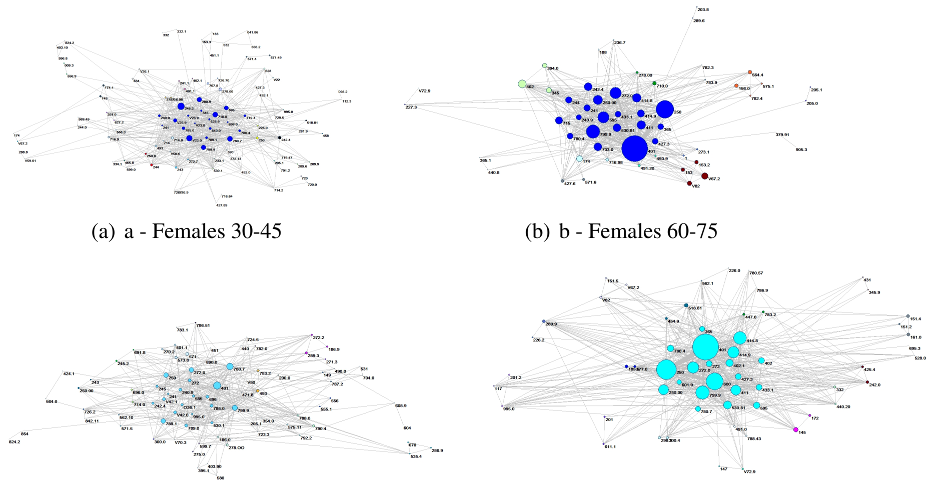


Fig. 1 Islands of size 2-20 in the comorbidity network. Node color= islands' partition; node size= nodes' degree.

Based on the data available, it is also possible to explore the effect of sex and age on the comorbidity network structure. To do that, the two-mode network can be divided into subnetworks according to these attributes associated with the prescriptions, and then specific one-mode networks can be derived and analyzed. By comparing the results of the islands algorithm for males and females divided by different age groups, different comorbidity patterns appear. For example, in the core of the main island of young females (Figure 2a), we found thyroiditis, gynecological problems, pregnancy, menstrual cramps, and cystitis; while on the periphery of the island, obesity, lipidosis, breast and thyroid cancer, arthritis and osteoarthritis were found. For older men (Figure 2d), in the core of the largest island we found arterial hypertension, prostatic hypertrophy, diabetes, heart disease, renal colic, and bronchitis; while, on the periphery, we found periodic check up after cancer, psychosis and depression, glaucoma, prostate cancer, and diverticula.



(a) a - Females 30-45

(b) b - Females 60-75

(c) c- Males 30-45

(d) d - Males 60-75

Fig. 2 Islands of size 2-20 in the comorbidity networks of females and males in the two age intervals. Node color= islands' partition; node size= nodes' degree.

4 Further developments

As the first results have shown, the prescription database is a very rich source of information about people, diagnosis prevalence, and temporal emergence of diagnosis as a proxy of incidence index. The possibility of handling this kind of data in terms of networks seems very promising (5). The graph representation allows many algorithmic tools from the network domain. Graphs can be built on patients, prescriptions, or diagnosis, according to different aims and kinds of investigation. For instance, a graph whose nodes are patients and has links established as common diagnosis or prescriptions could be used to predict diagnosis as new links between nodes. Alternatively, building the graph on prescriptions tied by common diagnoses allows us to highlight specific occurrences of comorbidity. In both cases, network techniques can be exploited to extract relevant information. Looking for cohesive sets of patients and diagnoses calls for community detection methods. Such communities are dense parts of a graph that could reveal important features like frequent patterns of comorbidity.

Beyond the Islands algorithm method here proposed for the treatment of large and dense comorbidity networks, another promising technique we are investigating for this case study is the extraction of the association rules for bipartite network data (1). This technique is strictly related to the aim of finding frequent item sets in a large dataset and commonly applied in transactional data for marketing strategy. Applying association rules in medical diagnosis can be used to assist physicians in making a diagnosis. Even if reliable diagnostic rules are difficult and may result in hypotheses with unsatisfactory predictions, which are too unreliable for critical medical applications (7). Serban et al. (11) has proposed a technique based on relational association rules and supervised learning methods. It helps to identify the probability of illness in a certain disease.

In the case of medical prescriptions, the collection of prescriptions per diagnosis can be arranged, as defined above, in the bipartite graph $\mathscr{B}$ and arranged in a transaction matrix where each prescription is a transaction defined on the set $\mathscr{V}_2$, so that $\mathscr{V}_1$ is the universal set of items and each prescription in $\mathscr{V}_2$ is a subset of $\mathscr{V}_1$. In this context, the aim stated in the association rules setting is to find set of diagnosis that are strongly correlated in the prescription database and is coherent with the notion of close neighbors in bipartite graphs. In this context, the aim stated in the association rules setting is to find a set of diagnoses that is strongly correlated in the prescription database and is coherent with the notion of close neighbors in bipartite graphs.

The usual metrics derived from the association rules context –*Support*, *Confidence*, and *Lift*– will be used to characterize the graph representation of the diagnosis item set, defined as the projection of the bipartite graph induced by the prescription database. For the sake of computational efficiency, the analysis could concentrate on the subset of diagnosis having a minimum Support *S*. On the other hand, if the graph is partitioned in *k* disjoint communities (islands or blocks), it makes sense to apply the association rules search on the separate subgraph induced by the partition. The more meaningful rules are usually sorted by decreasing the value of *Lift*. The first important rules among diagnosis will help to uncover association between frequent patterns of diagnosis co-occurrence in the whole set of prescriptions.

References

[1] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216. ACM, New York (1993)

[2] Batagelj, V.: Social network analysis, large-scale. In: Encyclopedia of Complexity and Systems Science, pp. 8245-8265. Springer, New York. (2009)

[3] Batagelj, V., Doreian, P., Ferligoj, A., Kejzar, N., : Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution (Vol. 2). John Wiley & Sons, United Kingdom (2014)

[4] Capobianco, E., Lio, P.: Comorbidity: a multidimensional approach. Trends Mol Med **19**, 515–521 (2013)

[5] Cavallo, P., Pagano, S., Boccia, G., De Caro, F., De Santis, M., Capunzo, M.: Network analysis of drug prescriptions. Pharmacoepidemiol Drug Saf **22**, 130–137 (2013)

[6] De Nooy, W., Mrvar, A., Batagelj, V.: Exploratory social network analysis with Pajek (Vol. 27). Cambridge University Press, Cambridge (2011)

[7] Gamberger, D., Lavrac, N., Jovanoski, V.: High confidence association rules for medical diagnosis, In: Proceedings of IDAMAP99, pp. 42–51 (1999)

[8] Mercer, S.W., Smith, S.M., Wyke, S., O'dowd, T., Watt, G.C.: Multimorbidity in primary care: developing the research agenda. Family Practice **26**, 7980 (2009) Available via DIALOG.
https://academic.oup.com/fampra/article/26/2/79/2367540. Cited 07 May 2018

[9] Pfaundler, M. and von Seht, L.: Uber Syntropie von Krankheitszustanden, Z. Kinderheilk. vol. **30**, 298–313 (1921)

[10] Puzyrev, V.P.: Genetic Bases of Human Comorbidity. Genetika **51**, 491–502 (2015)

[11] Serban, G., Czibula, I. G., Campan A.: A Programming Interface for Medical diagnosis Prediction. Studia Univ. Babes-Bolyai, Informatica **LI**, 21–30 (2006)

[12] Valderas, J.M., Starfield, B., Sibbald, C., Salisbury, M. Roland: Defining comorbidity: implications for understanding health and health services. Ann Fam Med, **7**, 357–63 (2009)

[13] Yurkovich, M., Avina-Zubieta J.A., Thomas J., Gorenchtein M., Lacaille D.: A systematic review identifies valid co-morbidity indices derived from administrative health data. J Clin Epidemiol **68**, 3–14 (2015)

Overlapping mixture models for network data (`manet`) with covariates adjustment

Modelli di mistura a gruppi sovrapposti per dati network (`manet`) con covariate

Saverio Ranciati and Giuliano Galimberti and Ernst C. Wit and Veronica Vinciotti

Abstract Network data often come in the form of *actor-event* information, where two types of nodes comprise the very fabric of the network. Examples of such networks are: people voting in an election, users liking/disliking media content, or, more generally, individuals - *actors* - attending events. Interest lies in discovering communities among these actors, based on their patterns of attendance to the considered events. To achieve this goal, we propose an extension of the model introduced in [5]: our contribution injects covariates into the model, leveraging on parsimony for the parameters and giving insights about the influence of such characteristics on the attendances. We assess the performance of our approach in a simulated environment.

Abstract *I dati network vengono spesso strutturati sotto forma di informazioni* attore-evento*, ovvero network dove esistono due tipologie di nodi. Alcuni esempi sono: persone che votano durante le elezioni, utenti che esprimono preferenza o meno su contenuti multimediali, o, più in generale, individui -* attori *- che partecipano a eventi. L'interesse risiede nel rilevare la presenza di gruppi fra questi attori, cluster che si differenzino per la propensione nel partecipare agli eventi in esame. A tale scopo, proponiamo un'estensione del modello introdotto in [5]: il nostro contributo contempla la presenza di covariate nel modello, sfruttando quindi un*

Saverio Ranciati
Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, Bologna, Italy, e-mail: saverio.ranciati2@unibo.it

Giuliano Galimberti
Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, Bologna, Italy, e-mail: giuliano.galimberti@unibo.it

Ernst C. Wit
Johann Bernoulli Institute for Mathematics and Computer Sciences University of Groningen, 9747 AG Groningen, The Netherlands, e-mail: e.c.wit@rug.nl

Veronica Vinciotti
Department of Mathematics, Brunel University, London UB83PH, UK,
e-mail: veronica.vinciotti@brunel.ac.uk

approccio parsimonioso e dando potenziali informazioni sull'effetto delle caratteristiche considerate. Valutiamo la performance del nostro approccio in un ambiente simulato.

Key words: Bayesian inference, bimodal network, MCMC, probit regression

1 Introduction

Network data are becoming increasingly available and a propelling force in the pursuit of new methodological approaches devoted to analyze the complexity behind interactions among units in a system. A review about the methods and models adopted in this research area can be found in [3]. Some of these data come in the form of individuals attending events, or, more generally, a network structure where two different types of nodes exist: these are also called two-mode networks, bimodal networks, or affiliation networks [6, Chapter 8]. We focus on those data describing people's behavior with respect to attending or not a set of events, and we aim to discover if communities exist within the network itself, communities that differ in patterns of preferences to attend each event. A recent approach to do model-based clustering in this context was proposed by [5]: motivated by a dataset about terrorists participating to meetings and bombings, the authors introduced a mixture model for network data called `manet`, where each unit is allowed to potentially belong to more than one community. We build on their contribution and propose an extension of their model, in order to accomodate for external information about the network, in the form of covariates describing characteristics of the units or the events.

The main contributions of the paper are:

- extending `manet` [5] by introducing covariates into the model formulation;
- eliciting how some existing regression techniques can be used in the covariates-adjusted `manet` approach;
- providing results on a simulation study about the performances of our proposed model.

The remainder of the manuscript is organized as follows: in Section 2, first we outline `manet` original formulation, to familiarize the reader with the model's structure, and then we introduce the proposed extension; in Section 3, performance of `manet` with covariates adjustment is explored in a simulated environment; finally, in Section 4, we discuss the contribution and hint at future research trajectories.

2 Covariates adjustment for `manet`

Network data are organized in an $n \times d$ matrix Y of observations y_{ij}, collecting the attendances of $i = 1,\dots,n$ units - also called *actors* - to $j = 1,\dots,d$ events. Each realization y_{ij} comes from a binary random variable, with $y_{ij} = 1$ meaning individual i attends event j, and zero otherwise. We want to cluster these n actors based on their attendances via a model-based framework, and mixture models prove to be a suitable approach to achieve the task [1]. In the traditional setting, clusters are mutually exclusive and have (prior) sizes given by $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_K)$. Usually, two conditions hold: (i) $\alpha_k \geq 0$, for each k; (ii) $\sum_{k=1}^{K} \alpha_k = 1$. Mixture models also have a hierarchical representation, attainable after introducing a unit-specific latent variable $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$: if actor i belongs to cluster k, the vector is full of zeros except for the k-th element $z_{ik} = 1$. Given the binary nature of response variables y_{ij}, for each actor $i = 1, \dots, n$ we have

$$\begin{aligned} \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}; a_1, \dots, a_K) \\ \mathbf{z}_i &\sim \text{Multinomial}(\mathbf{z}_i; \alpha_1, \dots, \alpha_K) \\ \mathbf{y}_i | (\mathbf{z}_i, \boldsymbol{\pi}) &\sim \prod_{k=1}^{K} \left(\prod_{j=1}^{d} \pi_{kij}^{y_{ij}} (1 - \pi_{kij})^{1-y_{ij}} \right)^{z_{ik}} \end{aligned} \tag{1}$$

for some hyper-parameters $(a_1, \dots, a_K)$; $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{id})$ is the attendance profile of the i-th actor to the d events, which we assume - given $\mathbf{z}_i$ - to be independent for all $j, j' = 1, \dots, d$ and $j \neq j'$. Vector $\boldsymbol{\pi}$ collects probabilities of attendance π_{kij} of a saturated model specification.

In many cases, one is interested in groups that are not mutually exclusive, allowing an *actor* to be allocated simultaneously to potentially more than a single cluster. We build on the approach suggested by [5], where a Bayesian **m**ultiple **a**llocation model for **net**work data (manet) is proposed. In manet, the hierarchical model in Equation 1 is modified by relaxing conditions on the cluster sizes $\boldsymbol{\alpha}$ and allocation vectors $\{\mathbf{z}_i\}$, allowing each actor to potentially belong to any number of the K clusters. The number of all possible group-allocating configurations is equal to $K^\star = 2^K$. Instead of working with the latent variables $\mathbf{z}_i$, a new $K^\star$-dimensional allocation vector $\mathbf{z}_i^\star$ is defined for each i. This vector satisfies $\sum_{h=1}^{K^\star} z_{ih}^\star = 1$, and a 1-to-1 correspondence exists between $\mathbf{z}_i$, which allocate actors into overlapping *parent* clusters, and $\mathbf{z}_i^\star$, which allocates actors into non-overlapping *heir* clusters. To this re-parametrization corresponds, in manet, the following hierarchical model

$$\begin{aligned} \boldsymbol{\alpha}^\star &\sim \text{Dir}(\boldsymbol{\alpha}^\star; a_1, \dots, a_{K^\star}), \\ \mathbf{z}_i^\star | \boldsymbol{\alpha}^\star &\sim \text{Multinom}(\mathbf{z}_i^\star; \alpha_1^\star, \dots, \alpha_{K^\star}^\star), \\ \mathbf{y}_i | \mathbf{z}_i^\star, \boldsymbol{\pi} &\sim \prod_{h=1}^{K^\star} \prod_{j=1}^{d} \Big[\text{Ber}\big(y_{ij}; \pi_{hij}^\star\big) \Big]^{z_{ih}^\star}. \end{aligned} \tag{2}$$

with prior $\pi_{kij} \sim \text{Beta}(\pi_{kij}; b_1, b_2)$, and (b_1, b_2) suitable hyper-parameters. For example, when $K = 2$, actor i may be assigned:

- to none of the two clusters, $\mathbf{z}_i = (0,0)$, corresponding to $\mathbf{z}_i^\star = (1,0,0,0)$;
- only to the first *parent* cluster, $\mathbf{z}_i = (1,0)$, corresponding to $\mathbf{z}_i^\star = (0,1,0,0)$;
- only to the second *parent* cluster, $\mathbf{z}_i = (0,1)$, corresponding to $\mathbf{z}_i^\star = (0,0,1,0)$;
- both of them $\mathbf{z}_i = (1,1)$, corresponding to $\mathbf{z}_i^\star = (0,0,0,1)$.

The advantage of working with re-parametrization in Equation 2 is that $\{\pi_{hij}\}$ are not additional parameters to be sampled, but probabilities of attendances produced by $\boldsymbol{\pi}$. For each actor i and event j, $\pi_{hij}^\star$ are computed via a function $\psi\big(\boldsymbol{\pi}_{\cdot ij}, \mathbf{z}_i\big)$, so that we obtain $\pi_{hij}^\star$ by looking at which parent clusters originated h, through the vector $\mathbf{z}_i$, and combining their corresponding probabilities $(\pi_{1ij}, \dots, \pi_{Kij})$. We consider $\psi(\cdot) \equiv \min(\cdot)$. For the simple case where $K = 2$, an actor i belonging to both clusters, $\mathbf{z}_i = (1,1)$, deciding whether to attend an event j or not, will do so with probability $\pi^{\star}_{hij} = \psi(\pi_{1ij}, \pi_{2ij}) = \min(\pi_{1ij}, \pi_{2ij})$. When $\mathbf{z}_i = (0,0)$, $\pi^{\star}_{hij} = 0$. The saturated manet demands inference on $(K \times n \times d)$ probabilities of attendance, where each π_{kij} has only one observation to update the prior information with. In [5] authors prescribe a more feasible formulation for manet by setting π_{kij} to be only event- and cluster-specific, defining thus a quasi-saturated model with $\pi_{kij} \equiv \pi_{kj}$.

In this manuscript, we propose an extension of manet which introduces parsimony by exploiting covariates information. These covariates could be characteristic related to an actor, such as, gender, age, etc, or features of an event, i.e. type of event, date, duration, and so forth. We define $\mathbf{x}_{i\cdot} = (x_{i1}, \ldots, x_{il}, \ldots, x_{iL})$ to be the L-dimensional vector of covariates for actor i, and $\mathbf{w}_{j\cdot} = (w_{j1}, \ldots, w_{jq}, \ldots, w_{jQ})$ the Q-dimensional vector of covariates for event j. For simplicity, we assume the non-categorical covariates to be standardized, i.e. zero mean and unit variance. Covariates enter the model through a *link* function as in the generalized linear models context [4]. We add the following layers to the Bayesian hierarchical formulation

$$\begin{aligned}
&\boldsymbol{\mu}_k \sim \mathrm{N}(\mu_k; 0, \sigma^2_{\mu}), \quad \boldsymbol{\beta}_k \sim \mathrm{N}_L(\boldsymbol{\beta}_k; \mathbf{0}_L, \sigma^2_{\beta} I_L), \quad \boldsymbol{\gamma}_k \sim \mathrm{N}_Q(\boldsymbol{\gamma}_k; \mathbf{0}_Q, \sigma^2_{\gamma} I_Q),\\
&\eta_{kij} = \mu_k + \sum_{l=1}^{L} \beta_{kl} x_{il} + \sum_{q=1}^{Q} \gamma_{kq} w_{jq},\\
&\pi_{kij}(\mathbf{x}_{i\cdot}, \mathbf{w}_{j\cdot}) = g^{-1}(\eta_{kij}),
\end{aligned}$$

where: η_{kij} is the linear predictor; g^{-1} is the normal distribution's cumulative function $\Phi(\cdot)$, leading to a probit formulation; $\mathrm{N}(\cdot)$ is the normal distribution's density function, with subscripts denoting the dimension of vector or matrix; $(\sigma^2_{\mu}, \sigma^2_{\beta}, \sigma^2_{\gamma})$ are hyper-parameters. Notice that, for g^{-1} set equal to the identity function and only $\{\mu_{kj}\}$ as parameters in the linear predictor, we revert back to the quasi-saturated manet. Once the probabilities π_{kij} are obtained from linear predictors η_{kij}, the corresponding heir parameters $\pi^{\star}_{hij}$ can be computed by means of the combining function $\psi(\cdot)$, exactly as in the formulation without covariates in [5]. Computing linear predictors requires regression coefficients and intercepts to be sampled, and this can be done separately for each cluster as they are independent. However, when an actor i belongs to multiple clusters, it is not univocally defined to which posterior distribution among the K sets of $(\mu_k, \beta_{k1}, \ldots, \gamma_{kQ})$ its likelihood term will contribute to. We follow the prescription of [5] to disentangle this issue. We introduce auxiliary variables $s(\boldsymbol{z}_i, \boldsymbol{\pi}) = s_{kij}$, such that, for a fixed (i, j) we have: $s_{kij} = z_{ik}$ if $\sum_{k=1}^{K} z_{ik} = 1$, whereas, if $\sum_{k=1}^{K} z_{ik} > 1$ then $\boldsymbol{s}_{ij\cdot}$ is a K-dimensional vector of zeros, except for $s_{ik_{\min},j} = 1$. Here $k_{\min}$ denotes the index corresponding to the *parent* cluster having the lowest value of η_{kij}, for a fixed event j and actor i. After introducing the auxiliary variables $\{s_{kij}\}$ into the model, the complete-data likelihood becomes

$$\mathscr{L}_{Y,Z}(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma}; S) = \prod_{k=1}^{K} \prod_{j=1}^{d} \prod_{i=1}^{n} \left\{ \left[\Phi(\eta_{kij}) \right]^{y_{ij}} \left[1 - \Phi(\eta_{kij}) \right]^{1-y_{ij}} \right\}^{s_{kij}}. \tag{3}$$

Equation 3 is similar to the likelihood of a binary regression model with probit link function. Also, Equation 3 highlights that, in general, the model can be cast in a regression framework and thus, potentially, other regression techniques and/or extensions can be further exploited to refine `manet`. An outline of the general idea behind the MCMC implementation is: (i) first, all the observations Y are stacked into a vector $\tilde{\mathbf{y}}$ of length $(n \cdot d) \times 1$; (ii) for each k, a vector $\tilde{\mathbf{y}}_k$ is obtained by filtering $\tilde{\mathbf{y}}$ with rule $s_{kij} = 1$, and then a Bayesian probit regression is performed. For this last part, we refer to the work of [2], where the authors discuss a hierarchical Bayesian formulation of the probit model and provide technical details in the manuscript's Appendix. The MCMC algorithm for `manet` with covariates adjustment is implemented in an `R` script, and code is available upon request.

3 Simulation study

We generate 50 independent datasets from a probit model, with $K = 2$ overlapping groups; sample size and number of events are fixed to $n = 100$ and $d = 15$. Covariates are: (i) actor-specific categorical covariate with two levels, same proportions for both levels (50/100), coded with a single binary variable x_{i1}; (ii) event-specific categorical covariate with three levels, same proportions for the three levels (5/15), coded with two binary variables w_{j1} and w_{j2}. For each replicated dataset, we run: (i) our algorithm, labelled `manet+cov`; (ii) `manet`, with the homonymous `R` package. Number of MCMC iterations is set to 10000 with burn-in window equal to 4000. Results are numerically reported in terms of Adjusted Rand Index and misclassification error rate, averaged across the replicated datasets for both models, showed in percentage. Adjusted Rand Index (ARI) is a measure in the range $[0, 1]$, with higher values indicating better performance. The misclassification error rate (MER) quantifies the proportion of wrongly allocated units, with smaller values indicating better performance. In terms of classification accuracy, `manet` and its covariates-adjusted extension attain comparable ARI and MER, with results slightly in favor of `manet+cov`: more specifically, average ARI is 79.06% for `manet` and 81.26% for `manet+cov`, while average MER is 8.92% for `manet` and 8.10% for `manet+cov`. This corroborates the idea of employing covariate information if available, as the performance of the parsimonious `manet+cov` is on par with the more flexible quasi-saturated `manet`. Results for `manet+cov` are also visualized (see Figure 1) through posterior distributions of the regression coefficients, plotted after aggregating chains from all the independent datasets. Despite the additional uncertainty introduced by combining MCMC samples from different datasets, the posterior distributions show good behavior in terms of location and scale: all the densities in Figure 1 are centered around the true values used to generate data, and exhibit limited dispersion.

4 Conclusions

We have proposed an extension of the model formulated in [5], in order to accomodate for additional information in the form of actor and/or event covariates. By cast-

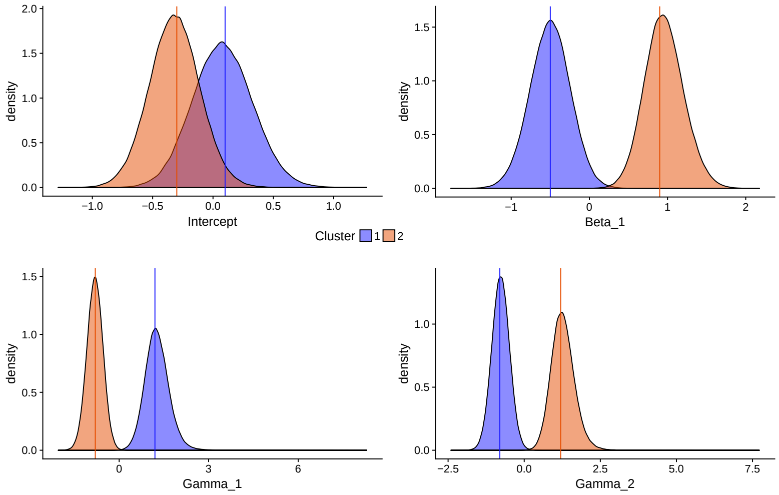


Fig. 1 Posterior distributions for the regression coefficients of `manet+cov`, computed after aggregating all the independent datasets' chains. True values are depicted as vertical lines in the plots, corresponding to parameters for: intercept, $\boldsymbol{\mu} = (0.1, -0.3)$; covariate x_{i1}, $\boldsymbol{\beta}_{\cdot 1} = (-0.5, 0.9)$; covariate w_{j1}, $\boldsymbol{\gamma}_{\cdot 1} = (1.2, -0.8)$; covariate w_{j2}, $\boldsymbol{\gamma}_{\cdot 2} = (-0.8, 1.2)$.

ing part of the inference problem into a binary regression framework, we highlighted the link between regression techniques and covariates-adjusted `manet`, paving the way for more interesting future refinements of the model - such as mixed effects, regularized regression, and so forth. We explored the performance of our proposed algorithm in a simulated environment, which shows appreciable preliminary outputs and encouraging results.

References

1. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer Science & Business Media, New York (2006)
2. Holmes, C.C., Held, L.: Bayesian auxiliary variable models for binary and multinomial regression. Bayesian analysis **1**(1), 145–168 (2006)
3. Kolaczyk, E.D.: Statistical Analysis of Network Data: Methods and Models. Springer, New York (2009)
4. McCulloch, C.E., Neuhaus, J.M.: Generalized linear mixed models. Wiley Online Library (2001)
5. Ranciati, S., Vinciotti, V., Wit, E.C.: Identifying overlapping terrorist cells from the noordin top actor-event network. arXiv preprint arXiv:1710.10319 (2017)
6. Wasserman, S., Faust, K.: Social network analysis: Methods and applications, vol. 8. Cambridge university press (1994)

New Challenges in the Measurement of Economic Insecurity, Inequality and Poverty

Social protection in mitigating economic insecurity

La protezione sociale per l'attenuazione dell'insicurezza economica

Alessandra Coli

Abstract Economic insecurity is generally conceived as a feeling of concern about the material conditions that may prevail in the future in case of adverse events, like job loss or sickness. The welfare state turns a set of individual risks (including sickness, job loss and other risks undermining economic security) into social risks by way of social protection. As a consequence, lower economic insecurity is expected where the welfare state is stronger. This paper explores this connection empirically, using official statistics. The analysis is run on a selection of European countries, whose social protection systems differ in terms of social spending levels, kinds of risks covered and rules for accessing social protection.

Abstract *L'espressione insicurezza economica indica generalmente un sentimento di preoccupazione per le condizioni economiche che potrebbero manifestarsi in futuro, come conseguenza di eventi avversi quali la perdita del lavoro o problemi di salute. Il welfare state trasforma una serie di rischi individuali (tra i quali la disoccupazione, la malattia e altri rischi che possono minacciare la stabilità economica) in rischi sociali, attraverso l'attivazione di misure di protezione sociale. Di conseguenza è ipotizzabile un livello minore di insicurezza economica laddove il sistema di welfare risulti più forte. Questo lavoro si propone di esplorare empiricamente la relazione tra insicurezza economica e sistema di welfare, utilizzano dati della statistica ufficiale. L'analisi è condotta su una selezione di paesi Europei che differiscono per livello di spesa sociale, per tipologia di rischi coperti e per regole di accesso al sistema di protezione sociale.*

Key words: Official statistics, welfare state, subjective well-being

Alessandra Coli
Dipartmento di Economia e Management, Università di Pisa, e-mail: alessandra.coli1@unipi.it

1 Introduction

Economic insecurity is generally conceived as a feeling of concern about the material conditions that may prevail in the future. Indeed, in case of adverse events, individuals may face difficulties in maintaining stable and satisfying living standards. Income instability is propelled by a variety of negative events, originating most commonly in the labour market and in the family. Particularly, thematic literature considers job loss, family dissolution, and poor health among the main triggers of economic insecurity.

The welfare state turns individual risks into social risks by way of social protection, whereby households and individuals are relieved from the financial burden of a number of risks and needs, including those mentioned above as causes of economic insecurity. As a consequence, lower economic insecurity is expected where the welfare state is stronger. This paper explores this connection empirically, using official statistics on social protection benefits and economic insecurity. To this end, an individual binary outcomes which separates "insecure" from "secure" households is modeled as a function of both family-level and country-level characteristics. The analysis is run on a selection of European countries, whose social protection systems differ considerably in terms of social spending levels, kinds of risks covered and rules for accessing social protection.

The paper is organized as follows. Section 2 discusses the connection between economic insecurity and welfare state. Section 3 describes variables and indicators used in the empirical analysis. Section 4 presents main results. Section 5 concludes.

2 Economic insecurity and the welfare state

Literature has proposed several definitions of the term "Economic insecurity". The common premise of all definitions is the idea that some economic misfortune might happen in the future and threaten people's quality of life. However, proposed definitions differ significantly from several points of view. According to some scholars, economic insecurity corresponds to the individual perception of the risk and the anxiety thereof ([11], [13],[2]), others identify economic insecurity with the probability of experiencing adverse events ([15], others with the outcome from the exposure to hardship causing economic losses ([14],[3]). Measures of economic insecurity proposed by literature are strikingly different as well, ranging from subjective to objective measures, based on different units of analysis, namely individual workers, households or countries. Osberg (2015) [11] makes a review of main alternative methodologies proposed.

In this work, the definition proposed by Osberg (1998) [11] is adopted, according to which economic insecurity is a state of anxiety produced by a lack of economic safety. Being anxiety an emotion, self-assessment of economic insecurity seems to be the most appropriate way to measure it. As suggested by Osberg (2015) [12], subjective anxiety should be assessed asking people if they are anxious with respect to different types of hazards. Unfortunately, current available surveys contain only broad few questions on “anxiety”, “worry” or “insecurity”, so that only proxy measures can be derived from existing data.

Economic insecurity has undoubtedly a relational dimension, since the sharing of individual risks within the household and the society allows one to smooth the risks themselves and their consequences. This aspect makes the choice of the unit of analysis an important preliminary issue. In line with one strand of the literature, this study considers economic insecurity as a household-level phenomenon since the individual feeling of uncertainty depends significantly on household composition [15].

However, the stability of economic life depends also on the social and institutional context where individuals live. Institutions regulate risk in different ways: making hazardous events less likely, moving the costs of a hazard from one actor to another or sharing the costs of a hazard across many actors. Welfare state deals with the last two kinds of interventions in particular. The amount of social spending represents a fundamental indicator of the institutional response to social risks, however other characteristics of the welfare state might affect the confidence of citizens in social protection. Literature on welfare regime typologies has proposed to focus on several dimensions([16], [5], [10]), like the composition of risks and needs covered, the quota of people covered and the rules for accessing benefits (whether means-tested or not). The kind of economic transactions through which benefits are delivered (whether benefits in cash or benefits in kind) and the origin of social spending, (whether public from general government taxes or private from employeer or employed social contributions) represent other characterizing aspects.

The following empirical analysis aims to explore the relationship between subjective economic insecurity and the level and kind of social protection delivered by the welfare state. In particular the aim is to assess which characteristics of welfare regimes help mitigate the feeling of economic insecurity of citizens.

3 Empirical analysis: data, variables and indicators

The data set derives from the combination of individual (micro) and country (macro) variables from different data sources. Micro data come from European Statistics on income and living conditions (Eusilc[1], [9]) and provide information on the social and demographic characteristics of the households and their members, on the level and kind of social benefits received and on the level of economic insecurity of the household. Macro data are used to detect the characteristics of European economic systems and their welfare system. Data are taken from Esspros [8] and Socx [1], which the main databases on social protection benefits of international official statistics.

[1] The responsibility for all conclusions drawn from the data lies entirely with the author.

Three main groups of variables are included: the outcome variable which allows us to separate “insecure” from “secure” households, a set of predictors relating to the country welfare system and a set of supplementary variables to control for household and country characteristics that may affect the outcome variable.

The outcome variable is called “INSECURITY” and derives from the combination of two measures of economic safety proposed by official statistics ([6]). The first is drawn from Eusilc module on wellbeing and concerns the respondent’s feeling about the financial situation of his/her family. The second variable is drawn from Eusilc standard module on household material deprivation and concerns the household’s capacity to afford an unexpected required financial expense, paying through its own resources. The amount of the expense is explicitly indicated in the questionnaire and depends on the national at-risk-at-poverty threshold. Based on these two variables, a proxy of household’s insecurity is defined. Particularly, a household is defined as insecure if the head of the household both expresses a score of 5 or lower (on a 0-10 scale) for the financial situation of the family and affirms that her/his household is not able to afford unexpected required expenses.

To characterize different social protection systems across countries, the following set of indicators is proposed: share of social benefits in kind over total benefits, share of means-tested benefits over total benefits, share of private social benefits over total social benefits, social expenditure per inhabitant in power purchased parities (henceforth PPP) and shares of social protection expenditure assigned respectively to health, sicks and disability, to old age and survivors, and to unemployment.

Finally, the following control variables/indicators are selected: number of members with tertiary education, number of members with up to lower secondary education, number of members with a job, number of unemployed members, number of retired members, number of old-age members, number of kids; furthermore, for each household, the amount of benefits received, the level of disposable income and equivalent disposable income (in PPP standard) and the tenure status of accomodation are considered. On the macro level, gross domestic product per inhabitant and household consumption expenditure per inhabitant, both expressed in standard PPP, are used to control for the country economy size.

Variables/indicators have been selected from a larger data set. Particularly, only those significantly associated with the measure of economic insecurity have been considered. Furthermore, some have been excluded to avoid multicollinearity problems when estimating the logistic model. Table 1 shows a synthetic description of variables.

The analysis concerns those European countries for which the complete data set is available, namely: Austria (AT), Belgium (BE), Czech Republic (CZ),Denmark (DK), Estonia (EE), Finland (FI), France (FR),Germany (DE), Greece (EL), Hungaria (HR),Ireland (IR), Italy (IT), Latvia (LV), Luxembourg (LU), Netherlands (NL), Norway (NO),Poland (PL), Portugal (PT), Slovakia (SK), Slovenia (SI), Spain (ES), Sweden (SE), Switzerland (CH), United Kingdom (UK). The choice of the year 2013 is motivated by the presence of an ad hoc module on subjective wellbeing in the 2013 Eusilc wave, which was essential to recover information on self-assessment of own household financial situation. Table 2 shows descriptive statistics

Table 1 Description of variables and indicators.

Name	*Description*	*Categories*
INSECURITY	Low level of satisfaction about the household financial situation AND inability to afford unexpected required expenses	0= not economically insecure, 1= economically insecure
FIN	Financial situation of the household	Score on a 0-10 sale
KIND	Share of in kind benefits over total benefits	Percentage values
MEANS	Share of means-tested benefits over total benefits	percentage values
PRIVATE	Share of private social benefits over total social benefits, 2011	Percentage values
SOCIAL	Social expenditure per inhabitant	Standards PPP
HEALTH	Share of social protection expenditure assigned to health, sicks and disability	Percentage values
OLDAGE	Share of social protection expenditure assigned to old age and survivors	Percentage values
FAMILY	Share of social protection expenditure assigned to family and children	Percentage values
UNEMP.	Share of social protection expenditure assigned to unemployment	Percentage values
NDEGREE	Number of members with tertiary education	Number
NLOW	Number of members with up to lower secondary education	Number
NWORK	Numbers of members with a job	Number
NUNEMP	Numbers of unemployed members	Number
NOLD	Number of members aged 75 or more	Number
NKIDS	Number of members aged less than 16	Number
BENEFITS	Benefits received as a quota of the country average disposable income of households, year 2012	Number
YF	Household disposable income	Standard PPP
HOUSE	The outright owner of the accomodation is a member of the household	0= no,1=yes
GDP	Gross domestic product per inhabitant	Standard PPP
CONS	Household consumption expenditure per inhabitant	Standard PPP

for the quantitative individual variables of the data set, whereas Table 3 points out differences among countries both for country-level variables and for the outcome variable.

4 Empirical analysis: methodology and results

The data set has a multilevel structure: individual families, on which economic insecurity is measured, are nested into countries, on which welfare state characteristics are observed. Given this data structure, multilevel models appear to be the natural

Table 2 Descriptive statistics on the quantitative variables of the dataset. Year 2013

	MEAN	MIN	MAX	SD	25%	50%	75%
FAMSIZE	2.32	1.00	28.00	1.31	1.00	2.00	3.00
NDEGREE	0.48	0.00	6.00	0.72	0.00	0.00	1.00
NLOWED	0.55	0.00	12.00	0.80	0.00	0.00	1.00
NKIDS	0.35	0.00	15.00	0.75	0.00	0.00	0.00
NOLD	0.20	0.00	4.00	0.49	0.00	0.00	0.00
NWORK	0.98	0.00	14.00	0.93	0.00	1.00	2.00
NUNEMP	0.14	0.00	9.00	0.42	0.00	0.00	0.00
NRETIRED	0.46	0.00	5.00	0.69	0.00	0.00	1.00
BENEFITS	0.34	-0.13	57.29	0.45	0.00	0.21	0.53
YF	28664	-226413	1602106	25652	13805	23160	37281
INSECURITY	0.26	0.00	1.00	0.44	0.00	0.00	1.00
FIN	6.00	0.00	10.00	2.46	5.00	6.00	8.00

Table 3 Avarage values of country-level variables and indicators Year 2013

COUNTRY	INSEC.	SOCIAL	KIND	MEANS	PRIVATE	HEALTH	UNEMP.	OLDAGE	FAMILY	CONS	GDP
AT	15.99	10084	30.24	8.40	6.74	32.47	5.46	50.45	9.68	18200	34571
BE	15.23	9114	31.81	5.17	5.88	36.55	11.70	40.94	7.47	16300	31549
CH	9.42	9968	32.43	7.30	26.26	39.34	3.64	47.74	6.00	22900	43349
CZ	27.63	4639	31.95	2.68	2.64	37.28	3.34	47.44	9.09	10900	21993
DE	25.03	9475	37.61	12.08	11.80	42.43	4.12	39.41	11.30	17800	32620
DK	12.63	10358	40.07	35.91	13.92	33.50	5.87	42.61	11.63	15800	33710
EE	32.59	3015	29.63	0.73	0.09	40.12	3.19	44.66	11.05	10100	19796
EL	37.42	4822	20.82	4.69	7.26	26.93	5.20	63.21	4.38	13000	18821
ES	29.22	5856	30.52	14.45	1.45	32.85	12.97	47.47	5.37	13700	23527
FI	10.01	9143	38.35	5.34	3.71	35.80	7.46	41.24	10.72	15800	29780
FR	20.90	9591	36.33	10.95	9.85	34.73	6.25	45.67	7.89	15500	28498
HU	47.52	3889	30.26	4.20	1.01	31.08	2.34	52.43	12.06	9100	17642
IR	38.18	7057	36.32	30.84	8.85	37.57	13.81	32.18	12.52	15500	34831
IT	26.31	7464	24.61	5.67	4.51	29.50	6.18	59.35	4.21	16000	25880
LU	14.44	14391	31.06	3.60	4.86	36.44	6.61	37.54	15.91	21100	68690
LV	53.13	2474	27.23	1.85	0.50	32.06	4.20	53.78	8.16	10100	16352
NL	9.25	10131	35.36	13.44	25.50	42.75	5.62	41.82	3.27	15800	35107
NO	9.32	11410	41.43	4.22	9.15	46.65	2.26	35.09	12.56	18800	48356
PO	35.28	3750	23.86	3.98	0.22	30.52	1.61	59.39	7.43	10800	17613
PT	37.44	5234	26.50	8.37	7.04	31.45	6.86	56.18	4.59	13000	20120
SE	11.26	9524	45.55	2.66	11.55	37.55	4.25	43.75	10.50	15200	32939
SL	35.33	5209	31.83	7.56	4.83	37.04	3.44	48.78	7.98	12000	21493
SK	28.76	3845	33.70	5.07	4.72	39.83	3.42	44.54	9.69	11400	20121
UK	27.54	7674	38.30	13.63	21.37	36.84	2.08	42.93	10.34	18100	28358

choice. Particularly, this analysis applies multilevel logistic regression with random intercept, in order to estimate the probability of experiencing economic insecurity, given some specific characteristics of the welfare state and other country-level and individual-level predictors.

Table 4 shows results of the multilevel logistic model. The model was fit to 200989 households within 24 countries, using standardized input data. A significant positive coefficient indicates that the predictor increases the probability of being classified as insecure, whereas a significant negative sign means that the predictor reduces the risk of economic insecurity. In case of categorical variables, the coefficient sign indicates whether the observed category increases (positive sign) or decreases (negative sign) the risk, compared to the reference category (see table 1 to check categories).

Table 4 Multilevel Logistic Regression

	Dependent variable:
	INSECURITY
FAMSIZE	0.417^{***}
NDEGREE	-0.295^{***}
NLOWED	0.170^{***}
NOLD	-0.081^{***}
NKIDS	-0.082^{***}
NWORK	-0.241^{***}
NUNEMP	0.230^{***}
NRETIRED	-0.272^{***}
BENEFITS	0.063^{***}
GDP	-0.096
CONS	0.204
SOCIAL	-0.555^{**}
HEALTH	-0.101
FAMILY/OLDAGE	0.404^{**}
MEANS	-0.002
PRIVATE	0.044
KIND	-0.308^{*}
YF	-1.619^{***}
HOUSE1	-0.765^{***}
Constant	-1.208^{***}
Observations	200,989
Log Likelihood	−85,016.570
Akaike Inf. Crit.	170,077.100
Bayesian Inf. Crit.	170,301.800
Note:	*p<0.1; **p<0.05; ***p<0.01

Our results show that some features of welfare systems significantly affects economic security, even when controlling for several characteristics of the household and of the country. Holding all other variables constant, the higher the level of per inhabitant social spending and the percentage of in kind services, the lower the probability of economic insecurity. Conversely, the risk of economic insecurity increases as the share of social benefits in favour of family overcome that assigned to the elderly. The controls generally perform in line with intuition. For instance, the presence of low-educated or unemployed members increases economic insecurity whereas the presence of high-educated or members with a job or retired decreases it. Furthermore, the risk of economic insecurity decreases with higher levels of household disposable income and when the outright owner of accommodation is a member of the household.

5 Conclusions

Improving the delivery of economic security is a socially important issue, as economic insecurity represents a major determinant of individual well-being.

This paper argues that social protection (namely the level and the kind of social benefits delivered), may play a relevant role in mitigating the extent of economic insecurity. An empirical analysis is run to investigate the relationship between economic insecurity (defined as the anxiety produced by a lack of economic safety) and main characteristics of the welfare state. In particular, a multilevel logistic regression (with random intercept) is fitted to 200989 households within 24 European countries. Welfare state characteristics are observed at the group (country) level, whereas the outcome is measured on individual families.

Results point out some interesting evidence. Holding all other variables constant, the higher the level of per inhabitant social spending and the percentage of in kind services, the lower the probability of economic insecurity. Conversely, the risk of economic insecurity increases as the share of social benefits in favour of family overcome that assigned to the elderly. Furthermore, household composition affects significantly household economic insecurity. Obviously, results depend on how economic insecurity is measured. Unfortunately, currently available statistics allow one to obtain only a proxy indicator of subjective economic insecurity in Europe.

References

1. Adema, W., Fron, P. , Ladaique, M.: Is the European Welfare State Really More Expensive? Indicators on Social Spending, 1980-2012; and a Manual to the Oecd Social Expenditure Database (Socx), Oecd Social, Employment and Migration Working Papers, No. 124, Oecd Publishing, http://dx.doi.org/10.1787/5kg2d2d4pbf0-en (2011).
2. Bossert, W.; D'Ambrosio, C. (2013) : Measuring economic insecurity, WIDER Working Paper, No. 2013/008, ISBN 978-92-9230-585-7
3. Chou Eileen Y.,Bidhan L. Parmar, and Adam D. Galinsky: Economic Insecurity Increases Physical Pain,Psychological Science, Vol 27, Issue 4, pp. 443 - 454, (2016)
4. D'Ambrosio C.,Rohde N.: The Distribution of Economic Insecurity: Italy and the U.S. over the Great Recession, Review of Income and wealth, Special Issue: Economic Insecurity: Challenges, Issues and Findings, Volume 60, (2014).
5. Esping-Andersen G. The three worlds of welfare capitalism, Cambridge, Polity Press, (1990).

6. Eurostat: Quality of life in Europe, facts and views, economic and physical safety, Statistics Explained, http://epp.eurostat.ec.europa.eu/statisticsexplained/, (2017).
7.
8. Eurostat: The European System Integrated Social Protection Statistics, Manual (2011).
9. Eurostat: EUSILC UDB 2013 -version 3 of January 2016.
10. Ferrera, M., Fargion, V., Jessoula, M.:Alle radici del Welfare all'italiana Origini e futuro di un modello sociale squilibrato. Saggi e ricerche, collana storica della banca dItalia. Marsilio, (2012).
11. Osberg L.: Economic Insecurity, SPRC Discussion Paper No. 88, Social Policy Research Centre, University of New South Wales, Sydney, Australia, (1998).
12. Osberg, L.: How Should One Measure Economic Insecurity?, OECD Statistics Working Papers, 2015/01,OECD Publishing, Paris (2015).
13. Stiglitz, J., Sen, A. Fitoussi: Report by the Commission on the Measurement of Economic Performance and Social Progress, (2010).
14. UNDESA: World Economic and Social Survey 2008: Overcoming Economic Insecurity, United Nations Department of Economic and Social Affairs, (2008).
15. Western, B., D. Bloome, B. Sosnaud and L. Tach: Economic Insecurity and Social Stratification, Annual Review of Sociology, Vol. 38, pp. 341-359,(2012).
16. Titmuss R.: Social policy. An introduction, London, Allen&Unwin,(1974).

Changes in poverty concentration in U.S. urban areas

Cambiamenti nella concentrazione della povertà nelle città americane

Francesco Andreoli and Mauro Mussini

Abstract This paper explores the changes in urban poverty concentration in U.S. cities in the 1980-2016 period. Since poverty is unevenly distributed between neighborhoods in a city, poverty concentration is measured by calculating the Gini index of neighborhood poverty headcount ratios. The change in the index is broken down into components along different dimensions, notably time and space.

Abstract *L'articolo esamina i cambiamenti nella concentrazione della povertà nelle aree urbane americane dal 1980 al 2016. La concentrazione della povertà è misurata con l'indice di Gini, calcolato per l'incidenza della povertà a livello di quartiere. La variazione dell'indice è scomposta secondo la dimensione spaziale e quella temporale.*

Key words: poverty, spatial concentration, decomposition

1 Introduction

Inequalities in American cities can be observed since income and opportunities are unevenly distributed within the cities [4, 2]. In the same American metro area there may be neighborhoods where most of the residents are at the bottom of the income distribution in the city, and other neighborhoods whose residents are mostly at the top. When poor individuals are more likely to live in some neighborhoods, poverty tends to be concentrated in such neighborhoods that offer fewer economic opportunities for their residents, causing a reduction in economic mobility [3]. The most

Francesco Andreoli
Luxembourg Institute of Socio-Economic Research, LISER. MSH, 11 Porte des Sciences, L-4366 Esch-sur-Alzette/Belval Campus, Luxembourg, e-mail: francesco.andreoli@liser.lu

Mauro Mussini
Department of Economics, University of Verona, Via Cantarane 24 - 37129 Verona, e-mail: mauro.mussini@univr.it

used measures of the degree of concentrated poverty across the neighborhoods in a city are based on the fraction of poor individuals who live in neighborhoods with high poverty levels. For example, according to the American Census definition, concentrated poverty can be measured as the fraction of poor individuals in the city who live in neighborhoods where at least 40% of residents are poor. We use a conventional inequality index to measure urban poverty concentration. This takes the form of the Gini index of neighborhood poverty headcount ratios. The more unequally distributed are poverty proportions across the city neighborhoods with respect to the citywide distribution, the larger is urban poverty concentration. This concentration measure captures a form of segregation of poor individuals across the city neighborhoods: when urban poverty concentration is high, there are neighborhoods with very high shares of local residents who are poor, and neighborhoods that are nearly poverty-free.

The Gini index of urban poverty is broken down into spatial components by using the Rey and Smith spatial decomposition of the Gini index [6]. In this way, we assess whether urban poverty is spatially concentrated within the city. We analyze the dynamics of concentrated poverty in American metro areas by considering the change in the Gini index of urban poverty from 1980 to 2016. First, building on Andreoli and Mussini [1], we break down the change over time in the Gini index of urban poverty into components that are attributable to different sources of the change in urban poverty concentration. Second, the change over time in each spatial component of the concentration index is decomposed.

2 Decomposing changes in concentrated poverty

We introduce some preliminary definitions and notation that are used to express the Gini index of urban poverty in a matrix form. This matrix formulation of the index is suitable for the decomposition along different dimensions we use in our analysis. Consider a city with n neighborhoods. Let $\mathbf{p} = (p_1, \dots, p_n)^T$ be the $n \times 1$ vector of neighborhood poverty headcount ratios sorted in decreasing order and $\mathbf{s} = (s_1, \dots, s_n)^T$ be the $n \times 1$ vector of the corresponding population shares. $\mathbf{1}_n$ being the $n \times 1$ vector with each element equal to 1, $\mathbf{P}$ is the $n \times n$ skew-symmetric matrix:

$$\mathbf{P} = \frac{1}{\bar{p}} \left(\mathbf{1}_n \mathbf{p}^T - \mathbf{p} \mathbf{1}_n^T \right) = \begin{bmatrix} \frac{p_1 - p_1}{\bar{p}} & \dots & \frac{p_n - p_1}{\bar{p}} \\ \vdots & \ddots & \vdots \\ \frac{p_1 - p_n}{\bar{p}} & \dots & \frac{p_n - p_n}{\bar{p}} \end{bmatrix}, \tag{1}$$

where $\bar{p}$ is the poverty headcount ratio for the whole city. Let $\mathbf{S} = diag\{\mathbf{s}\}$ be the $n \times n$ diagonal matrix with diagonal elements equal to the population shares in $\mathbf{s}$, and $\mathbf{G}$ be a $n \times n$ G-matrix (a skew-symmetric matrix whose diagonal elements are equal to 0, with upper diagonal elements equal to -1 and lower diagonal elements equal to 1) [7]. The matrix formulation of the Gini index of urban poverty is

$$G(\mathbf{s},\mathbf{p}) = \frac{1}{2} tr\left(\tilde{\mathbf{G}}\mathbf{P}^{T}\right), \tag{2}$$

where the matrix $\tilde{\mathbf{G}} = \mathbf{SGS}$ is the weighting G-matrix, a generalization of the G-matrix [1, 5].

2.1 The components of the change in concentrated poverty

Assume that the distributions of poor and non-poor individuals between the neighborhoods in a city are observed at different times, notably t and $t+1$. Let $\mathbf{p}_t$ be the $n \times 1$ vector of the t neighborhood poverty headcount ratios sorted in decreasing order and $\mathbf{s}_t$ be the $n \times 1$ vector of the corresponding neighborhood population shares. Let $\mathbf{p}_{t+1}$ be the $n \times 1$ vector of the $t+1$ neighborhood poverty headcount ratios sorted in decreasing order and $\mathbf{s}_{t+1}$ be the $n \times 1$ vector of the corresponding population shares. The change in the degree of concentrated poverty between t and $t+1$ can be measured by the difference between the Gini index in $t+1$ and the Gini index in t [1]:

$$\Delta G = G(\mathbf{s}_{t+1},\mathbf{p}_{t+1}) - G(\mathbf{s}_t,\mathbf{p}_t) = \frac{1}{2} tr\left(\tilde{\mathbf{G}}_{t+1}\mathbf{P}_{t+1}^{T}\right) - \frac{1}{2} tr\left(\tilde{\mathbf{G}}_t\mathbf{P}_t^{T}\right). \tag{3}$$

As shown by Andreoli and Mussini [1], equation 3 can be decomposed to separate the components attributable to changes in neighborhood population shares, ranking of neighborhoods by poverty level and disparities between neighborhood poverty headcount ratios. To decompose ΔG some additional definitions and notation are needed. Let $\mathbf{p}_{t+1|t}$ be the $n \times 1$ vector of $t+1$ neighborhood poverty headcount ratios sorted in decreasing order of the respective t neighborhood poverty headcount ratios, and $\mathbf{B}$ be the $n \times n$ permutation matrix rearranging the elements of $\mathbf{p}_{t+1}$ to obtain $\mathbf{p}_{t+1|t}$. Let $\lambda = \bar{p}_{t+1}/\bar{p}_{t+1|t}$ be the ratio of the actual $t+1$ poverty headcount ratio in the whole city to the fictitious $t+1$ poverty headcount ratio which is the weighted average of $t+1$ neighborhood poverty headcount ratios where the weights are the corresponding population shares in t. The elements of $\mathbf{P}_{t+1|t} = \left(1/\bar{p}_{t+1|t}\right)\left(\mathbf{1}_n\mathbf{p}_{t+1|t}^{T} - \mathbf{p}_{t+1|t}\mathbf{1}_n^{T}\right)$ are the relative pairwise differences between the neighborhood poverty headcount ratios in $\mathbf{p}_{t+1|t}$. The decomposition of ΔG is

$$\Delta G = \frac{1}{2} tr\left(\mathbf{W}\mathbf{P}_{t+1}^{T}\right) + \frac{1}{2} tr\left(\mathbf{R}\lambda\mathbf{P}_{t+1}^{T}\right) + \frac{1}{2} tr\left(\tilde{\mathbf{G}}_t\mathbf{D}^{T}\right) = W + R + D, \tag{4}$$

where $\mathbf{W} = \tilde{\mathbf{G}}_{t+1} - \lambda\tilde{\mathbf{G}}_{t|t+1}$, $\mathbf{R} = \tilde{\mathbf{G}}_{t|t+1} - \mathbf{B}^{T}\tilde{\mathbf{G}}_t\mathbf{B}$ and $\mathbf{D} = \mathbf{P}_{t+1|t} - \mathbf{P}_t$. In equation 4, W is the component measuring the change in concentrated poverty due to changes in the distribution of population between neighborhoods. R is the re-ranking component that is greater than zero when at least two neighborhoods exchanged their ranks in the distribution of neighborhood poverty headcount ratios between t and $t+1$. D is the component measuring the change in relative disparities between neighborhood poverty headcount ratios. D is positive (negative) when relative disparities between neighborhood poverty headcount ratios increased (decreased) over time [1].

2.1.1 The role of the change in poverty incidence

Since component D would not reveal changes in neighborhood poverty headcount ratios if they all changed in the same proportion, this component is split into two further terms: one measuring the change in the poverty headcount ratio in the whole city, the second measuring the changes in relative disparities between neighborhood poverty headcount ratios by assuming that the poverty headcount ratio in the city is unchanged between t and $t+1$. Let c be the change in the poverty headcount ratio in the city by assuming that neighborhood population shares are unchanged over time:

$$c = \frac{\bar{p}_{t+1|t} - \bar{p}_t}{\bar{p}_t}. \tag{5}$$

Let $\mathbf{p}^c_{t+1|t} = \mathbf{p}_t + c\mathbf{p}_t$ be the vector of neighborhood poverty headcount ratios we would observe in $t+1$ if the change in every neighborhood poverty headcount ratio was equal to c in relative terms. Vector $\mathbf{p}_{t+1|t}$ can be expressed as

$$\mathbf{p}_{t+1|t} = \mathbf{p}^c_{t+1|t} + \mathbf{p}^\delta_{t+1|t}, \tag{6}$$

where the elements of vector $\mathbf{p}^\delta_{t+1|t}$ are the element-by-element differences between vectors $\mathbf{p}_{t+1|t}$ and $\mathbf{p}^c_{t+1|t}$. Since $\mathbf{p}^c_{t+1|t} = \mathbf{p}_t + c\mathbf{p}_t$, $\mathbf{p}_{t+1|t}$ can be rewritten as

$$\begin{aligned} \mathbf{p}_{t+1|t} &= \underbrace{\mathbf{p}_t + \mathbf{p}^\delta_{t+1|t}}_{\mathbf{p}^e_{t+1|t}} + c\mathbf{p}_t \\ &= \mathbf{p}^e_{t+1|t} + c\mathbf{p}_t, \end{aligned} \tag{7}$$

where the elements of $\mathbf{p}^e_{t+1|t}$ account for disproportionate changes in neighborhood poverty headcount ratios from t to $t+1$, as $\mathbf{p}^e_{t+1|t}$ would equal $\mathbf{p}_t$ if there were no disproportionate changes in neighborhood poverty headcount ratios. Given equations 5 and 7, matrix $\mathbf{P}_{t+1|t}$ can be written as

$$\begin{aligned} \mathbf{P}_{t+1|t} &= \left(1/\bar{p}_{t+1|t}\right)\left(\mathbf{1}_n \mathbf{p}^T_{t+1|t} - \mathbf{p}_{t+1|t}\mathbf{1}^T_n\right) \\ &= \frac{1}{1+c}\mathbf{P}^e_{t+1|t} + \frac{c}{1+c}\mathbf{P}_t. \end{aligned} \tag{8}$$

Since matrix $\mathbf{D}$ in equation 4 is obtained by subtracting $\mathbf{P}_t$ from $\mathbf{P}_{t+1|t}$, $\mathbf{D}$ can be rewritten as

$$\mathbf{D} = \mathbf{P}_{t+1|t} - \mathbf{P}_t \tag{9}$$

$$
\begin{aligned}
&= \frac{1}{1+c}\mathbf{P}^{e}_{t+1|t} + \frac{c}{1+c}\mathbf{P}_t - \mathbf{P}_t \\
&= \underbrace{\left(\frac{1}{1+c}\right)}_{C}\underbrace{\left(\mathbf{P}^{e}_{t+1|t} - \mathbf{P}_t\right)}_{\mathbf{E}} \\
&= C\mathbf{E}.
\end{aligned}
$$

By replacing **D** in equation 4 with its expression in equation 9, the decomposition of the change in concentrated poverty becomes

$$\Delta G = \frac{1}{2}tr\left(\mathbf{W}\mathbf{P}^{T}_{t+1}\right) + \frac{1}{2}tr\left(\mathbf{R}\lambda\mathbf{P}^{T}_{t+1}\right) + C\frac{1}{2}tr\left(\tilde{\mathbf{G}}_t\mathbf{E}^{T}\right) = W + R + CE. \tag{10}$$

C in equation 10 measures the change in the poverty headcount ratio for the whole city, and E captures the change in relative disparities between neighborhood poverty headcount ratios once the effect of the change in the proportion of poor people in the city has been removed. In other words, E is a “pure” component of disproportionate change between neighborhood poverty headcount ratios.

2.2 Spatial decomposition of the change in concentrated poverty

The components of the change in concentrated poverty described in Sect. 2.1 can be broken down into spatial components by using the Rey and Smith approach to the spatial decomposition of the Gini index [6]. Building on Andreoli and Mussini [1], the spatial components of ΔG, W, R and E are obtained. Let $\mathbf{N}_t$ be the $n \times n$ spatial weights matrix having its (i, j)-th entry equal to 1 if and only if the (i, j)-th element of $\mathbf{P}_t$ is the relative difference between the poverty headcount ratios of two neighboring neighborhoods, otherwise the (i, j)-th element of $\mathbf{N}_t$ is 0. Using the Hadamard product,[1] the relative pairwise differences between the poverty headcount ratios of neighboring neighborhoods can be selected from $\mathbf{P}_t$:

$$\mathbf{P}_{N,t} = \mathbf{N}_t \odot \mathbf{P}_t. \tag{11}$$

Since $\mathbf{P}^{e}_{t+1|t}$ and $\mathbf{P}_t$ are defined by the ordering of neighborhoods in t, $\mathbf{N}_t$ also selects the relative pairwise differences between neighboring neighborhoods from $\mathbf{P}^{e}_{t+1|t}$:

$$\mathbf{P}^{e}_{N,t+1|t} = \mathbf{N}_t \odot \mathbf{P}^{e}_{t+1|t}. \tag{12}$$

Given that $\mathbf{E} = \mathbf{P}^{e}_{t+1|t} - \mathbf{P}_t$, the Hadamard product between $\mathbf{N}_t$ and $\mathbf{E}$ is a matrix with nonzero elements equal to the elements of $\mathbf{E}$ pertaining to neighboring neighborhoods:

[1] Let **X** and **Y** be $k \times k$ matrices. The Hadamard product $\mathbf{X} \odot \mathbf{Y}$ is defined as the $k \times k$ matrix with the (i, j)-th element equal to $x_{ij}y_{ij}$.

$$\mathbf{E}_N = \mathbf{P}^e_{N,t+1|t} - \mathbf{P}_{N,t} = \mathbf{N}_t \odot \left(\mathbf{P}^e_{t+1|t} - \mathbf{P}_t\right) = \mathbf{N}_t \odot \mathbf{E}. \tag{13}$$

Let $\mathbf{P}_{N,t+1}$ be the matrix whose nonzero elements are the relative pairwise differences between the poverty headcount ratios of neighboring neighborhoods in $t+1$:

$$\mathbf{P}_{N,t+1} = \mathbf{N}_{t+1} \odot \mathbf{P}_{t+1}. \tag{14}$$

The decomposition of the change in the neighbor component of concentrated poverty is obtained by replacing $\mathbf{P}_{t+1}$ and $\mathbf{E}$ in equation 10 with $\mathbf{P}_{N,t+1}$ and $\mathbf{E}_N$, respectively:

$$\begin{aligned}\Delta G_N &= \frac{1}{2}tr\left(\mathbf{W}\mathbf{P}^T_{N,t+1}\right) + \frac{1}{2}tr\left(\mathbf{R}\lambda\mathbf{P}^T_{N,t+1}\right) + C\frac{1}{2}tr\left(\tilde{\mathbf{G}}_t\mathbf{E}^T_N\right) \\ &= W_N + R_N + CE_N.\end{aligned} \tag{15}$$

$\mathbf{J}_n$ being the matrix with diagonal elements equal to 0 and extra-diagonal elements equal to 1, the matrix with nonzero elements equal to the relative pairwise differences between the $t+1$ poverty headcount ratios of non-neighboring neighborhoods is

$$\mathbf{P}_{nN,t+1} = (\mathbf{J}_n - \mathbf{N}_{t+1}) \odot \mathbf{P}_{t+1}. \tag{16}$$

The matrix selecting the elements of $\mathbf{E}$ related to the pairs of non-neighboring neighborhoods is

$$\mathbf{E}_{nN} = (\mathbf{J}_n - \mathbf{N}_t) \odot \mathbf{E}. \tag{17}$$

The decomposition of the change in the non-neighbor component of concentrated poverty is obtained by replacing $\mathbf{P}_{t+1}$ and $\mathbf{E}$ in equation 10 with $\mathbf{P}_{nN,t+1}$ and $\mathbf{E}_{nN}$, respectively:

$$\begin{aligned}\Delta G_{nN} &= \frac{1}{2}tr\left(\mathbf{W}\mathbf{P}^T_{nN,t+1}\right) + \frac{1}{2}tr\left(\mathbf{R}\lambda\mathbf{P}^T_{nN,t+1}\right) + C\frac{1}{2}tr\left(\tilde{\mathbf{G}}_t\mathbf{E}^T_{nN}\right) \\ &= W_{nN} + R_{nN} + CE_{nN}.\end{aligned} \tag{18}$$

Given equations 15 and 18, the spatial decomposition of the change in concentrated poverty is

$$\Delta G = W_N + W_{nN} + R_N + R_{nN} + C\left(E_N + E_{nN}\right). \tag{19}$$

3 The dynamics of concentrated poverty in American cities

We use information on income and population distributions within U.S. metro areas over the 1980-2016 period from the U.S. Census Bureau database. Information about population counts, income levels and family composition at a very fine spatial grid is taken from the decennial census Summary Tape File 3A. Census tracts are the spatial units of observation, and poverty headcount ratios at the federal poverty line provided by the U.S. Census Bureau are calculated. The 1980-2016 period is divided into five sub-periods to observe the dynamics of each component of the change in concentrated poverty. Since the changes in the population distribution within cities played a minor role in the change in concentrated poverty, we focus our attention on the components measuring the re-ranking effect (R) and the effect of the disproportionate change between census tract poverty headcount ratios (E) for the largest three American metro areas: New York, Los Angeles and Chicago. Concentrated poverty decreased in each of the three metro areas during the period considered, with the largest reduction in Chicago ($\Delta G = -0.11088$) where concentrated poverty ($G = 0.54921$) was greater than in the other two cities in 1980. The degree of concentrated poverty was 0.49669 in New York and 0.41069 in Los Angeles in 1980. Figure 1 shows the spatial decomposition of E. The poverty headcount ratios of non-neighboring census tracts in Chicago have become less unequal, especially during the decade from 2000 to 2010. The decrease in disparities between the poverty headcount ratios of non-neighboring census tracts has been less pronounced in New York and Los Angeles. The decrease in disparities between the poverty headcount ratios of neighboring census tracts in New York has been greater than in the other two cities.

Figure 2 shows the spatial components of the re-ranking effect. The largest re-ranking effect occurred between non-neighboring census tracts in Chicago during the 2000-2010 sub-period. This re-ranking effect partly offset the effect of the reduction in inequality between the poverty headcount ratios of non-neighboring cen-

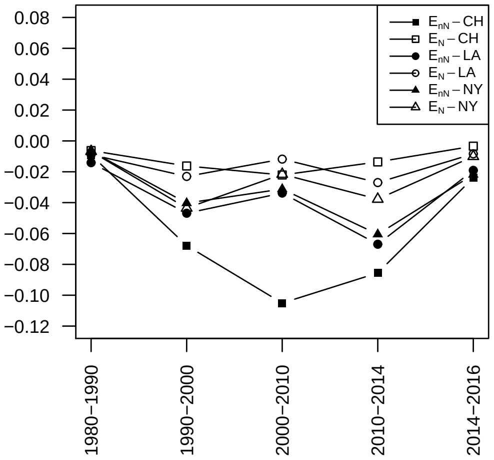


Fig. 1 Spatial components of E in Chicago (CH), Los Angeles (LA) and New York (NY) in the 1980-2016 period.

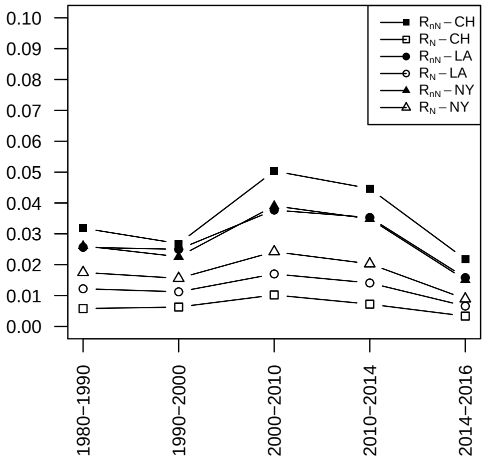


Fig. 2 Spatial components of R in Chicago (CH), Los Angeles (LA) and New York (NY) in the 1980-2016 period.

sus tracts in the city in that decade, especially in view of the increase in poverty incidence in the city ($C = 0.79812$) that weakened the effect of the reduction in inequality between census tract poverty headcount ratios.

References

1. Andreoli, F., Mussini, M.: A spatial decomposition of the change in urban poverty concentration. In: Petrucci, A., Rosanna, V. (eds.) Proceedings of the Conference of the Italian Statistical Society. SIS 2017 Statistics and Data Science: new challenges, new generations, pp. 59-64. Firenze University Press, Florence (2017)
2. Andreoli, F., Peluso, E.: So close yet so unequal: Spatial inequality in American cities. LISER, Working Paper 2017-11 (2017).
3. Chetty, R., Hendren, N., Kline, P. Saez, E.: Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. The Quarterly Journal of Economics, **129**, 1553–1623 (2014)
4. Moretti, E.: Real wage inequality. American Econonomic Journal: Applied Economics, **5**, 65–103 (2013)
5. Mussini, M., Grossi, L.: Decomposing changes in CO_2 emission inequality over time: The roles of re-ranking and changes in per capita CO_2 emission disparities. Energy Economics, **49**, 274–281 (2015)
6. Rey, S., Smith, R.: A spatial decomposition of the Gini coefficient. Letters in Spatial and Resource Sciences, **6**, 55–70 (2013)
7. Silber, J.: Factor components, population subgroups and the computation of the Gini index of inequality. The Review of Economics and Statistics, **71**, 107–115 (1989)

Evaluating sustainability through an input-state-output framework: the case of the Italian provinces

Valutazione delle sostenibilità attraverso un sistema input-state-output: analisi sulle province Italiane

Achille Lemmi[1], Laura Neri[2], Federico M. Pulselli[3]

Abstract In line with the recommendation of monitoring local context, in this paper we propose to investigate regional (NUTS2) and provincial (NUTS3) economic systems in a schematic and usable way: three different indicators are used to take into account resource use (input), societal organization (state) and to quantify the outputs of the system (output). A fuzzy cluster analysis is applied to the input-state-output indicator framework, that, as a whole, represents the interconnection of the three aspects of sustainability, namely environmental, social and economic. This framework is a useful and comprehensive tool for investigating and monitoring local context economic systems.

Abstract *In linea con le direttive di monitorare sistemi come le economie nazionali in un contesto locale, questo lavoro, propone l'analisi dei sistemi economici regionali (NUTS2) e provinciali (NUTS3) in modo semplice schematico e fruibile: tre diversi indicatori sono utilizzati per tenere conto dell'uso delle risorse (input), organizzazione della società (stato) e per quantificare l'output del sistema (output). Un'analisi cluster sfocata viene applicata agli indicatori del framework input-state-output, che, nel suo complesso, rappresentano l'interconnessione dei tre aspetti della sostenibilità, in particolare ambientale, sociale* ed economica.

Key words: Sustainability, input-state-output, fuzzy cluster analysis

[1] Achille Lemmi, ASESD Tuscan Universities Research Centre "Camilo Dagum", lemmiachille@virgilio.it

[2] Laura Neri, Department of Economics & Statistics, University of Siena, laura.neri@unisi.it

[3] Federico M. Pulselli, Ecodynamics Group, Department of Earth, Environmental and Physical Sciences, University of Siena, federico.pulselli@unisi.it

1 Introduction

In recent years there has been an increasing interest in the measurement of collective phenomena at the local level. The EU Committee of the Regions (2014) strongly recommend local authorities to define their own “2020 vision”, based on a territorial dimension, overcoming the present top-down approach of country targets fitting all regions irrespectively. Italy is subdivided into 20 regions (NUTS2) representing the first-level of administrative divisions; the country is further subdivided into 107 provinces (NUTS3). Though progressive measures are trying to eliminate this intermediate administrative level, provinces still play an important role in planning, coordination and cooperation at local level in connection with municipalities and other local bodies. According to OECD Regional Well-being (2016, http://www.oecd.org/cfe/regional-policy/hows-life-country-facts-italy.pdf) “Italy has the largest regional disparities among the OECD countries in safety, with the Aosta Valley ranking in the top 1% and Sicily in the bottom 10% of the OECD regions. Important regional differences are found also in jobs, environment, community, civic engagement, income and access to services”. In line with the recommendation of monitoring local context, in this paper we propose to investigate regional (NUTS2) and provincial (NUTS3) economic systems in a schematic and usable way: three indicators are used to take into account resource use, societal organization and to quantify the outputs of the system. This framework is consistent with an input-state-output scheme (I-S-O, Pulselli et al., 2015), representing the ordered triad environment–society–economy. A three-storey pyramid represents the mutual relationships among the three dimensions of sustainability, rotating the pyramid clockwise, the succession of the stages is oriented from left to right, consistently with the I-S-O framework

2 Data and Methods

In this framework different combinations of indicators can be used to account for the input-state-output. The study here presented is referred to provincial areas, so the preliminary challenge to face is the data availability. Then, the aim of the research is to produce an “objective” classification of the Italian provinces in terms of the three aspects of sustainability. Such classification should be useful for designing and delivering policy responses to economic, environmental and social needs at the local level.

2.1 The indicators

The input indicator should be representative of what a system extracts/obtains, directly or indirectly, from the environment. Referring to provincial areas, which are sub-regional systems, poor datasets are systematically produced, especially in the environmental field. Therefore, no encompassing methods, like emergy evaluation (as in Pulselli et al., 2015) or ecological footprint, or other environmental accounting methods can apply. In this case we used an aggregation of energy consumption measures, collected from two institutional databases (Terna: National electric network; DGERM: Ministry of Economic Development). In particular, we selected electricity consumption and sale of a set of fuel types, for all the provinces of Italy. In order to aggregate these measures, we calculated the equivalent in terms of CO_2 emission to show both the use of resources (electricity and fuels) and the environmental pressure (emissions) on the other. The result is an estimation of gross CO_2 emission due to almost all the items composing the energy sector. In order to monitor the environmental pressure of human activities on each provincial territory and compare provinces, the amount of CO_2 per unit area is computed. This choice helps determine the contribution of human actions (that imply energy consumption) to climate change independently of the number of inhabitants in each area. To encompass the characteristics of the state, a synthetic indicator describing a form of societal organization should be used. Considering the critical importance of reducing unemployment, in order to drive toward inclusive society, an indicator related to the labour market seems to be appropriate. Again, availability and reliability of data at NUTS3 level is the critical issue to face: it has been retained that the most reliable official statistics should be the Labour Force Survey, so the unemployment rate has been chosen as key state indicator.Gross Domestic Product (GDP) is the principal aggregate for measuring economic development/growth of a country/region. In this analysis we maintain this logic, considering the GDP per inhabitants in purchasing parity power as output indicator of the I-S-O system.

2.2 Methods

Our goal is to produce an “objective” classification of the Italian provinces in terms of the three aspects of sustainability, namely environmental, social and economic, according to the chosen triad: CO_2 per unit area, unemployment rate and GDP per inhabitants in purchasing parity power.

Data clustering is recognized as a statistical technique for classifying data elements into different groups (known as clusters) in such a way that the elements within a group possess high similarity while they differ from the elements in a different group. By using the triad of indicators, the clustering algorithm starts from an initial partition of the provinces into a fixed number of groups, where each group is initially randomly chosen. The grouping is then updated: based on the distance between every single observation and the reference objects of each group, every observation is reallocated to the closest group aiming to produce a classification that is reasonably “objective” and “stable”. The classification obtained by using the crisp cluster analysis suffers for both a poor homogeneity within group and a lacking separation between the groups. For this reason, we explored the clustering procedure by using a soft clustering known as Fuzzy Cluster Analysis, a very important clustering technique based on fuzzy logic. In case of soft clustering techniques, fuzzy sets are used to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In many situations, fuzzy clustering is more natural than hard clustering.

Objects on the boundaries between several clusters are not forced to fully belong to one of the cluster, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The most popular algorithm, the fuzzy c-means (developed by Dunn in 1973 and improved by Bezdek in 1981), aims at minimizing an objective function –the weighted within-groups sum of square- whose (main) parameters are the membership degrees and the parameters determining the localisation as well as the shape of the clusters. Objects are assigned to clusters according to membership degrees in [0,1]: 0 is where the data point is at the farthest possible point from a cluster's center and 1 is where the data point is the closest to the center. Each of the c clusters is represented by a cluster center. These centers are chosen randomly in the beginning, then each data vector is assigned to the nearest prototype according to a suitable similarity measure and each center is replaced by the centre of gravity of those data assigned to it. The alternating assignment of data to the nearest center and the update of the cluster centres is repeated until the algorithm converges, i.e., no more changes happen.

Although the extension from crisp to fuzzy clustering seems to be an obvious concept, it turns out that to actually obtain membership degrees between zero and one, it is necessary to introduce a so-called fuzzifier in fuzzy clustering. Usually, the fuzzifier is simply used to control how much clusters are allowed to overlap. This fuzzifier function creates an area of crisp membership values around a prototype while outside of these areas of crisp membership values, fuzzy values are assigned. The analysis has been performed by using the fuzzy clustering with polynomial fuzzifier (Frank Klawonn and Frank Hoppner, 2003).

A problem that frequently occurs in real data analysis is the presence of one or more observation presenting anomalous values, i.e. outliers. Such a subset, that may be referred to as noise, tends to disrupt clustering algorithms making difficult to detect the cluster structure of the remaining domain points. According to the adopted approach the first k standard clusters are homogeneous, whereas the noise cluster, serving as a "garbage collector", contains the outliers and is usually not formed by objects with homogeneous.

All data have been standardized before performing the cluster analysis. The analysis has been conducted by using R and specifically the R package *fclust* (Giordani, Ferraro, 2018). The package provides the cluster solution, cluster validity index and plots and also the visualization of fuzzy clustering results.

The analysis has been performed by using the fuzzy clustering with polynomial fuzzifier with noise cluster. For assessing cluster validity, some have been evaluated: here, just the so called partition coefficient (PC) is presented. Given that the closer to unity the PC index the "crisper" the clustering is and that value close to $1/n_c$ (where n_c is the number of clusters) indicates that there is no clustering tendency, a PC=0.92, indicates a significant cluster structure.

As regard to the visual inspection of fuzzy clustering results -VIFCR- (Klawonn et al., 2003) is a scatter plot where, for each object, the coordinates u_1 and u_2 denote, respectively, the highest and the second highest membership degrees. All points lie within the triangle with vertices (0,0), (0.5,0.5) and (1,0). In the ideal case of (almost) crisp membership degrees all points are near the vertex (1,0). This graph has been evaluated for different partition and the partition with three cluster, plus the noise one, containing just one province (Milan), seems to be the best one (Fig.1).

The algorithm applied to reach the fuzzy clustering solution, assigns an objects to clusters only if the corresponding member function degree is greater than 0.5. In this way the closest hardest partition can be identified by assigning an object to the cluster according to the maximal membership function (>0.5) and the characteristics of each cluster can be identified. The cardinalities and the average membership function of the closest hardest cluster are reported in Table 1, as well as the average of each indicator considered in the clustering procedure. Specifically, in Table 2, the list of provinces belonging to each cluster, according to the maximal membership function, are reported; it is worth to point out that cluster 4, the noise one, includes just Milan.

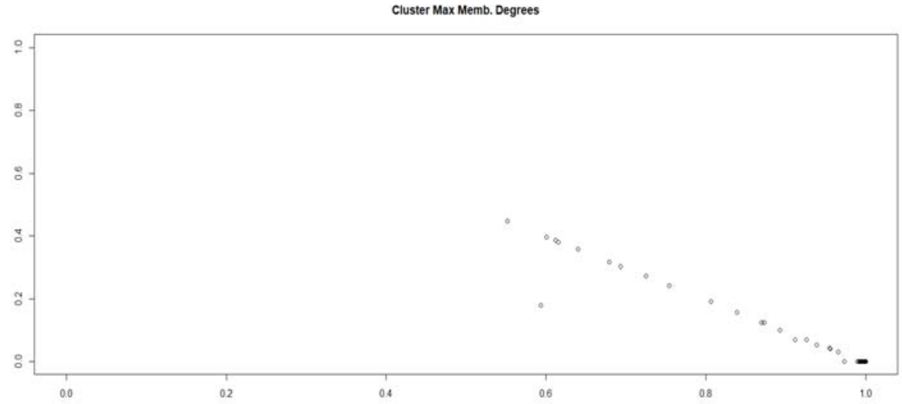


Figure 1: Scatter plot: for each observation, the coordinates u_1 and u_2 denote, respectively, the highest and the second highest membership degrees.

Table 1: Size, average membership function and average values for the indicators by cluster

Cluster	*size*	*m.f.*	*CO₂_area*	*Unempl_rate*	*pps_ab*
1	27	0.96	2272.88	7.75	30866.67
2	44	0.94	912.13	10.31	24854.34
3	35	0.97	804.13	19.78	17046.71
4	1	1	12755.85	7.68	44493.13

In order to have a vision of the distribution of each indicator within each cluster, the boxplots can be observed (Figure 2).

PCA plot (Figure 3) is a very useful tools to visualize the data: this has nothing to do with the type of clustering algorithm or the accuracy of the algorithm used, however it is a useful representation to recognize the utility of the fuzzy clustering, given that the clusters, are clearly separated but the borderline units provinces in this study, are very close.

Table 2: The list of provinces belonging to the four clusters according to the maximal membership function

Cluster	***Provinces***
1	Aosta,Bergamo,Bologna,Bolzano,Brescia,Como,Cremona,Firenze, Forlì-Cesena, Genova,Lecco,Livorno,Mantova, Modena,Padova, Parma,Prato,Ravenna,Reggio-Emilia,Roma,Trento,Treviso,Trieste,Varese,Venezia,erona,Vicenza
2	Alessandria,Ancona,Arezzo,Ascoli,Piceno,Asti,Belluno,Biella,Chieti,Cuneo, Ferrara,Frosinone,Gorizia,Grosseto,Imperia,L'Aquila,La Spezia, Latina,Lodi, Lucca,Macerata,Massa Carrara,Novara,Nuoro,Pavia,Perugia,Pesaro Urbino, Pescara,Piacenza,Pisa,Pistoia,Pordenone,Potenza,Rieti,Rimini,Rovigo,Savona, Siena,Sondrio,Teramo,Terni,TorinoUdine,Verbano-Cusio-Ossola,Vercelli
3	Agrigento,Avellino,Bari,Benevento,Brindisi,Cagliari,Caltanissetta,Campobasso, Carbonia-I.,Caserta,Catania,Catanzaro,Cosenza, Crotone,Enna,Foggia,Isernia, Lecce,Matera,Medio,Campidano,Messina,Napoli,Ogliastra,Olbia-T.,Oristano, Palermo,Ragusa,Reggio C., Salerno,Sassari,Siracusa,Taranto,Trapani,Vibo Valentia,Viterbo
4	Milano

3. An attempt of taxonomy according to the triad

Cluster composition (Table 2) and average values of indicators (Tab. 1) suggest a hig h heterogeneity among clusters emphasizing the existence of economic disparities.
In Cluster 1 there are 27 provinces: Rome, plus 26 provinces located in the North or Central Italy. Comparing Cluster 1 with Cluster 2 and 3, we can observe the highest a verage level of CO_2 per area, the highest GDP per capita and the lowest unemployme nt rate. Observing Fig. 2, we can also realize that the unemployment rate is quite ho mogeneous within the cluster while the GDP per capita and the level of CO_2 per area present the largest variability if compared with the variability of Cluster 2 and 3. Thi s result is justified considering the presence of very peculiar provinces in this cluster, like Bolzano (belonging to an autonomous region) with the highest GDP per capita (40000 Euro) (out of Milan), and the lowest unemployment rate among all the Italian provinces, and the lowest level of CO_2 per area within Cluster 1. and Livorno, the pr ovince presenting the lower membership function for this cluster (0.64), maybe due t o its GDP which is closer to the average GDP per capita computed for Cluster 2.
Cluster 2 is the most heterogeneous cluster as regard to the geography of provinces. I t is composed mainly by provinces located in the North and Central Italy plus provin ces of Lazio (out of Rome and Viterbo) and Abruzzo plus Nuoro and Potenza. All th e provinces in Cluster2, out of Torino, have small medium dimension. This cluster is characterized by high variability in GDP, ranging from 17900 for Nuoro to 29900 fo r Siena and low variability in CO_2. Torino presents the lowest membership function (0.55) among all the provinces, being a borderline unit between Cluster 2 and 1. Clus ter 3 is composed by whole Calabria, Campania, Molise, Puglia, Sicilia and Sardegn a (out of Nuoro), plus Viterbo (Lazio). This cluster is very similar to the one, labelle d as “Minimal system with high social risks” by Bertin (2012) in his classification co ncerning the welfare state systems of the Italian regions. Cluster 3 is characterized by the lowest average level of CO_2 emission per area, although such level is biased

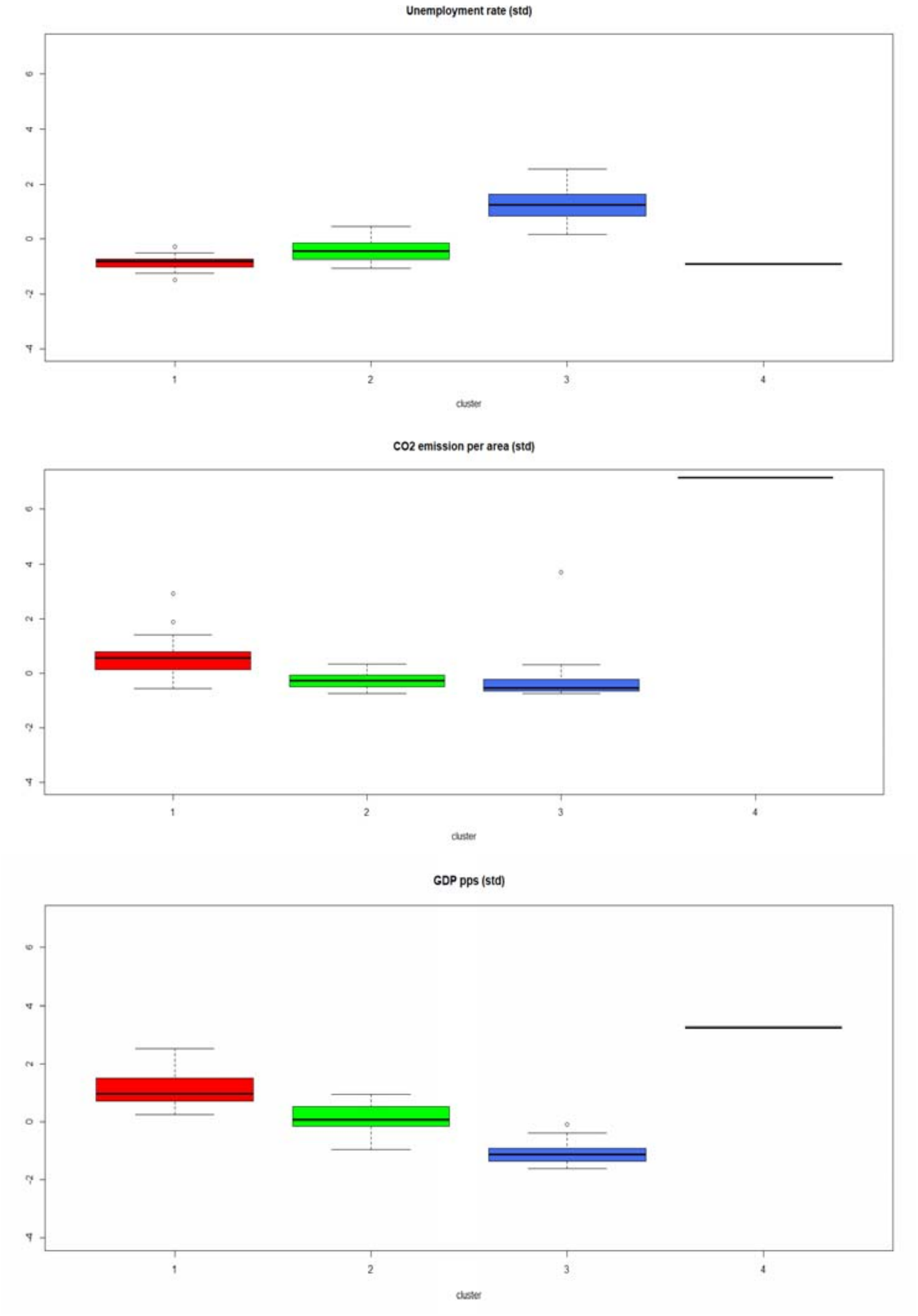


Figure 2: Boxplot-cluster structure

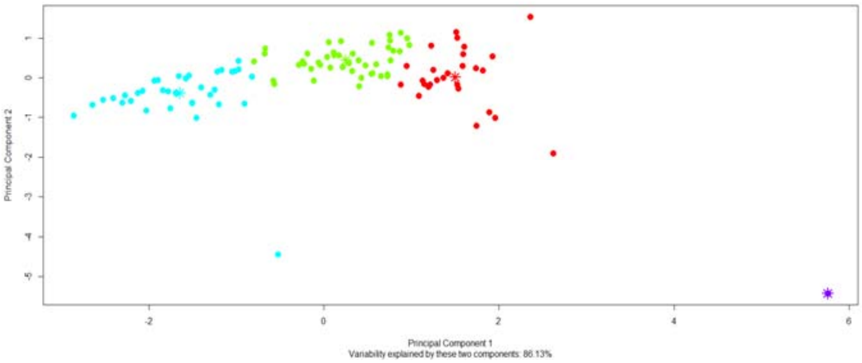


Figure 3: Principal component plot of the cluster structure

due to the membership of Napoli, presenting the highest level of CO_2 among all Italian provinces (out of Milan). As regard to the unemployment rate and GDP, the cluster presents an evident disparity with respect to the others. As regard to GDP per capita, the level of Cagliari (23400), the highest value within cluster 3, is lower than the minimum value in Cluster 1 (25500) registered for Livorno. It is interesting to observe that Napoli is a borderline unit between cluster 3 and 2.

Finally, Cluster 4, the one containing just Milan, is for sure a true outlier with respect to the level of CO2 emission per area (more than ten times the average values of all the other provinces i.e.12755.85 vs 1218.29) and GDP per capita which in Milan is nearly double than the average values of all the other provinces (44493 vs 23812).

The analysis conducted confirm the well-known dualism, resulted in a North-South divide in GDP per capita and in labour-market performance, adding a new element: the Southern Italian provinces are homogeneous with respect to the considered characteristics whilst the Northern and Central provinces are not homogeneous even if they belong to the same region. This result suggests that local policies can be better aligned and tailored to specific local opportunities and challenges. Moreover, the clustering structure obtained can be an useful tool to adopt, considering that provinces belonging to the same cluster could share ways to generate better outcomes for environment, jobs and the economy, trying to reach provinces of their cluster.

References

1. Bertin, G.: Crises and change processes of the welfare systems, Italian Sociological Review, vol. 2, n. 1, pp. 1–13 (2012)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Ac. Pub., Norwell, MA, (1981)
3. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3, 3, 32-57 (1974)
4. Klawonn, F., Höppner, F.: What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier. In: Advances in intelligent data analysis, pp. 254-264. LNCS (2003)
5. Giordani, P., Ferraro, M.B.: fclust: Fuzzy Clustering, R package version 1.1.2, http://cran.r-project.org/package=fclust.
6. Pulselli, F.M., Coscieme, L., Neri, L., Regoli, A., Sutton, P., Lemmi, A., Bastianoni, S.: The World Economy in a cube: a more rational structural representation of Sustainability. Glob. Environ. Change 35, pp. 41–51 (2015).

New Methods and Models for Ordinal Data

Weighted and unweighted distances based decision tree for ranking data.

Alberi decisionali per ranking data basati su distanze pesate e non pesate

Antonella Plaia, Simona Buscemi, Mariangela Sciandra

Abstract Preference data represent a particular type of ranking data (widely used in sports, web search, social sciences), where a group of people gives their preferences over a set of alternatives. Within this framework, distance-based decision trees represent a non-parametric tool for identifying the profiles of subjects giving a similar ranking. This paper aims at detecting, in the framework of (complete and incomplete) ranking data, the impact of the differently structured weighted distances for building decision trees. The traditional metrics between rankings don't take into account the importance of swapping elements similar among them (element weights) or elements belonging to the top (or to the bottom) of an ordering (position weights). By means of simulations, using weighted distances to build decision trees, we will compute the impact of different weighting structures both on splitting and on consensus ranking. The distances that will be used satisfy Kemenys axioms and, accordingly, a modified version of the rank correlation coefficient τ_x, proposed by Edmond and Mason, will be proposed and used for assessing the trees' goodness.

Abstract *I dati di preferenza rappresentano un particolare tipo di ranking data (ampiamente usati nello sport, nella ricerca sul web, nelle scienze sociali), dove un gruppo di persone dá le sue preferenze su un set di alternative. In questo contesto, gli alberi decisionali basati sulle distanze rappresentano uno strumento non parametrico per identificare i profili di soggetti che forniscono un ranking simile. Questo "articolo" mira ad indagare, nel contesto di ordinamenti completi e incompleti, quale sia l'impatto delle differenti distanze pesate sulla costruzione di alberi decisionali. Le tradizionali metriche tra ordinamenti non prendono in considerazione l'importanza di scambiare elementi simili tra di loro (pesi di item) o elementi che stanno in cima o in coda a una classifica (pesi di posizione). Usando le distanze pesate per la costruzione degli alberi, condurremo uno studio di sim-*

Antonella Plaia, Simona Buscemi, Mariangela Sciandra
Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Palermo, Viale delle Scienze, Edificio 19, 90128 Palermo, Italy, e-mail: `antonella.plaia(simona.buscemi, mariangelasciandra)@unipa.it`

ulazione per misurare l'impatto di differenti sistemi di peso sia sugli splitting sia sull'individuazione del consensus ranking. Le distanze che saranno usate rispettano gli assiomi di Kemeny.

Key words: weighted distances, ranking, Kemeny, consensus, trees

1 Introduction

Distances between rankings and the rank aggregation problem have received a growing consideration in the past few years. Ranking and classifying are two simplified cognitive processes usefull for people to handle many aspects in their life. When some subjects are asked to indicate their preferences over a set of alternatives (items), ranking data are called preference data. One great issue of interest in literature is: what can be done to identify, through subject-specific characteristics, the profiles of subjects having similar preferences? In order to answer to this question, different solutions have been proposed: distance-based tree models [15], distance-based multivariate trees for ranking [4], log-linearized Bradley-Terry models [9] and a semi-parametric approach for recursive partitioning of Bradley-Terry models for incorporating subject covariates [18]. Lee and Yu (2010) [15] investigated the development of distance-based models using decision tree with weighted distances, where weights are related to items. The traditional metrics between rankings don't take into account the importance of swapping elements similar among them (element weights) or elements belonging to the top (or to the bottom) of an ordering (position weights). Kumar and Vassilvitskii (2010) [14] provided an extended measure for Spearman's Footrule and Kendall's τ, embedding weights relevant to the elements or to their position in the ordering. The purpose of this paper is to investigate the effect of different weighting vectors on the tree. A particular attention is given to the weighted Kemeny distance and to the consensus ranking process for assigning a suitable label to the leaves of the tree. The stopping criterion for detecting the optimum tree is a properly modified Tau_x [10].

The rest of the paper is organized as follows: Section 2 introduces different metrics between rankings, their properties and their weighted extension; Section 3 introduces the weighted correlation coefficient; after a brief view on decision trees, in Section 4 we perform our analysis through a simulation study and, in the end, a short conclusion is presented (Section 5).

2 Distances between rankings

Ranking data arise when a group of n individuals (experts, voters, raters etc) shows their preferences on a finite set of items (k different alternatives of objects, like movies, activities and so on). If the k items are ranked in k distinguishable ranks, a complete ranking or linear ordering is achieved [8]. A ranking π is, in this case, one of the $k!$ possible permutations of k elements, containing the preferences given by the judge to the k items. When some items receive the same preference, then a tied ranking or a weak ordering is obtained. In real situations, many times it happens that not all items are ranked: partial rankings, when judges are asked to rank only a subset of the whole set of items, and incomplete rankings, when judges can freely choose to rank only some items. In order to get homogeneous groups of subjects having similar preferences, it's natural to measure the spread between rankings through dissimilarity or distance measures among them. Within the metrics proposed in literature to compute distances between rankings, the Kemeny distance will be here considered [13]. The Kemeny distance (K) between two rankings π and σ is a city-block distance defined as:

$$K(\pi,\sigma) = \frac{1}{2}\sum_{r=1}^{k}\sum_{s=1}^{k}|a_{rs} - b_{rs}| \tag{1}$$

where a_{rs} and b_{rs} are the generic elements of the $k \times k$ score matrices associated to π and σ respectively, assuming value equal to 1 if the item r is preferred to or tied with the item s, -1 if the item s is preferred to the item r and 0 if $r = s$.
K is in a one-to-one correspondence, $\tau = 1 - 2d/D_{max}$, to the rank correlation coefficient τ_x proposed by [10] defined as:

$$\tau_x(\pi,\sigma) = \frac{\sum_{r=1}^{k}\sum_{s=1}^{k} a_{rs}b_{rs}}{k(k-1)}. \tag{2}$$

2.1 Weighted distances

Kumar and Vassilvitskii (2010) [14] introduced two aspects essential for many applications involving distances between rankings: positional weights and element weights. In short, i) the importance given to swapping elements near the head of a ranking could be higher than the same attributed to elements belonging to the tail of the list or ii) swapping elements similar between themselves should be less penalized than swapping elements which aren't similar. In this paper, we deal with case i) and consider the weighted version of the Kemeny metric. For measuring the weighted distances, the non-increasing weights vector $w = (w_1, w_2, ..., w_{k-1})$ constrained to $\sum_{p=1}^{k-1} w_p = 1$ is used, where w_p is the weight given to position p in he ranking.

Given two generic rankings of k elements, π and σ, the Weighted Kemeny distance was provided by [11] as follows:

$$K^w(\sigma,\pi) = \frac{1}{2}\left[\sum_{\substack{r,s=1\\ r<s}}^{k} w_r|a_{rs}^{(\sigma)} - b_{rs}^{(\sigma)}| + \sum_{\substack{r,s=1\\ r<s}}^{k} w_r|b_{rs}^{(\pi)} - a_{rs}^{(\pi)}|\right], \tag{3}$$

where (σ) states to follow the σ ranking and (π), similarly, orders according to π (see [16] for more details).

3 A new suitable rank correlation coefficient

In this paper we propose a new consensus measure, suitable for position weighted rankings. It represents an extension of τ_x proposed by [10] that handles linear and weak rankings when the position occupied by the items is relevant. It is defined as:

$$\tau_x^w(\pi,\sigma) = \frac{\sum_{r<s}^{k} a_{rs}^{\sigma} b_{rs}^{\pi} w_r + \sum_{r<s}^{k} a_{rs}^{\pi} b_{rs}^{\sigma} w_r}{Max[K^w(\sigma,\pi)]}, \tag{4}$$

where the denominator represents the maximum value for the Kemeny weighted distances, equal to:

$$Max[K^w(\pi,\sigma)] = \sum_{r=1}^{m-1} (m-1) w_r \cdot n. \tag{5}$$

A consensus measure has to satisfy conditions like unanimity, anonymity and neutrality, i.e. the consensus in every subset of individuals is maximum if and only if all opinions are the same and the degree of consensus is not affected by permutations of the voters or permutations of the alternatives, respectively. Furthermore, it could fulfill some other properties such as maximum dissension, reciprocity and homogeneity, i.e.: in each subset of two subjects, the minimum consensus is achieved if the preferences are linear orderings and each one is the reverse of the other one; if all individual orderings are reversed then level of consensus doesn't change and, in the end, if a subset of agents is replicated, then the consensus in that group doesn't change [11]. By simulations, we verified the fulfillment of these properties.

4 Decision Trees and Simulation Study

Decision trees are non parametric recursive statistical tools used for classification and prediction issues. The most known decision tree methodology is applied when the response variable is categorical or quantitative. Recently the procedure has been extended to rankings as response variable. For more details see [16]. In this paper we will use the weighted Kemeny distance (3) as impurity function and τ_x^w (4) as a measure of goodness of the tree. In particular, we are interested in evaluating the effect both on the splits and on the leaf labels of different weighting vectors w. For this reason, following [7], we consider a theoretical population partition of the predictor space (X_1 and X_2): Fig. 1 shows one of the nine datasets considered in the simulation plan, with $X_1 \sim U(0,10)$ and $X_2 \sim U(0,6)$. The number of rankings falling in each group was defined by a random number drawn from a normal distribution $N(10,2)$ and each number was divided by the summation of all of them, obtaining a relative frequency distribution for each sub-partition.

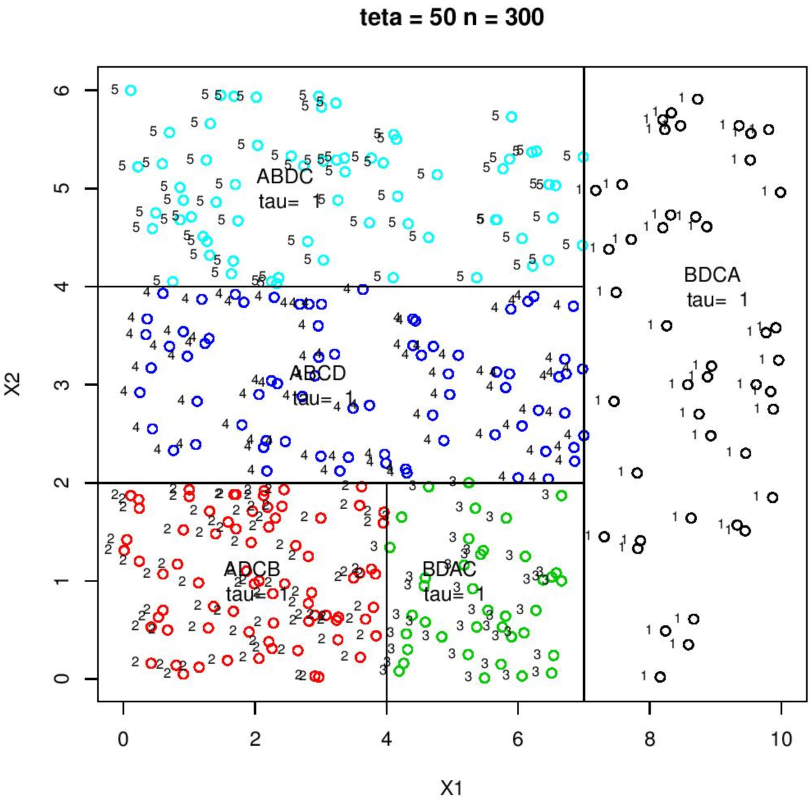
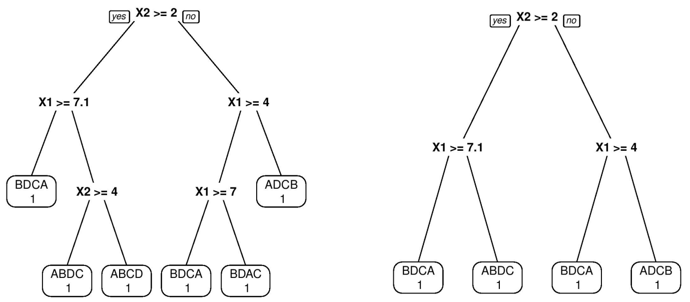


Fig. 1: Generation of homogeneous groups of ranking

The rankings of $k = 4$ items of each sub-partition were generated from a Mallows Model [12], varying the dispersion parameter θ, according to three different level of noise (low with $\theta = 50$, medium with $\theta = 2$ and high with $\theta = 1$). Considering three levels for the sample size (50, 100 and 300), the experimental design counts 3x3=9 different experiments. For each dataset, five different weighting vectors are considered : $w_1 = (1/3, 1/3, 1/3)$, $w_2 = (3/6, 2/6, 1/6)$, $w_3 = (1/2, 1/2, 0)$, $w_4 = (2/3, 1/3, 0)$ and $w_5 = (1, 0, 0)$.

With reference to the data in Fig. 1 (corresponding to $\theta = 50$ and $n = 300$), Fig. 2 reports two of the five trees obtained: in particular, Fig. 2a shows the tree corresponding to w_1, which perfectly recreates the original partition of the predictor space; Fig. 2b corresponds to w_3 and, as expected, does not perform the two splits $X \gtreqless 4$ and $X \gtreqless 7$ (the couples of rankings below each of the split in fig. 2a do not differ for the first two positions).



(a) Decision tree with w_1 weights (b) Decision tree with w_3 weights

Fig. 2: Decision tree models for weighted rankings

5 Conclusion

In this paper, we have focused on distance-based decision trees for ranking data, when the position occupied by items is relevant. We have proposed the weighted Kemeny distance as impurity function and a relative proper weighted consensus measure to be computed in each leaf and for the whole tree. Our methodology found to be capable of identifying correctly homogeneous groups of rankings for the relevant positions (according to the weighting structure). Future developments could be an analytical study of the properties of the new consensus measure and a replication of the same analyses both with an increasing number of items and in the case of weak orderings.

References

1. Amodio, S. and D'Ambrosio, A. and Siciliano, R.: Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. European Journal of Operational Research, 249(2), 667-676 (2016)

2. Breiman, L. and Friedman, J. and Olshen, R. and Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, 1984.
3. Cheng, W and Hühn, J. and Hüllermeier, E.: Decision Tree and Instance-Based Learning for Label Ranking. In:Léon Bottou and Michael Littman, Proceedings of the 26th International Conference on Machine Learning, pages 161-168, Montreal. Omnipress. (2009)
4. D'Ambrosio, A.: Tree based methods for data editing and preference rankings. Ph.D. thesis, Universitá degli Studi di Napoli ”Federico II” (2007)
5. D'Ambrosio, A. and Amodio, S.: ConsRank: Compute the Median Ranking(s) According to the Kemeny's Axiomatic Approach. R package version 1.0.2. (2015)
6. D'Ambrosio, A. and Amodio, S. and Iorio, C.: Two algorithms for finding optimal solutions of the Kemeny rank aggregation problem for full rankings. Electronic Journal of Applied Statistical Analysis, 8(2). (2015)
7. DAmbrosio, A., and Heiser, W. J.: A recursive partitioning method for the prediction of preference rankings based upon kemeny distances. Psychometrika 81.3 774-794. (2016)
8. Cook, W.D.: Distance based and ad hoc consensus models in ordinal preference ranking. European Journal Operation Research 172:369385. (2006)
9. Dittrich, R., Hatzinger, R., Katzenbeisser, W.: Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. Journal of the Royal Statistical Society C (Appl Stat) 47(4):511525 (1998)
10. Edmond, E. J. and Mason, D. W.: A new rank correlation coefficient with application to the concensus ranking problem. Journal of Multi-criteria decision analysis, 11, 17-28. (2002)
11. García-Lapresta, J. L. and Pérez-Román, D.: Consensus measures generated by weighted Kemeny distances on weak orders. In: Procceedings of the 10th International Conference on Intelligent Systems Design and Applications, Cairo. (2010)
12. Irurozki, E., Calvo, B., Lozano, J. A.: PerMallows: An R Package for Mallows and Generalized Mallows Models. Journal of Statistical Software, 71(12), 1-30. doi:10.18637/jss.v071.i12 (2016)
13. Kemeny, J. G. and Snell, J. L.: Preference rankings an axiomatic approach. MIT Press. (1962)
14. Kumar, R. and Vassilvitskii, S.:Generalized Distances Between Rankings. In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pages 571-580, New York, NY, USA. ACM (2010)
15. Lee, P. H. and Yu, P. LH.: Distance-based tree models for ranking data. Computational Statistics & Data Analysis, 54(6), 1672-1682. (2010)
16. Plaia, A. and Sciandra, M.: Weighted distance-based trees for ranking data. Advances in Data Analysis and Classification, pages 1-18. Springer, https://doi.org/10.1007/s11634-017-0306-x (2017)
17. Sciandra, M. Plaia, A. and Capursi, V.: Classification trees for multivariate ordinal response: an application to Student Evaluation Teaching. Quality & Quantity, pages 1-15. (2016)
18. Strobl, C., Wickelmaier, F., Zeileis, A.: Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. Journal of Educational and Behavioral Statistics 36(2):135153. (2011)
19. Therneau, T. and Clinic, M.: User written splitting functions for RPART (2015)
20. Therneau, T. and Atkinson, B. and Ripley, B: rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10 (2015)
21. Yu, P. LH. and Wan, W. M. and Lee, P. H.: Decision tree modeling for ranking data. In Preference Learning, pages 83-106. Springer. (2010)

A dissimilarity-based splitting criterion for CUBREMOT

Carmela Cappelli, Rosaria Simone and Francesca Di Iorio

Abstract CUBREMOT (CUB REgression MOdel Trees) is a model-based approach to grow trees for ordinal responses that relies on a class of mixture models for evaluations and preferences (CUB). The original proposal considers deviances in log-likelihood to partition observations. In the present paper a new splitting criterion is introduced that, among the significant splitting variables, chooses the one that maximizes a dissimilarity measure. This choice is tailored to generating child nodes as far apart as possible with respect to the estimated probability distributions. An application to real data on Italians' trust towards the European Parliament taken from the official survey on daily life conducted by the Italian National Institute of Statistics (ISTAT) in 2015 is presented and discussed in comparison with alternative methods.
Abstract *Nel presente lavoro viene proposto un nuovo criterio di split per la procedura* CUBREMOT *(CUB REgression MOdel Trees).* CUBREMOT *è uno strumento per crescere alberi per risposte ordinali ai cui nodi sono associati modelli mistura per le valutazioni e preferenze (modelli* CUB*) e che utilizza un criterio di split basato sulla differenza in log-verosimiglianza. Il criterio di split alternativo che viene qui introdotto utilizza invece un indice di dissimilarità per generare, attraverso lo split di un nodo padre, nodi figli che siano il più distanti possibile in termini della distribuzione di probabilità stimata. La validità dell'approccio e il confronto con altri metodi sono mostrati mediante l'applicazione a dati reali sulla fiducia verso il Parlamento Europeo sulla base dell'indagine multiscopo condotta nel 2015 dall'ISTAT.*

Key words: Tree -based methods; Ordinal Responses; Dissimilarity measure

Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò, 22, 80138 Napoli, e-mail: carcappe@unina.it, e-mail: rosaria.simone@unina.it, e-mail: fdiiorio@unina.it

1 Introduction

In the spirit of the model-based partitioning approach [9], CUBREMOT [2, 3] is a tool for growing trees for ordinal responses in which every node is associated with a CUB model [4]. This approach to model preferences, judgements and perceptions is based on the idea that discrete choices arise from a psychological process that involves a personal *feeling* and an inherent *uncertainty* both possibly related to explanatory covariates.
The splitting criterion employed in CUBREMOT computes the log-likelihood increment from the father node to the child nodes for each possible split, and at the given step chooses the one that maximizes such deviance. Thus, this criterion selects the covariate that entails the most plausible values for CUB parameters in the child nodes among the variables that are significant for at least one of the model components at the father node.
We propose a further splitting criterion that focuses on the dissimilarity between child nodes, aiming at generating child nodes as far apart as possible with respect to the probability distributions estimated by CUB models. Both splitting criteria generate a model-based tree whose terminal nodes provide different profiles of respondents, which are classified into nodes according to levels of feeling and uncertainty conditional to the splitting covariates. In what follows, we briefly recall the main features of CUBREMOT, then we present the new splitting criterion and we illustrate the results of an application to data from the official survey on daily life conducted by the Italian National Institute of Statistics (ISTAT) in 2015 focusing on Italians' trust towards the European Parliament.

2 Background and Methodology

CUB models paradigm [4] designs the data generating process yielding to a discrete choice on a rating scale as the combination of a *feeling* and an *uncertainty* component. The resulting mixture prescribes a shifted Binomial distribution for feeling to account for substantial likes and agreement and assigns a discrete Uniform distribution for uncertainty to shape heterogeneity. Then, if R_i denotes the response of the i-th subject to a given item of a questionnaire,

$$Pr(R_i = r|\pi_i, \xi_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r}(1-\xi_i)^{r-1} + (1-\pi_i)\frac{1}{m}, \quad r = 1, \dots, m,$$

where the model parameters π_i and ξ_i are called uncertainty and feeling parameter, respectively. Covariates may be included in the model in order to relate feeling and/or uncertainty to respondents' profiles. Customarily, a logit link is considered:

$$logit\,(\pi_i) = \boldsymbol{x}_i\boldsymbol{\beta}; \qquad logit\,(\xi_i) = \boldsymbol{w}_i\boldsymbol{\gamma}\,, \tag{1}$$

where $\boldsymbol{x}_i, \boldsymbol{w}_i$ are the values of selected explanatory variables for the i-th subject. If no covariate is considered neither for feeling nor for uncertainty, then $\pi_i = \pi$ and $\xi_i = \xi$ are constant among subjects. Estimation of CUB models relies on likelihood methods and on the implementation of the Expectation-Maximization (EM) algorithm.
In CUBREMOT, CUB models are employed in the top-down partitioning algorithm that grows the tree as follows. According to binary recursive partitioning, each of the available covariates is sequentially transformed into suitably splitting variables or binary questions which are Boolean condition on the value (or categories) of the covariate where the condition is either satisfied ("yes") or not satisfied ("no") by the observed value of that covariate (for details see [1]). In this respect any split s can be seen as a dummy variable.
Then, for a given node $k \geq 1$ with size n_k, a CUB without covariates is fitted, whose log-likelihood at the final ML estimates $(\hat{\pi}_k, \hat{\xi}_k)$ is denoted by $\mathscr{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k)$. Then, a CUB with splitting variable s is tested: if it is significant for at least one component, it implies a split into a left and right child nodes that will be associated with the conditional distributions $R|s = 0$ with parameter values $(\hat{\pi}_{2k}, \hat{\xi}_{2k})$ and $R|s = 1$ with parameter values $(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})$, respectively, Thus, the splitting criterion proposed in [2, 3] at the given step chooses the split that maximizes the deviance:

$$\Delta \mathscr{L}_k = \left[\mathscr{L}_{n_{2k}}(\hat{\pi}_{2k}, \hat{\xi}_{2k}) + \mathscr{L}_{n_{2k+1}}(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})\right] - \mathscr{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k). \tag{2}$$

Indeed, such difference measures the improvement in log-likelihood yielded by the inclusion of the significant splitting variable and the best split, being associated with the maximum log-likelihood increment, provides the child nodes characterized by the most plausible values for CUB parameters.

Here we propose an alternative splitting criterion based on the concept of dissimilarity between child nodes: the proposal considers a proper version of the normalized index proposed by [8] that compares the estimated probability distribution with the observed relative frequencies and it is generally considered in the framework of CUB models as a goodness of fit measure. Specifically, aiming at the generation of child nodes that are the farthest apart from each other in terms of distribution of the responses, in the set $\mathscr{S}_k = \{s_{k,1}, \dots, s_{k,l}\}$ of the l significant splitting variables for node k, a CUBREMOT is grown by choosing, at each step, the split maximizing the distance between the estimated CUB probability distributions $\hat{p}_{2k}$ and $\hat{p}_{2k+1}$ for the child nodes in terms of the dissimilarity measure:

$$Diss(2k, 2k+1) = \frac{1}{2}\sum_{r=1}^{m} |\hat{p}_{2k} - \hat{p}_{2k+1}|. \tag{3}$$

The choice of this normalized index entails that, as long as CUB models estimated at the child nodes provide an adequate fitting, the splitting variable generates an optimal partition of the father node in terms of the chosen distance. In particular, the resulting terminal nodes determine well-separated profiles of respondents, in terms of both feeling (agreement, preferences, and so on) and uncertainty (indecision, het-

erogeneity).
Note that, up to now, no retrospective pruning is implemented for CUBREMOT , as two natural stopping rules are available: node partitioning stops (i.e. a node is declared terminal) if either none of the available covariates is significant or the sample size is too small to support a CUB model fit.

3 Application

In order to grow a CUBREMOT using the defining splitting criterion in [2, 3], data from the yearly multiscope survey on daily life run by ISTAT in 2015 have been considered. The data set and its detailed description are available at:
www.istat.it/it/archivio/129916.
Here, the chosen response variable is *Trust in EU Parliament* (TEP for short) and it has been collected on a Likert type scale with 11 categories, ranging from 0 = 'I totally distrust it', to 10 = "I have absolute trust in it": as customarily, it has been forward shifted to the range 1-11 for CUB models fitting. For illustrative purposes, only a subset of the available covariates has been given in input to the procedure. Moreover, for the sake of saving space, the CUBREMOT growth has been stopped to three levels and only node 7 and 10 have been declared terminal according to the stopping rules defined in section 2. The tree is displayed in Figure 1, highlighting that the following covariates affect evaluations and discriminate response patterns:

1. *Political Talk* (PT), an ordinal factor with levels from 1= "On daily basis" to 6 = "Never" to assess the frequency of involvement in political talks and discussion;
2. *Economic Satisfaction* (*ES*): an ordinal factor asking interviewees to assess their satisfaction towards their wealth status within the previous 12 months, on a balanced scale with levels 1 ="Very satisfied", 2 = "Fairly Satisfied", 3 = "Little Satisfied" up to 4 = "Not at all satisfied";
3. *General Satisfaction* (*GS*): an ordinal factor asking interviewees to assess their overall life satisfaction on a rating scale ranging from levels 0 ="Not all satisfied" up to 10 = "Extremely Satisfied";
4. *Trust in Italian Parliament* (TIP): an ordinal variable asking respondents to rate their perceived trust in the Italian Parliament, collected on the same scale as the chosen response variable TEP.

For each node, the number of observations n_k, the estimated CUB parameters π and ξ, as well as the dissimilarity between the estimated CUB probabilities for the descending split (*DissB*) are reported while Figure 2 shows fitted (vertical bars) and estimated distributions at selected nodes by reporting also their dissimilarity. We might conclude that *Trust towards the Italian Parliament* plays a prominent role in the understanding and modelling of *Trust towards the European Parliament* but it interacts with the perception of economic well-being and general satisfaction as well as with the direct involvement of the respondents in political talks. In addition, having chosen the CUB paradigm as the root of the model-based approach, nodes

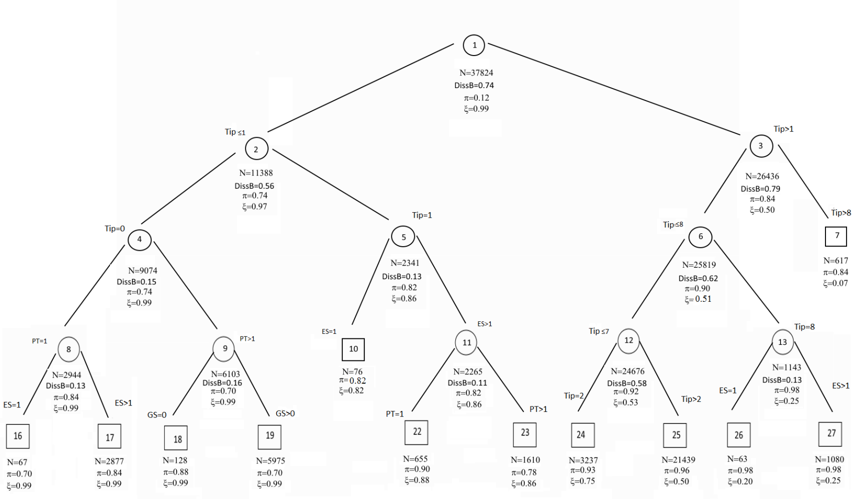


Fig. 1 CUBREMOT for *Trust for European Parliament* assuming a dissimilarity splitting rule

can be discriminated both in terms of trust (feeling) and in terms of indecision and heterogeneity of the respondents. For instance, at the first split the procedure recognizes a major difference between those with an extremely low trust towards the Italian Parliament ($TIP \leq 1$) and those giving higher evaluations: as shown by Figure 2, these two groups correspond to people with extremely low trust towards the EU Parliament and people with intermediate evaluations, respectively. From Nodes 10, 13, one derives that satisfaction for the economic status is associated, in general, with a higher trust (as measured by $1-\xi$) but also with a higher indecision (Node 16). People claiming to be always involved in political talks ($PT = 1$) are in general more resolute and homogeneous in the responses (Node 8 against Node 9) and also less trustful of the European Parliament (Node 22 against Node 23).

4 Final remarks

In a comparative perspective, a tree for the chosen response variable has been grown using the `RpartScore` package [5], which implements the work of [7] to deal with ordinal responses as an extension of the `Rpart` package. In this respect, since it is a common belief that, when the number of categories is high as in the selected case study, the response can be treated as numeric, a tree using standard `Rpart` has also been grown. In both cases, the only splitting variable able to grow the tree is TIP, no other covariate is selected even when relaxing the parameters that control the growth of the tree. On the contrary, the proposed approach allows to disclose several drivers of the responses at different levels and with different strength. Also notice that the

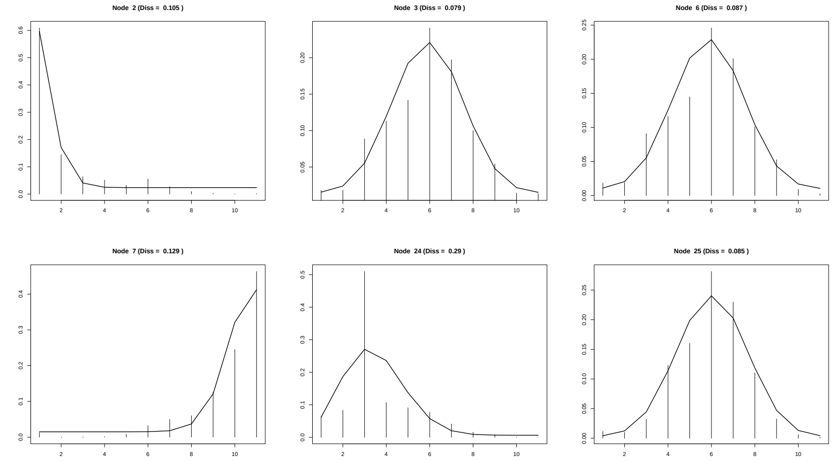


Fig. 2 Observed and fitted probability distributions at selected nodes

dissimilarity-based splitting criterion grows a different tree with respect to the log-likelihood splitting criterion but they both allow to disentangle various determinants of the response assuming specific decision rules on variable importance. Ongoing research involves a deep comparison with other tree-based methods based on simulation studies as well as the implementation of retrospective pruning while future research will be devoted to a more flexible modelling of the node distributions by considering extensions of CUB models.

References

1. Breiman L., Friedman J.H., Olshen R.A, Stone C.J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks: Monterey (CA).
2. Cappelli, C., Simone, R. and Di Iorio, F. Model-based trees to classify perception and uncertainty: analyzing trust in European institutions, *under review*, 2017.
3. Cappelli, C., Simone, R., Di Iorio, F. (2017). Growing happiness: a model-based tree. In SIS 2017. Statistics and Data Science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society, Florence 28–30 June 2017, 261-266.
4. D'Elia A., Piccolo D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**, 917–934.
5. Galimberti G., Soffritti G., Di Maso M. (2012). Classification Trees for Ordinal Responses in R: The RpartScore Package, *Journal of Statistical Software*, **47**, 1–25.
6. Iannario M., Piccolo D., Simone, R. (2017). CUB: A Class of Mixture Models for Ordinal Data. R package, version 1.1.1 http://CRAN.R-project.org/package=CUB.
7. Picarretta R. (2008). Classification trees for ordinal variables, *Computational Statistics*, **23**, 407–427.
8. Leti, G. (1983). *Statistica descrittiva*. Il Mulino, Bologna.
9. Zeileis A., Hothorn T., Hornik K. (2008). Model-Based Recursive Partitioning, *Journal of Computational and Graphical Statistics*, **17**, 492–514.

Constrained Extended Plackett-Luce model for the analysis of preference rankings

Cristina Mollica and Luca Tardella

Abstract Choice behavior and preferences typically involve numerous and subjective aspects that are difficult to be identified and quantified. For this reason, their exploration is frequently conducted through the collection of ordinal evidence in the form of ranking data. Multistage ranking models, including the popular Plackett-Luce distribution (PL), rely on the assumption that the ranking process is performed sequentially, by assigning the positions from the top to the bottom one (*forward order*). A recent contribution to the ranking literature relaxed this assumption with the addition of the discrete *reference order* parameter, yielding the novel *Extended Plackett-Luce model* (EPL). In this work, we introduce the EPL with order constraints on the reference order parameter and a novel diagnostic tool to assess the adequacy of the EPL parametric specification. The usefulness of the proposal is illustrated with an application to a real dataset.

Key words: Ranking data, Plackett-Luce model, Bayesian inference, Data augmentation, Gibbs sampling, Metropolis-Hastings, model diagnostics

1 Introduction

A *ranking* $\pi = (\pi(1), \dots, \pi(K))$ of K items is a sequence where the entry $\pi(i)$ indicates the rank attributed to the i-th alternative. Data can be equivalently collected in the ordering format $\pi^{-1} = (\pi^{-1}(1), \dots, \pi^{-1}(K))$, such that the generic component $\pi^{-1}(j)$ denotes the item ranked in the j-th position. Regardless of the adopted

Cristina Mollica
Dipartimento di Metodi e Modelli per il Territorio, l'Economia e la Finanza, Sapienza Università di Roma e-mail: `cristina.mollica@uniroma1.it`

Luca Tardella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma e-mail: `luca.tardella@uniroma1.it`

format, ranked observations are multivariate and, specifically, correspond to permutations of the first K integers.

The statistical literature concerning ranked data modeling and analysis is reviewed in [3] and, more recently, in [1]. Several parametric distributions on the set of permutations $\mathscr{S}_K$ have been developed and applied to real experiments. A popular parametric family is the *Plackett-Luce* model (PL), belonging to the class of the so-called *stagewise ranking models*. The basic idea is the decomposition of the ranking process into $K-1$ stages, concerning the attribution of each position according to the *forward order*, that is, the ordering of the alternatives proceeds sequentially from the most-liked to the least-liked item. The implicit forward order assumption has been relaxed by [4] in the *Extended Plackett-Luce model* (EPL). The PL extension relies on the introduction of the *reference order* parameter indicating the rank assignment order. In this work, we investigate a restricted version of the EPL with order constraints for the reference order parameter representing a meaningful rank attribution process and we also introduce a novel diagnostic to assess the adequacy of the EPL assumption as the actual sampling distribution of the observed rankings.

2 The Extended Plackett-Luce model with order constraints

2.1 Model specification

The implicit assumption in the PL scheme is the forward ranking order, meaning that at the first stage the ranker reveals the item in the first position (most-liked alternative), at the second stage she assigns the second position and so on up to the last rank (least-liked alternative). [4] suggested the extension of the PL by relaxing the canonical forward order assumption, in order to explore alternative meaningful ranking orders for the choice process and to increase the flexibility of the PL parametric family. Their proposal was realized by representing the ranking order with an additional model parameter $\rho=(\rho(1),\dots,\rho(K))$, called reference order, where the entry $\rho(t)$ indicates the rank attributed at the t-th stage of the ranking process. Thus, ρ is a discrete parameter given by a permutation of the first K integers and the composition $\eta^{-1}=\pi^{-1}\rho$ of an ordering with a reference order yields the sequence $\eta^{-1}=(\eta^{-1}(1),\dots,\eta^{-1}(K))$ which lists the items in order of selection, such that the component $\eta^{-1}(t)=\pi^{-1}(\rho(t))$ corresponds to the item chosen at stage t and receiving rank $\rho(t)$. The probability of a generic ordering under EPL can be written as

$$\mathbf{P}_{\text{EPL}}(\pi^{-1}|\rho,\underline{p})=\mathbf{P}_{\text{PL}}(\pi^{-1}\rho|\underline{p})=\prod_{t=1}^{K}\frac{p_{\pi^{-1}(\rho(t))}}{\sum_{v=t}^{K}p_{\pi^{-1}(\rho(v))}}\qquad \pi^{-1}\in\mathscr{S}_K, \tag{1}$$

Hereinafter, we will shortly refer to (1) as EPL$(\rho,\underline{p})$. The quantities p_i's are the support parameters and are proportional to the probabilities for each item to be ranked in the position indicated by the first entry of ρ.

Differently from [4], we focus on a restriction $\tilde{\mathscr{S}}_K$ of the whole permutation space S_K for the reference order parameter. Our choice can be explained by the fact that, in a preference elicitation process, not all the possible $K!$ orders seem to be equally natural, hence plausible. Often the ranker has a clearer perception about her extreme preferences (most-liked and least-liked items), rather than middle positions. In this perspective, the rank attribution process can be regarded as the result of a sequential "top-or-bottom" selection of the positions. At each stage, the ranker specifies either her best or worst choice among the available positions at that given step. With this scheme, the reference order can be equivalently represented as a binary sequence $\underline{W} = (W_1,\ldots,W_K)$ where the generic W_t component indicates whether the ranker makes a top or bottom decision at the t-th stage, with the convention that $W_K = 1$. One can then formalize the mapping from the restricted permutation ρ to $\underline{W}$ with the help of a vector of non negative integers $\underline{F} = (F_1,\ldots,F_K)$, where F_t represents the number of top positions assigned before stage t. In fact, by starting from positing by construction $F_1 = 0$, one can derive sequentially

$$W_t = I_{[\rho(t)=\rho_{\mathrm{F}}(F_t+1)]} = \begin{cases} 1 & \text{at stage } t \text{ the top preference is specified,} \\ 0 & \text{at stage } t \text{ the bottom preference is specified,} \end{cases}$$

where $I_{[E]}$ is the indicator function of the event E and $F_t = \sum_{v=1}^{t-1} W_v$ for $t = 2,...,K$. Note that, since the forward and backward orders $(\rho_{\mathrm{F}},\rho_{\mathrm{B}})$ can be regarded as the two extreme benchmarks in the sequential construction of ρ, this allows us to understand that $\rho_F(F_t+1)$ corresponds to the top position available at stage t. Conversely, $B_t = (t-1)-F_t$ is the number of bottom positions assigned before stage t and thus, symmetrically, one can understand that $\rho_B(B_t+1)$ indicates the bottom position available at stage t.

The binary representation of the reference order suggests that, under the constraints of the "top-or-bottom" scheme, the size of $\tilde{\mathscr{S}}_K$ is equal to 2^{K-1}. The reduction of the reference order space into a finite set with an exponential size, rather than with a factorial cardinality, is convenient for at least two reasons: i) it leads to a more intuitive interpretation of the support parameters, since they become proportional to the probability for each item to be ranked either in the first or in the last position and ii) it facilitates the construction of a Metropolis-Hastings (MH) step to sample the reference order parameter.

2.2 Bayesian estimation of the EPL via MCMC

Inference on the EPL and its generalization into a finite mixture framework was originally addressed from the frequentist perspective in [4]. Here we consider the original MCMC methods recently developed by [6] to solve Bayesian inference for the constrained EPL.

In the Bayesian domain, the data augmentation with the latent quantitative variables $\underline{y} = (y_{st})$ for $s = 1,\ldots,N$ and $t = 1,\ldots,K$ crucially contributes to make it tractable analytically the inference for the EPL . The auxiliary variables y_{st}'s are assumed to be conditionally independent on each other and exponentially distributed with rate parameter equal to the normalization term of the EPL, see also [5]. For the prior specification, independence of $\underline{p}$ and ρ is assumed together with independent Gamma densities for the support parameters, motivated by the conjugacy with the model, and a discrete uniform distribution on $\tilde{\mathscr{S}}_K$ for the reference order. [6] presented a tuned joint Metropolis-within-Gibbs sampling (TJM-within-GS) to perform approximate posterior inference, where the simulation of the reference order is accomplished with a MH algorithm relying on a joint proposal distribution on ρ and $\underline{p}$, whereas the posterior drawings of the latent variables y's and the support parameters are performed from the related full-conditional distributions. At the generic iteration $l+1$, the TJM-within-GS iteratively alternates the following simulation steps

$$\begin{aligned}
\rho^{(l+1)}, \underline{p}' &\sim \text{TJM}, \\
y_{st}^{(l+1)} | \pi_s^{-1}, \rho^{(l+1)}, \underline{p}' &\sim \text{Exp}\left(\sum_{i=1}^{K} \delta_{sti}^{(l+1)} p_i'\right), \\
p_i^{(l+1)} | \underline{\pi}^{-1}, \underline{y}^{(l+1)}, \rho^{(l+1)} &\sim \text{Ga}\left(c+N, d+\sum_{s=1}^{N}\sum_{t=1}^{K} \delta_{sti}^{(l+1)} y_{st}^{(l+1)}\right).
\end{aligned}$$

3 EPL diagnostic

Simulation studies confirmed the efficacy of the TJM-within-GS to recover the actual generating EPL, together with the benefits of the SM strategy to speed up the MCMC algorithm in the exploration of the posterior distribution. However, we were surprised to verify a less satisfactory performance of the TJM-within-GS in terms of posterior exploration in the application to some real-world examples, such as the famous `song` dataset analyzed by [2]. Since the joint proposal distribution relies on summary statistics, the posterior sampling procedure is expected to work well as long as the data are actually taken from an EPL distribution. So, the unexpectedly bad behavior of the MCMC suggested to conjecture that, for such real data, the EPL does not represent the true (or in any case an appropriate) data generating mechanism. This has motivated us to the develop some new tools to appropriately check the model mis-specification issue.

Suppose we have some data simulated from an EPL model. We expect the marginal frequencies of the items at the first stage to be ranked according to the order of the corresponding support parameter component. On the other hand, although computationally demanding to be evaluated in terms of their closed form formula we expect the marginal frequencies of the items at the last stage to be ranked according to the reverse order of the corresponding support parameter component. After proving such a statement one can then derive that the ranking of the marginal frequencies of the items corresponding to the first and last stage should sum up to

$(K+1)$, no matter what their support is. Of course, this is less likely to happen when the sample size is small or when the support parameters are not so different of each other. In any case, one can define a test statistic by considering, for each couple of integers (j,j') candidate to represent the first and the last stage ranks, namely $\rho(1)$ and $\rho(K)$, a discrepancy measure $T_{jj'}(\pi)$ between $K+1$ and the sum of the rankings of the frequencies corresponding to the same item extracted in the first and in the last stage. Formally, let $\underline{r}_j^{[1]} = (r_{j1}^{[1]},\dots,r_{jK}^{[1]})$ and $\underline{r}_{j'}^{[K]} = (r_{j'1}^{[K]},\dots,r_{j'K}^{[K]})$ be the marginal item frequency distributions for the j-th and j'-th positions, to be assigned respectively at the first [1] and last [K] stage. In other words, the generic entry $r_{ji}^{[s]}$ is the number of times that item i is ranked j-th at the s-th stage. The proposed EPL diagnostic relies on the following discrepancy

$$T_{jj'}(\pi) = \sum_{i=1}^{K} |(\text{rank}(\underline{r}_j^{[1]})_i + \text{rank}(\underline{r}_{j'}^{[K]})_i - (K+1))|,$$

implying that the smaller the test statistics, the larger the plausibility that the two integers (j,j') represent the first and the last components of the reference order. To globally assess the conformity of the sample with the EPL, we consider the minimum value of $T_{jj'}(\pi)$ over all the possible rank pairs satisfying the order constraints

$$T(\pi) = \min_{(j,j')\in\mathscr{P}} T_{jj'}(\pi), \tag{2}$$

where $\mathscr{P} = \{(j,j') : j \in \{1,K\} \text{ and } j \neq j'\}$.

3.1 Applications to real data

We fit the EPL with reference order constraints to the `sport` dataset of the `Rankcluster` package, where N=130 students at the University of Illinois were asked to rank K=7 sports in order of preference: 1=Baseball, 2=Football, 3=Basketball, 4=Tennis, 5=Cycling, 6=Swimming and 7=Jogging. We estimated the Bayesian EPL with hyperparameter setting $c = d = 1$, by running the TJM-within-GS for 20000 iterations and discarding the first 2000 samplings as burn-in phase. We show the approximation of the posterior distribution on the reference order in Figure 1, where it is apparent that the MCMC is mixing sufficiently fast and there is some uncertainty on the underlying reference order. The modal reference order is (7,1,2,3,4,6,5), with slightly more than 0.4 posterior probability. However, when we compared the plausibility of the observed diagnostic statistic with the reference distribution under the fitted EPL, we got a warning with a bootstrap classical p-value approximately equal to 0.011. This should indeed cast some doubt on the use of PL or EPL as a suitable model for the entire dataset. In fact, we have verified that, after suitably splitting the dataset into two groups according to the EPL mixture methodology suggested by [4] (best fitting 2-component EPL mixture

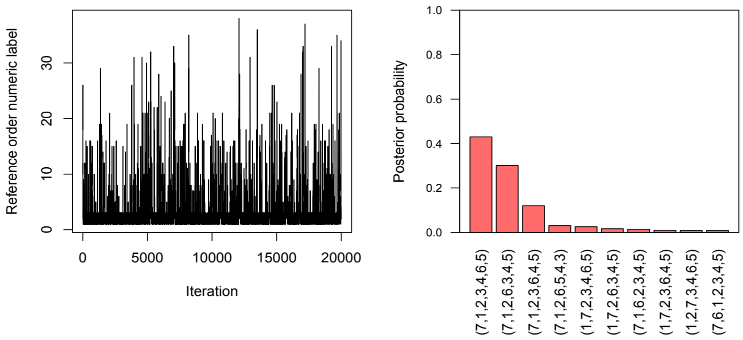


Fig. 1: Traceplot (left) and top-10 posterior probabilities (right) for the reference order parameter.

with BIC=2131.20), we have a different more comfortable perspective for using the EPL distribution to separately model the two clusters. The modal reference orders are (1,2,3,4,5,6,7) and (1,2,3,7,4,5,6) and the estimated Borda orderings are (7,6,4,5,3,1,2) and (1,2,3,4,6,7,5), indicating opposite preferences in the two subsamples towards team and individual sports. In this case, no warning by the diagnostic tests applied separately to the two subsamples is obtained, since the resulting p-values are 0.991 and 0.677.

4 Conclusions

We have addressed some relevant issues in modelling choice behavior and preferences. In particular, we have further explored the idea in [4] related to the use of the reference order specifying the order of the ranks sequentially assigned by introducing monotonicity restrictions on the discrete parameter to describe a "top-or-bottom" attribution of the positions. Our contribution allows to gain more insights on the sequential mechanism of formation of preferences, whether or not it is appropriate at all and whether it privileges a more or less natural ordered assignment of the most extreme ranks. Additionally, some issues experienced when implementing a well-mixing MCMC approximation motivated us to derive a diagnostic tool to test the appropriateness of the EPL distribution, whose effectiveness has been checked with an application to a real example.

References

1. Alvo M, Yu PL (2014). *Statistical methods for ranking data*. Springer.
2. Critchlow DE, Fligner MA, Verducci JS (1991). "Probability models on rankings." *Journal of Mathematical Psychology*, **35**(3), 294–318.
3. Marden JI (1995). *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall. ISBN 0-412-99521-2.
4. Mollica C, Tardella L (2014). "Epitope profiling via mixture modeling of ranked data." *Statistics in Medicine*, **33**(21), 3738–3758. ISSN 0277-6715. doi:10.1002/sim.6224.
5. Mollica C, Tardella L (2017). "Bayesian mixture of Plackett-Luce models for partially ranked data." *Psychometrika*, **82**(2), 442–458. ISSN 0033-3123. doi:10.1007/s11336-016-9530-0.
6. Mollica C, Tardella L (2018). "Algorithms and diagnostics for the analysis of preference rankings with the Extended Plackett-Luce model." *arXiv preprint: http://arxiv.org/abs/1803.02881*.

A prototype for the analysis of time use in Italy
Un prototipo di analisi sull'uso del tempo in Italia

Stefania Capecchi and Manuela Michelini

Abstract Our study focuses on a sub-sample of the Italian time use survey, where respondents are asked to evaluate to what extent each moment of the day is enjoyable. A mixture model framework is implemented to highlight the main components of the data generating process by which interviewees express their point subjective well-being (BESP) towards daily activities. A prototypical proposal is presented to understand interactions between individual evaluation of each activity and subjects' covariates.

Abstract *Lo studio approfondisce i risultati dell'Indagine sull'uso del tempo per un sottoinsieme di rispondenti individuando un modello mistura per la valutaizone espressa dagli intervistati in merito al "Benessere Soggettivivo Puntuale" (BESP). Il prototipo di analisi qui proposto consente di stimare l'effetto congiunto sulle risposte della natura dell'attività svolta e delle caratteristiche dei soggetti anche in termini di eterogeneità.*

Keywords: Time use survey, Point subjective well-being, Mixture models

1 Introduction

Our study aims to investigate people's moods and feelings towards daily activities, stemming from a sub-sample of the Italian time use survey, where respondents are asked to evaluate to what extent each moment of the day is enjoyable, the so-called *Point subjective well-being* (BESP). The following question is considered: "Is this

Stefania Capecchi
Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò, 22, I-80138 Napoli, Italy,
Manuela Michelini
Directorate for Social Statistics and Population Census, Department for Statistical Production, Istat, Viale Liegi, 13, I-00198 Rome, Italy,
e-mail: stefania.capecchi@unina.it; manuela.michelini@istat.it

a pleasant moment?", taking into account the specific activity and the context. Preliminary results are provided exploiting a model to explicitly take into account both feeling and heterogeneity in response patterns and their relationships with activities and subjects' characteristics.

The paper is organized as follows: in the next Section the survey design is briefly sketched; rationale for the prototype model-based approach are presented in Section 3; main results are interpreted in Section 4. Final considerations conclude the work.

2 Main survey features

Time use studies are an important section of current surveys in several contexts [4, 11, 10]. In Italy, time use survey is carried out five-yearly by Istat (the Italian National Institute of Statistics) as a part of an integrated system of social surveys, the Multipurpose Surveys on Households. Referred to the resident population in private households, it is a large sample survey (about $19,000$ families and $45,000$ individuals aged more than 11 years) whose core aim is to learn about the way each respondent allocates her/his time. Interviewees are asked to fill in a daily journal (24 hours, divided into 144 intervals, 10 minutes each), specifying their activities in detail [7]. The survey covers all daily life aspects which, in this study, have been clustered in 10 main activities: *personal care; job, education; houseworking; caregiving; social life; sport and leisure; games and hobbies; TV, radio and reading; travel*. In the following, reference is made to 2014 survey.

Exploratory results provide measures on the amount of time people use to spend in various daily activities on an average day of the year. In 2014 the average day of the population in Italy was as follows: 48.7% of 24 hours was dedicated to personal care (sleeping, eating, and so on), 8.8% to paid work, 3.6 % to study, 12.6% to houseworking, 21% to free time and 5.2% to travel (see [7] for more details).

For the first time, the 2014 survey measures the level of enjoyment associated with daily activities, rating the feeling on a scale from -3 to $+3$, with -3 meaning "not enjoyable at all" and $+3$ "very enjoyable". Work and study turn out to be the least pleasant activities where leisure time is the most enjoyable one, although the rate is decreasing with age. Cultural activities, sports and outdoor leisure, social life, considered as a whole, made the day much more enjoyable.

In fact, individual, activity-related, and temporal dimensions should be jointly considered in a comprehensive model and this study explores such detailed information in order to propose a first probabilistic approach to examine survey results. Given the experimental nature of the proposal, a random sub-sample of 894 adult subjects is considered (aged over 20) and the activities are classified according to the above mentioned 10 groups. Then, individual characteristics as gender, age, education level, marital status, presence of children, number of household members and self-assessed economic condition, are registered. Finally, the day is considered as a working one or a public holiday.

3 Modelling framework for the affective component

A possible approach considers observed data on BESP as the realization of a process by which respondents choose ordinal ratings to denote perceived well-being when involved in each activity. These sequences of affective assessments are the results of a complex interaction of subjective, environmental and time-dependent circumstances. Then, as a first instance, it seems interesting to compute the probability of each rating as a function of the activity itself as well as of the individual characteristics by taking the ordinal structure of responses into account. As a matter of fact, for each i-th subject, a sub-sample of 12 responses –one per hour, out of 144 responses, picked in the time interval from $08:00$ a.m. to $08:00$ p.m. – has been selected. Thus, time effect and serial correlation are partially removed and (almost) conditional independence may be assumed.

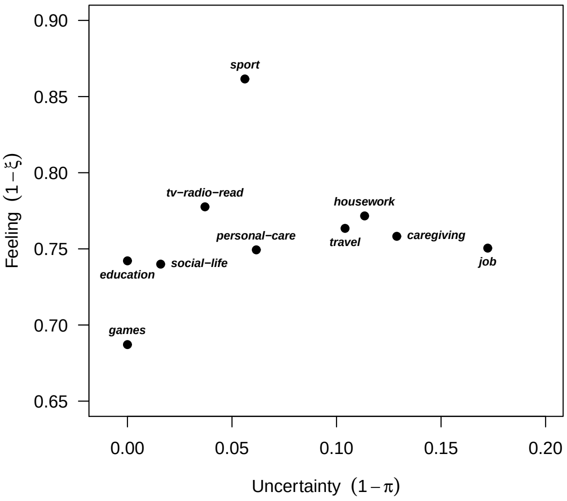


Fig. 1 Estimated models of the expressed enjoyment for different activities

A class of mixture models is applied to parameterize both *feeling* and *indecision/heterogeneity* components in the response pattern [8, 3, 9]; more extensive discussions of the approach –also in different contexts– are in [2, 1]. In particular, the parameters may be easily related to subjects' covariates and to the performed activity. By assuming conditional independence, all the observations may be considered as a whole. These models are estimated and checked by an R package available on the CRAN repository [6] and the visualization of the estimated distributions is an effective added value of the approach.

From a formal point of view, original response R_i^* are transformed into $R_i = R_i^* + 4$ in order to get the first $m = 7$ integers as support. Then, responses are linked to W_i subjects' covariates and A_k activities, for $i = 1,\dots,n$ and $k = 1,\dots,K$ by means of a logistic function.

Thanks to this structure, it is possible to visualize and compare the enjoyment declared by the sampled subjects in performing different activities in terms of both feeling and heterogeneity as in Figure 1. Sports and leisure are the most appreciated activities whereas games and hobbies get the lowest level of feeling. Other mandatory activities, such as those of caregiving and housework, travel and job (the latter with a comparatively higher level of heterogeneity) receive appreciation. In fact, heterogeneity in response patterns is very limited and feeling may be considered as the prominent component in this case study.

4 A more complex modelling structure

A more complex model implies that responses are considered as jointly conditional to the type of activity performed and subjects' characteristics. Thus, the basic model is:

$$Pr\left(R_i = r \mid W_i, A_{i,k}\right) = \pi\, b_r(\xi_{i,k}) + (1-\pi)\,\frac{1}{m}\,, \quad r = 1,2,\dots,m\,, \tag{1}$$

where $b_r(\xi_{i,k})$ is a shifted Binomial distribution, for $i = 1,2,\dots,n$ and

$$logit(1-\xi_{i,k}) = \gamma_0 + \sum_{j=1}^{J} \gamma_j\, w_{i,j} + \sum_{h=1}^{H} \delta_h A_{i,h}\,,$$

Here, the dummy $A_{i,k} = 1$ if the i-th subject is performing the k-th activity in the selected sequences of responses. Then, the comparison between log-likelihoods of nested models (computed at maxima) solves in a likelihood ratio test (LRT) to infer on the significance of the covariates and/or the activities.

Table 1 Log-likelihoods and *LRT* of different models

Models	*Log-likelihood*	*LRT*	*g*
Benchmark	−17196.91		
Activity dummies	−17120.62	152.58	5
Subjects' covariates	−17018.59	356.64	7
Omnibus	−16959.35	475.12	11

Summarizing several intermediate steps, Table 1 reports such investigations and compares a benchmark model (a crude model of all the responses without covariates) with respect to: i) a model including only activities coded as dummies; ii) a model with only subjects' covariates and their possible interactions, if significant; iii) a comprehensive (*omnibus*) model where both subjective variables and activities dummies are considered. LRTs are shown with their degrees of freedom (*g*) and all of them are highly significant.

From the model including only activities (here non reported for brevity), it turns out that :

- Activities gathered in the constant (that is: *personal care –mainly, meals–; job; education; caregiving; social life*) exert the prevailing effect on the responses;
- watching tv, listening to the radio and reading have positive effects on well-being;
- houseworking and sport activities positively affect response pattern;
- quite moderate is the role of mobility;
- playing games and hobbies have an adverse effect.

Although the consideration of activities significantly improves log-likelihood, the role of subjects' covariates is comparatively by far more relevant if considering LRT. Then, the joint specification of these explanatory components implies an *omnibus model* implemented with all the significant covariates, that is:

$$\begin{cases} 1-\hat{\pi} &= \underset{(0.006)}{0.107} \\ logit(1-\hat{\xi}_i) &= \underset{(0.045)}{0.838} + \underset{(0.027)}{0.080} Act1_i + \underset{(0.076)}{0.713} Act2_i - \underset{(0.144)}{0.350} Act3_i + \underset{(0.060)}{0.119} Act4_i \\ & \quad + \underset{(0.029)}{0.109} Act5_i + \underset{(0.035)}{0.191} Gender_i + \underset{(0.035)}{0.146} Married_i \\ & \quad + \underset{(0.044)}{0.272} Holiday_i - \underset{(0.002)}{0.013} Education_i + \underset{(0.009)}{0.070} Components_i \\ & \quad - \underset{(0.045)}{0.213} Gender_i \times Married_i, \ + \underset{(0.057)}{0.134} Married_i \times Holiday_i \end{cases}$$

Here, significant activities are denoted as: `Act1` (*houseworking*), `Act2` (*sport and leisure*), `Act3` (*games and hobbies*), `Act4` (*Tv, radio and reading*), `Act5` (*mobility*); as a consequence, the impact on the constant derives from all the other activities, `Act0` (*personal care; job; education; caregiving; social life*). Thus, *ceteris paribus* with subjects' covariates, the impact on the enjoyment of the activities may be estimated as follows: `Act0`= 0.838; `Act1`= 0.919; `Act2`= 1.552; `Act3`= 0.489; `Act4`= 0.957; `Act5`= 0.947. It is confirmed that the largest and smallest contributions to enjoyment are given by *sport and leisure* and *games and hobbies*, respectively. Notice that holiday significantly impacts on the response with an additional $+0.272$ and a significant positive interaction with married people, estimated as 0.134.

The subjects' characterization influences the positive responses with respect to women, married people and household's components. Gender and marital status interact on the responses so that the final impact, *ceteris paribus*, is zero and 0.191 for unmarried men and women, respectively, and 0.146 and 0.124 for married men and women, respectively. Then, generally speaking, married women express lower enjoyment in their activities.

5 Concluding remarks

Time use survey is considered strategic for the knowledge of the population's lifetime as it allows to detect everyday life organization in society and in families. Moreover, distributions of affective states, expressed through specific tasks and duties evaluation, provide valuable information on people's moods, in general and across different activities. When clustering population groups and specific activities, such an investigation may offer important complements to other measures of well-being.

According to these perspectives, to address and evaluate policy interventions, even in the fields of work-life balance and sustainability, the proposed modelling prototype is suitable to give useful insights to establish priority hierarchies at both individual and aggregated level. Thus, further research should be pursued to exploit the dynamic content of the expressed enjoyment during day as related to different activities. This objective might be achieved by generalizing the proposed models and including both longitudinal aspects and switching expressed modalities.

Acknowledgements. This work has been implemented within CUBREMOT project financially supported by University of Naples Federico II.

References

1. Capecchi, S., Iannario, M., Simone, R. (2018), Well-Being and Relational Goods: A Model-Based Approach to Detect Significant Relationships, *Social Indicators Research*, 135(2), 729–750.
2. Capecchi, S., Piccolo, D. (2015), Investigating the determinants of job satisfaction of Italian graduates: a model-based approach, *Journal of Applied Statistics*, 43(1), 169–179.
3. D'Elia, A., Piccolo, D. (2005), A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, 49, 917–934.
4. European Commission (2009), Harmonised European time use surveys: 2008 guidelines, *Eurostat Methodologies and Working papers*, Luxembourg: Office for Official Publications of the European Communities.
5. Iannario, M., Piccolo, D. (2016), A comprehensive framework of regression models for ordinal data, *METRON*, 74, 233–252.
6. Iannario, M., Piccolo, D., Simone, R. (2018), CUB: A Class of Mixture Models for Ordinal Data. R package version 1.2.0 http://CRAN.R-project.org/package=CUB
7. ISTAT (2016), I tempi della vita quotidiana. Statistiche Soddisfazione dei cittadini. Report, 23 novembre 2016.
8. Piccolo, D. (2003), On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, 85–104.
9. Piccolo, D., D'Elia, A. (2008), A new approach for modelling consumers' preferences, *Food Quality and Preference*, 19, 247–259.
10. Gershuny, J. (2011), Time-Use Surveys and the Measurement of National Well-Being, Centre for Time-Use Research Department of Sociology, University of Oxford.
11. Ricroch, L. (2011), The enjoyable moments of day-to-day life, INSEE Premiere papers, 1378.

New Perspectives in Supervised and Unsupervised Classification

Robust Updating Classification Rule with applications in Food Authenticity Studies

Robust Updating Classification Rule con applicazioni a studi di autenticità degli alimenti

Andrea Cappozzo, Francesca Greselin and Thomas Brendan Murphy

Abstract In food authenticity studies the central concern is the detection of products that are not what they claim to be. Here, we introduce robustness in a semi-supervised classification rule, to identify non-authentic sub-samples. The approach is based on discriminating observations with the lowest contributions to the overall likelihood, following the *impartial trimming* established technique. Experiments on real data, artificially adulterated, are provided to underline the benefits of the proposed method.

Abstract *Negli studi di autenticità degli alimenti risulta cruciale saper riconoscere prodotti contraffatti. In questo paper si adotta un approccio robusto per modificare una regola di classificazione semi-supervised e poter quindi identificare potenziali adulterazioni. L'approccio basato sulla selezione delle osservazioni che danno minore contributo alla verosimiglianza globale, seguendo tecniche ben note di impartial trimming. Esperimenti su dati reali, artificialmente adulterati, evidenziano l'efficacia del metodo proposto.*

Key words: Robust Statistics; Impartial trimming; Model-based classification; Semi-supervised method; Food Authenticity

1 Introduction and Motivation

Nowadays, meticulous consideration is devoted to the food market, therefore, analytical methods for food identification are needed to protect food quality and prevent its illegal adulteration. In a standard classification framework, hypothesized trust-

Andrea Cappozzo • Francesca Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: a.cappozzo@campus.unimib.it; francesca.greselin@unimib.it

Thomas Brendan Murphy
School Of Mathematics & Statistics and Insight Research Centre, University College Dublin, e-mail: brendan.murphy@ucd.ie

worthy learning data are employed to build a decision rule. However, in a context in which the final aim is to detect potentially adulterated samples, also the learning data may be unreliable and thus it can strongly damage the classifier performance [9]. Especially if the training size is small, mislabelled data in the learning phase can be detrimental for the decision phase. The aforementioned problem is known as “label noise” and it is not new in the statistical learning literature: a discussion was already reported in [11]. We refer the reader to [3] for a review of related work on this topic.

Considering the aforementioned issues in dealing with food data, the present work introduces a robust semi-supervised model-based classification method. The methodology arises as a modification of the framework first developed in [4], here endowed with robust techniques. The rest of the manuscript is organized as follows: in Section 2 the proposed Robust Updating Classification Rule is introduced and an EM algorithm for parameter estimation is detailed in Section 3. Section 4 describes the data reduction procedure for the spectra of raw homogenized meat samples; the proposed method is then applied to a scenario with adulterated labels and benchmark results are considered. The paper concludes with some considerations for future research.

2 Robust Updating Classification Rule

The aim of the proposed method is to construct a model where possibly adulterated observations are correctly classified as such, whilst preventing them to bias parameter estimation. To account for useful information about group heterogeneity that may be contained also in the unlabelled samples, we adopt a semi-supervised approach. This methodology was originally developed in [4], the present work employs functional PCA [14] for data reduction and incorporates robust estimation for outlier detection, with the specific role of identifying the adulterated samples. Our conjecture is that the illegal subsample is revealed by selecting observations with the lowest contributions to the overall likelihood, and that impartial trimming [7] prevents their bad influence on parameter estimation for authentic samples. The updating classification rule is modified for providing reliable estimates even when there are mislabelled samples in the training data. Additionally, given the semi-supervised nature of the methodology, outlying observations in the test data can also be discarded in the estimation procedure.

Denote the labelled data by x_n; $n = 1,\dots,N$, and their associated label variables l_{ng}, $g = 1\dots G$ and $n = 1,\dots N$ where $l_{ng} = 1$ if observation n comes from group g and $l_{ng} = 0$ otherwise. Likewise, denote the unlabelled data by y_m, $m = 1,\dots,M$ and their associated unknown labels z_{mg}, $g = 1\dots G$ and $m = 1,\dots M$. Both labelled and unlabelled data are p-dimensional. We construct a procedure for maximizing the *trimmed observed-data likelihood*:

$$L_{trim}(\pi,\theta|\boldsymbol{x}_N,\boldsymbol{l}_N,\boldsymbol{y}_M)=\prod_{n=1}^{N}\left[\prod_{g=1}^{G}(\pi_g f(x_n|\theta_g))^{l_{ng}}\right]^{\zeta(x_n)}\prod_{m=1}^{M}\left[\sum_{g=1}^{G}\pi_g f(y_m|\theta_g)\right]^{\eta(y_m)} \tag{1}$$

where π_g denotes the vector of mixing proportions, θ_g represent the parameters of the gth mixture component and $\zeta(\cdot)$, $\eta(\cdot)$ are 0-1 trimming indicator functions, that tell us whether observation x_n and y_m are trimmed off or not. A fixed fraction α_l and α_u of observations, respectively belonging to the labelled and unlabelled data, is unassigned by setting $\sum_{n=1}^{N}\zeta(x_n)=\lceil N(1-\alpha_l)\rceil$ and $\sum_{m=1}^{M}\eta(y_m)=\lceil M(1-\alpha_u)\rceil$. The less plausible observations, under the currently estimated model, are therefore tentatively trimmed out at each iteration that leads to the final estimate. α_l and α_u represent the *trimming level* for the *training* and *test* set, respectively, accounting for possible adulteration in both datasets. In our approach a final value of $\zeta(x_n)=0$, as well as $\eta(y_m)=0$, corresponds to identify x_n and y_m, respectively, as illegal observations. We consider the case in which $f(\cdot|\theta_g)$ indicates the multivariate normal density distribution, where $\theta_g=(\mu_g,\Sigma_g)$ respectively denotes the mean vector and the covariance matrix in the G mixture components. For performing the maximization of (1), an EM algorithm [5] is employed and different constraints on the eigenvalue decomposition of the covariance matrices are considered for parsimony [1]. In addition, the singularity issues that may be introduced in the case of heteroscedastic covariance matrices (i.e., with volume and/or shape free to vary across components) are avoided considering a restriction on the eigenvalues on the matrices Σ_g. Particularly, we fix a constant $c\geq 1$ such that

$$M_n/m_n\leq c \tag{2}$$

where $M_n=\max_{g=1\ldots G}\max_{j=1\ldots p}\lambda_j(\Sigma_g)$ and $m_n=\min_{g=1\ldots G}\min_{j=1\ldots p}\lambda_j(\Sigma_g)$, $\lambda_j(\Sigma_g)$ being the eigenvalues of the matrix Σ_g. Such restriction leads to a well-defined maximization problem [8].

3 The EM algorithm

The EM algorithm for implementing the robust updating classification rule involves the following steps:

- *Initialization:* set $k=0$. Find starting values by using model-based discriminant analysis. That is, find $\hat{\pi}^{(0)}$ and $\hat{\theta}^{(0)}$ using only the labelled data through standard approaches, such as *mclust* routines [6]. If the selected model allows for heteroscedastic Σ_g and (2) is not satisfied, the constrained maximization is also applied, see [8] for details.
- *EM Iterations:* Denote by $\hat{\theta}^{(k)}$ the parameters at the current iteration of the algorithm.
  - *Step 1 - Concentration*: after computing the quantities $D_g(y_m,\hat{\theta}^{(k)})=\hat{\pi}_g^{(k)}f(y_m|\hat{\theta}_g^{(k)})$, the trimming procedure is implemented by discarding the $\lceil N\alpha_l\rceil$ observations x_n with smaller values of

$$D(x_n|\hat{\theta}^{(k)}) = \sum_{g=1}^{G} f(x_n|\hat{\theta}_g^{(k)})^{l_{ng}} \quad n = 1,\dots,N$$

and discarding the $\lceil M\alpha_u \rceil$ observations y_m with smaller values of

$$D(y_m|\hat{\theta}^{(k)}) = \max\{D_1(y_m|\hat{\theta}^{(k)}),\dots,D_G(y_m|\hat{\theta}^{(k)})\} \quad m = 1,\dots,M$$

- *Step 2 - Expectation*: for each non-trimmed observation y_m the posterior probabilities

  $$\hat{z}_{mg}^{(k+1)} = \frac{D_g(y_m|\hat{\theta}^{(k)})}{\sum_{t=1}^{G} D_t(y_m|\hat{\theta}^{(k)})} \text{ for } g = 1\dots G \text{ and } m = 1,\dots M$$

  are computed.
- *Step 3 - Constrained Maximization*: the parameters are updated, based on the non-discarded observations and their cluster assignments:

  $$\hat{\pi}_g^{(k+1)} = \frac{\sum_{n=1}^{N}\zeta(x_n)l_{ng} + \sum_{m=1}^{M}\eta(y_m)\hat{z}_{mg}^{(k+1)}}{\lceil N(1-\alpha_l)\rceil + \lceil M(1-\alpha_u)\rceil} \quad \text{for } g = 1\dots G$$

  $$\hat{\mu}_g^{(k+1)} = \frac{\sum_{n=1}^{N}\zeta(x_n)l_{ng}x_n + \sum_{m=1}^{M}\eta(y_m)\hat{z}_{mg}^{(k+1)}y_m}{\sum_{n=1}^{N}\zeta(x_n)l_{ng} + \sum_{m=1}^{M}\eta(y_m)\hat{z}_{mg}^{(k+1)}} \quad \text{for } g = 1\dots G$$

  The estimation of the variance covariance matrices depends on the considered constraints on the eigenvalue decomposition [1].
  If $\lambda_j\left(\hat{\Sigma}_g^{(k+1)}\right)$, $g = 1\dots G$, $j = 1\dots p$ do not satisfy (2) the constrained maximization described in [8] must be applied.
- *Step 4 - Convergence of the EM algorithm*: the Aitken acceleration estimate of the final converged maximized log-likelihood is used to determine convergence of the EM algorithm [2]. If convergence has not been reached, set $k = k+1$ and repeat steps 1-4.

The final estimated values $\hat{z}_{mg}$ provide a classification for the unlabelled observations y_m, assigning observation m into group g if $\hat{z}_{mg} > \hat{z}_{mg'}$ for all $g' \neq g$. Final values of $\zeta(x_n) = 0$, and $\eta(y_m) = 0$, classify x_n and y_m respectively, as illegal observations.

4 Meat samples: classification results in presence of adulteration

The algorithm described in Section 3 is employed in performing classification for the meat dataset [10]. This dataset reports the electromagnetic spectrum from a total of 231 homogenized meat samples, recorded from 400-2498 *nm* at intervals of 2 *nm*. Figure 1 reports the spectra for each meat type, measured as the amount of light reflected by the sample at a given wavelength.

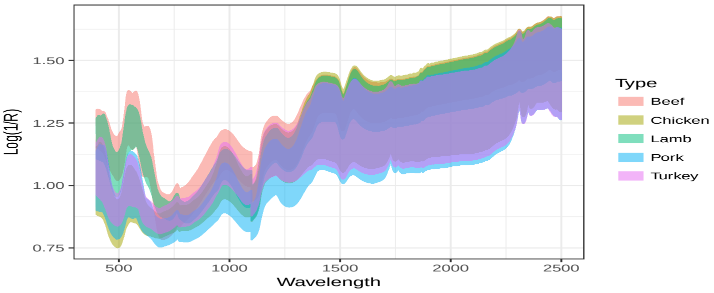


Fig. 1 Functional representation of the NIR spectra of five homogenized meat types, meat dataset

To reduce the dimension of the data using a functional data analysis approach, we perform functional Principal Component Analysis (fPCA) and retain the first 15 scores vectors. Details on the employed procedure can be found in [14].

The robust updating classification rule is employed for classifying the meat samples. To do so, we divided the data into a training sample and a test sample. We investigated the effect of different proportions for the data in terms of classification accuracy. Particularly, 3 split proportions have been considered: 50% - 50% , 25% - 75% and 10% - 90% for training and test set, respectively, within each meat group. Additionally, for each split a 8% of pork observations in the training set were wrongly labelled as beef, for artificially creating an adulteration scenario. Results confronting the misclassification rate for the original [4] and robust updating classification rule are reported in Table 1.

Table 1 Average correct classification rates for the unlabelled five meat groups (after data reduction by using fPCA) for 50 random splits in training and test data, employing robust and non-robust updating classification rule. Standard deviations are reported in parentheses. Results on the original dataset (without adulteration) are reported in the rightmost columns, for a comparison.

	Adulterated Dataset		Original Dataset	
	Upclassify	Robust Upclassify	Upclassify	Robust Upclassify
50% Tr - 50% Te	84.42 (4.49)	91.51 (3.89)	91.20 (3.11)	95.39 (1.74)
25% Tr - 75% Te	79.55 (4.29)	93.55 (1.30)	85.75 (3.80)	93.63 (4.74)
10% Tr - 90% Te	66.37 (8.64)	86.97 (11.87)	78.59 (5.04)	87.24 (6.99)

The average misclassification rates reported in Table 1 highlight the improvement in employing the robust version of the method, whenever noise labels are present in the training set. To compare results between robust and non-robust method, the trimmed observations were classified a-posteriori according to the Bayes rule, and assigned to the component g having greater value of $D_g(y_m, \hat{\theta}) = \hat{\pi}_g f(y_m|\hat{\theta}_g)$. As expected, the negative effect due to mislabelling increases when the training sample size is small. Labelled and unlabelled trimming levels were set equal to 0.1 and 0.05 respectively for the Robust Upclassify method. Interestingly, on average, higher classification rates are obtained for the 25% - 75% training test split: the robust methodology perfectly identified the mislabelled units in each of the 50 splits. For the 50% - 50% and 10% - 90% case the robust method detected on average respectively 85% and 88% of the mislabelled units. According to Mclust nomenclature, the EEE and the VVE models were almost always chosen in each scenario: model selection was performed through *trimmed BIC* [13]. As a last worthy note, in Table 1 we underline the positive impact in terms of classification rate of a small proportion of impartial trimming ($\alpha_l = \alpha_u = 0.05$) also in the case of an unadulterated training sample, fostering the employment of the robust version of the algorithm.

Further research directions will consider the integration of a wrapper approach for variable selection, along the lines of [12], and the adoption of robust mixtures of factor analyzers for jointly performing classification and dimensionality reduction.

References

1. H. Bensmail and G. Celeux. Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition. *Journal of the American Statistical Association*, 91(436):1743–1748, 1996.
2. D. Bohning, E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Inst. Statist. Math*, 46(2):373–388, 1994.
3. C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42:2649–2658, 2009.
4. N. Dean, T. B. Murphy, and G. Downey. Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(1):1–14, 2006.
5. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
6. M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, XX(August):1–29, 2016.
7. L. A. García-Escudero, A. Gordaliza, and C. Matrán. Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2):434–449, 2003.
8. L. A. García-Escudero, A. Gordaliza, and A. Mayo-Iscar. A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, 8(1):27–43, 2014.
9. T. Krishnan and S. Nandy. Efficiency of discriminant analysis when initial samples are classified stochastically. *Pattern Recognition*, 23(5):529–537, jan 1990.
10. J. McElhinney, G. Downey, and T. Fearn. Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *Journal of Near Infrared Spectroscopy*, 7(3):145–154, 1999.
11. G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*, volume 544 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, mar 1992.
12. T. B. Murphy, N. Dean, and A. E. Raftery. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics*, 4(1):396–421, mar 2010.
13. N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299–308, sep 2007.
14. B. W. Ramsay, James, Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2005.

A robust clustering procedure with unknown number of clusters

Una procedura di cluster analysis robusta con un numero di cluster incognito

Francesco Dotto and Alessio Farcomeni

Abstract A new methodology for robust clustering without specifying in advance the underlying number of Gaussian clusters is proposed. The procedure is based on iteratively trimming, assessing the goodness of fit, and reweighting. The forward version of our procedure is initialized with a high trimming level and $K = 1$ populations. The procedure is then iterated throughout a fixed sequence of decreasing trimming levels. New observations are added at each step and, whenever a goodness of fit rule is not satisfied, the number of components K is increased. A stopping rule prevents our procedure from using outlying observations. Additional use of a backward criterion is discussed.

Abstract *In questo lavoro viene introdotta una metodologia per la cluster analysis robusta che non richiede la specificazione a priori del numero di cluster gaussiani. La procedura si basa, iterativamente, sul trimming, la valutazione della bontà di adattamento ed il reweighting. La sua versione forward viene inizializzata fissando un livello di trimming alto e $K = 1$ popolazioni sottostanti. In seguito la procedura viene iterata all'interno di una griglia fissata di livelli trimming decrescenti. Ad ogni passo vengono reinserite osservazioni e, laddove l'adattamento peggiori sostanzialmente, il numero di componenti K viene aumentato. Una regola di arresto garantisce che valori anomali non vengano usati per stimare i parametri. Si discute infine un criterio aggiuntivo di tipo backward.*

Key words: Trimming, Reweighting, Robustness

Francesco Dotto
Univeristy of Rome "Roma Tre", Via Silvio D'amico 77, 00145 Roma, e-mail: francesco.dotto@uniroma3.it

Alessio Farcomeni
Univerity of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma e-mail: alessio.farcomeni@uniroma1.it

1 Introduction

Model based clustering procedures for multivariate data can be inconsistent in presence of contamination. A trimming step is often used to guarantee robustness. Impartial trimming is based on discarding a fixed proportion α of observations lying far from their closest centroid. A detailed review of robust clustering may be found in part II of [7]. The procedure of [10] is based on pre-specifying α and the number of clusters K. Simultaneously fixing the tuning parameters is still an open problem. First, it shall be noticed that the two parameters are clearly intertwined. Indeed the optimal α depends on the chosen K and also the *vice versa* holds. We propose here to use iterative reweighting to obtain robust cluster analysis without having to specify in advance these two tuning parameters. Our approach is related to the forward search philosophy (e.g., [1]), but with substantial differences. The rest of the paper is as follows. In Section 2.1 we briefly review the `tclust` methodology and its reweighted version. Our new proposal for robust clustering is presented in Section 3 while its application to real data example is provided in Subsection 3.2. Finally Section 4 contains the concluding remarks and the further directions of research.

2 Trimming approach to cluster analysis and its reweighted version

2.1 The `tclust` methodology

Within this subsection we briefly present the `tclust` methodology, [10], whose `R` implementation is presented in [9] and its reweighted version introduced in [4]. Let $x_i \in \mathbb{R}^p$ be a sample point, $f(\cdot)$ the multivariate normal density, μ_j and Σ_j be location and scatter parameters, respectively, of the j-th group. Additionally let $g_{\psi_i}(\cdot)$ be the contaminating density and K the number of groups. Then the likelihood function associated to the spurious outliers model is given by:

$$\left[\prod_{j=1}^{K}\prod_{i\in R_j} f(x_i;\mu_j;\Sigma_j)\right]\left[\prod_{i\notin R_j} g_{\psi_i}(x_i)\right] \tag{1}$$

Additionally it must be pointed out that, in equation (1), $R = \bigcup_{j=1}^{K} R_j$ represents the set of the clean observations and is such that $\#R = \lceil n(1-\alpha)\rceil$ and only the clean data give a contribution to the likelihood function, while, noise component, whose likelihood is given by the right hand side of equation (1) give no contribution to the likelihood function. The maximum likelihood estimator of (1) exists if and only if the following condition on the contaminating density holds:

$$\arg\max_{\mathscr{R}} \max_{\mu_j,\Sigma_j} \prod_{j=1}^{k} \prod_{i \in R_j} f(x_i;\mu_j,\Sigma_j) \subseteq \arg\max_{\mathscr{R}} \prod_{i \notin \cup_{j=1}^{k} R_j} g_{\psi}(x_i) \tag{2}$$

As pointed out in [5], condition (2) states that identification of clean observations by maximization of the right hand term of (2) identifies the same observations as would identification of contaminated observations by maximizing the part of the likelihood corresponding to the noise. Thus, once clean observations are identified by maximizing the right hand term of (2), then the contaminated entries are optimally identified.
Additionally, if the condition (2) holds, the MLE of the likelihood function (1) has a simple representation and its maximization reduces to the maximization of:

$$\sum_{j=1}^{K} \sum_{i \in R_j} \log f(x_i;\mu_j,\Sigma_j) \tag{3}$$

2.2 The `tclust` without specifying α in advance

We now focus our attention to two tuning parameters, that are required to be fixed by the user in order to appy the `tclust` methodology: the trimming level α and the number of clusters K. In [4] is introduced a contribution, known as reweighted `tclust` or `rtclust` for the sake of brevity, designed to avoid the specification of the trimming level α. The idea behind such contribution is starting with a high trimming level α_0 the `tclust`, for which efficient computing is possible ([9]). Once the procedure is initialized, L decreasing trimming levels $\alpha_1 > \alpha_2 > \ldots > \alpha_L$ are fixed; then the `rtclust` algorithm proceeds, for each $l = 1, 2 \ldots, L$, as follows:

1. *Initialization:* Set the initial parameters' set $\pi_1^0,...,\pi_k^0,\pi_{k+1}^0$, $\mu_1^0,...,\mu_k^0$ and $\Sigma_1^0,...,\Sigma_k^0$ obtained by applying the `tclust` with a high trimming level α_0.
2. *Reweighting process:* Consider $\alpha_l = \alpha_0 - l \cdot \varepsilon$ with $\varepsilon = (\alpha_L - \alpha_0)/L$ for $l = 1,...,L$

2.1 *Fill the clusters:* Given $\pi_1^{l-1},...,\pi_k^{l-1},\pi_{k+1}^{l-1}$, $\mu_1^{l-1},...,\mu_k^{l-1}$ and $\Sigma_1^{l-1},...,\Sigma_k^{l-1}$ from the previous step, let us consider

$$D_i = \min_{1 \le j \le k} d^2_{\Sigma_j^{l-1}}(x_i, \mu_j^{l-1}) \tag{4}$$

and sort these values as $D_{(1)} \le ... \le D_{(n)}$. Take the sets

$$A = \{x_i : D_i \le D_{([n(1-\alpha_l)])}\} \text{ and } B = \{x_i : D_i \le \chi^2_{p,\alpha_L}\}$$

Now, use the distances in (4) to obtain a partition $A \cap B = \{H_1,...,H_k\}$ with

$$H_j = \left\{ x_i \in A \cap B \;:\; d_{\Sigma_j^{l-1}}(x_i, \mu_j^{l-1}) = \min_{q=1,\dots,k} d_{\Sigma_q^{l-1}}(x_i, \mu_q^{l-1}) \right\}.$$

2.2 *Update cluster weights* The proportion of contamination is estimated by computing

$$\pi_{k+1}^l = 1 - \frac{\#B}{n}.$$

Given $n_j = \#H_j$ and $n_0 = n_1 + ... + n_k$ the cluster weights are estimated by computing:

$$\pi_j^l = \frac{n_j}{n_0}\left(1 - \pi_{k+1}^l\right). \tag{5}$$

2.3 *Update locations and scatters:* Update the cluster centers by taking μ_j^l equal the sample mean of the observations in H_j and the scatter by computing the sample covariance matrix of the observations in H_j multiplied by its consistency factor.

3. *Output of the algorithm:* $\mu_1^L,...,\mu_k^L$ and $\Sigma_1^L,...,\Sigma_k^L$ are the final parameters estimates for the normal components. From them, final assignments are done by computing

$$D_i = \min_{1 \le j \le k} d^2_{\Sigma_j^L}(x_i, \mu_j^L),$$

for $i = 1,...,n$. Observations assigned to cluster j are those in H_j with

$$H_j = \left\{ x_i : d_{\Sigma_j^L}(x_i, \mu_j^L) = \min_{q=1,\dots,k} d_{\Sigma_q^L}(x_i, \mu_q^L) \text{ and } D_i \le \chi^2_{p,\alpha_L} \right\}$$

and the trimmed observations are observations not assigned to any of these H_j sets (i.e., those observations with $D_i > \chi^2_{p,\alpha_L}$).

There are, in our opinion, two great advantages in using `rtclust`. First, as shown in th simulation study and the theoretical properties reported in [4], high robustness with high efficiency can be reached at the same time. Secondly no much tuning is required. Indeed the final estimated contamination level is independent to the initial trimming α_0 and the assumptions on constraint on the eigenvalues can be relaxed after the initialization. It shall be noticed that, besides the required number of groups K - to which are dedicated the further sections of the paper - the parameter α_L my need tuning too. Such parameter establishes how far the outliers are supposed to be placed with respect to the bulk of the data. Such choice is pretty subjective and strongly depends on the context of application only heuristics. We only recall the guidelines provided in [2] for generally tuning in robust statistics and the contribution provided in [6] where an example of the tuning of the parameter α_L is provided.

3 The `tclust` without specifying α and K in advance

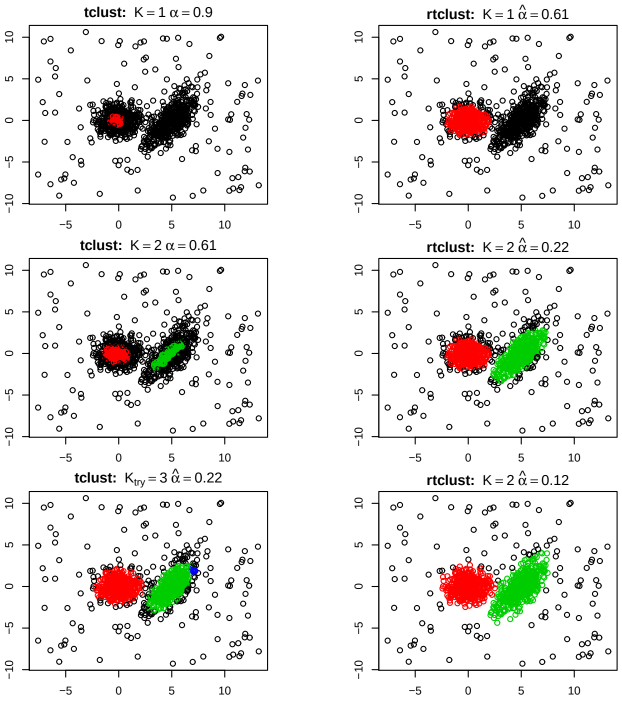
3.1 Introduction

We now outline an automatic methodology based on reweighting that does not need the imposition of the desired number of groups by the user. We do so by applying the reweighting logic to both reinsert the wrongly discarded observations and increase the number of groups, if required. To do so, we resort the forward search philosophy outlined in [1]. In practice, we start by applying the `tclust` method with $K = 1$ and $\alpha = .9$ imposed. Then, we apply the reweighting approach outlined in [4] to estimate the true contamination level $\hat{\varepsilon}$ given $K = 1$ population. Once we can rely on a "precise" estimate of the contamination level we try to increase number of groups imposing $K_{try} = K + 1$ and a trimming level equal to $\hat{\varepsilon}$. The goodness of fit of this new proposed model is evaluated by computing the proportion of observations that are flagged as outlying in the new proposed model that were not flagged as outlying at the previous step. The underlying idea is the following. If a considerably high proportion of observations initially considered clean at the previous step, are recognized as outlying in the current step as a higher number of underlying groups is imposed, the this means that a high dense region of points (a potential cluster) has been trimmed off in the previous step. Algorithmically speaking we alternate the `tclust` and the `rtclust` up to convergence within the steps described in Algorithm 1. A graphical counterpart is provided in Figure 1.

Algorithm 1

```
   Initialization:
1. Fix: $K_0 = 1, \alpha_0 = .9$ and $\rho \in [0.01, 0.05]$
2. Let $mod_{rew}$ be the output of rtclust with $K_0$ and $\alpha_0$ imposed
   Update:
3. Take $\hat{\varepsilon}$ estimated by the model $mod_{rew}$ and set $K_{try} = K + 1$.
4. Launch the tclust with $\alpha = \hat{\varepsilon}$ and $K = K_{try}$ imposed.
5. Take $\pi_{new}$: the proportion of observations flagged as outlying by $mod_{try}$.
   Stopping rule:
6. If $\pi_{new} \leq \rho$ then stop. Else, if $\pi_{new} > \rho$:
   • $K = K + 1$
   • Calculate $mod_{rew}$ by launching rtclust with K imposed.
   • Repeat steps 3-6.
   Final output:
7. Return the output of $mod_{rew}$ as the final output of the algorithm.
```

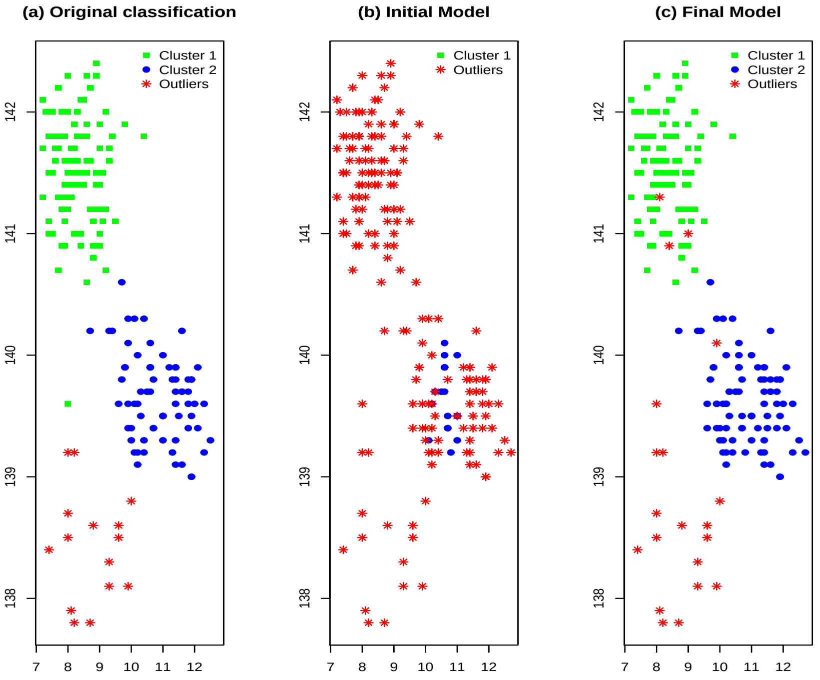
Fig. 1 The application of Algorithm 1 to a 2- dimensional simulated composed by $K = 2$ clusters and a proportion of contaminating points equal to 0.10.



3.2 A real data application

In this Section we apply the proposed iterative reweighting approach to the 6-dimensional "Swiss Bank Notes" data set presented in [8] which describes certain features of 200 Swiss 1000-franc bank notes divided in two groups: 100 genuine and 100 counterfeit notes. This is a well known benchmark data set. In [8], it is pointed out that the group of forged bills is not homogeneous since 15 observations arise from a different pattern and are, for that reason, outliers. As stated in Algorithm 1 we start by imposing a trimming level $\alpha_0 = .9$ and $K = 1$ clusters. The obtained results, that are briefly summarized in Figure 2, are in substantial agreement with characteristics described in [8]. Indeed $K = 2$ are automatically estimated by the algorithm while the estimated proportion of outliers is slightly overestimated: 10% of outliers are recognized by the algorithm while in [8] the declared percentage of outliers is equal to 7.5%.

Fig. 2 Fourth against the sixth variable of the Swiss Bank Notes data set. (a) The original classification. (b) The initial classification obtained by imposing $K = 1$ and $\alpha_0 = .9$. (c) The final output of Algorithm 1



4 Concluding Remarks

We outlined a robust procedure for clustering data that does not need the specification in advance of the required number of groups and of the proportion of outlying observations. We are aware of the fact that, as pointed out in [11], *There are no unique objective "true" or "best" clusters in a dataset. Clustering requires that the researchers define what kind of clusters they are looking for.* In conclusion, as pointed out in [3], we do not think that a fully automatized way to fix simultaneously all the parameters is to be expected. Indeed, the outlined methodology can be viewed as an additional tool to be combined with researchers' specification and a priori informations to provide a better understanding of the phenomenon of interest.

References

1. Atkinson, A.C., Riani, M., Cerioli, A.: Cluster detection and clustering with random start forward searches. Journal of Applied Statistics pp. 1–22 (2017)
2. Cerioli, A., Riani, M., Atkinson, A.C., Corbellini, A.: The power of monitoring: how to make the most of a contaminated multivariate sample. Statistical Methods & Applications pp. 1–29 (2018)
3. Dotto, F., Farcomeni, A., García-Escudero, L.A., Mayo-Iscar, A.: A fuzzy approach to robust regression clustering. Advances in Data Analysis and Classification **11**(4), 691–710 (2017)
4. Dotto, F., Farcomeni, A., García-Escudero, L.A., Mayo-Iscar, A.: A reweighting approach to robust clustering. Statistics and Computing **28**(2), 477–493 (2018)
5. Farcomeni, A.: Robust constrained clustering in presence of entry-wise outliers. Technometrics **56**, 102–111 (2014)
6. Farcomeni, A., Dotto, F.: The power of (extended) monitoring in robust clustering. Statistical Methods & Applications pp. 1–10
7. Farcomeni, A., Greco, L.: Robust methods for data reduction. CRC press (2016)
8. Flury, B., Riedwyl, H.: Multivariate Statistics. A Practical Approach. Chapman and Hall, London (1988)
9. Fritz, H., García-Escudero, L., Mayo-Iscar, A.: tclust: An R package for a trimming approach to cluster analysis. J Stat Softw **47** (2012). URL http://www.jstatsoft.org/v47/i12
10. García-Escudero, L., Gordaliza, A., Matrán, C., Mayo-Iscar, A.: A general trimming approach to robust cluster analysis. Ann Stat **36**, 1324–1345 (2008)
11. Hennig, C., Liao, T.F.: How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C (Applied Statistics) **62**(3), 309–369 (2013)

Issues in joint dimension reduction and clustering methods

Metodi congiunti di riduzione della dimensionalità e classificazione automatica: alcuni aspetti applicativi

Michel van de Velden, Alfonso Iodice D'Enza and Angelos Markos

Abstract Joint data reduction (JDR) methods consist of a combination of well established unsupervised techniques such as dimension reduction and clustering. Distance-based clustering of high dimensional data sets can be problematic because of the well-known curse of dimensionality. To tackle this issue, practitioners use a principal component method first, in order to reduce dimensionality of the data, and then apply a clustering procedure on the obtained factor scores. JDR methods have proven to outperform such sequential (tandem) approaches, both in case of continuous and categorical data sets. Over time, several JDR methods followed by extensions, generalizations and modifications have been proposed, appraised both theoretically and empirically by researchers. Some aspects, however, are still worth further investigation, such as *i*) the presence of mixed continuous and categorical variables; *ii*) outliers undermining the identification of the clustering structure. In this paper, we propose a JDR method for mixed data: the method in question is built upon existing continuous-only and categorical-only JDR methods. Also, we appraise the sensitivity of theproposed method to the presence of outliers.

Abstract *Abstract in Italian* I metodi congiunti di sintesi dei dati (Joint data reduction, JDR) rappresentano una combinazione di approcci di analisi non supervisionata quali tecniche fattoriali e classificazione automatica. La classificazione automatica, quando basata sulla distanza tra le osservazioni, diventa di difficile applicazione quando queste siano descritte da un elevato numero di variabili, a causa della *maledizione della dimensionalità*. Per aggirare tale problema, è pratica comune ridurre la dimensionalità dei dati utilizzando tecniche fattoriali, e applicare successi-

Michel van de Velden
Erasmus University of Rotterdam, visiting Professor at Università della Campania Luigi Vanvitelli
e-mail: vandevelden@ese.eur.nl

Alfonso Iodice D'Enza
Università di Cassino e del Lazio Meridionale e-mail: iodicede@unicas.it

Angelos Markos
Democritus University of Thrace e-mail: amarkos@eled.duth.gr

vamente la classificazione automatica sui fattori ottenuti in precedenza. In letteratura è stato dimostrato empiricamente che a tale approccio sequenziale è preferibile utilizzare dei metodi JDR. Il filone di ricerca sui metodi JDR comprende varianti, generalizzazioni e confronti teorici ed empirici tra i diversi metodi proposti. Tuttavia, alcuni aspetti applicativi non sono ancora stati oggetto di studio: in particolare, si fa riferimento ai casi di dataset che contengano sia variabili quantitative che qualitative, e alla presenza di valori anomali. L'obiettivo del presente contributo è quello di presentare un metodo di JDR che sia applicabile ad insiemi di dati di tipo misto. Inoltre, si vuole valutare la robustezza delle soluzioni ottenute in presenza di valori anomali.

Key words: Dimension reduction, cluster analysis, mixed data sets

1 Joint dimension reduction and clustering methods

Distance-based clustering methods aim at defining groups such that observations that belong to the same group are similar to each other. The distance or dissimilarity measure being used depends on the nature of the considered variables. When the set of observations is described by a large number of variables, it becomes difficult to calculate meaningful pair-wise dissimilarities and, hence, to define clusters. To overcome this problem, methods that combine dimension reduction with cluster analysis have been proposed.

The most straightforward approach is to apply dimension reduction prior to clustering, the latter being therefore applied to the scores obtained in the first step. In this two-step approach, however, two different criteria are optimized: in particular, while dimension reduction aims at defining a reduced set of combinations of the original variables that maximize the original variability, cluster analysis aims at maximize the between-groups variability. This may lead to the cluster masking problem (e.g., van Buuren and Heiser, 1989; De Soete and Carroll, 1994; Vichi and Kiers, 2001) and several solutions have been proposed that proposed a combined optimization of the two steps. We refer to this class of methods as joint data reduction (JDR).

Methods for JDR have been proposed that deal with continuous and categorical data. In particular, for continuous (or, interval) data we consider reduced K-means (De Soete and Carroll, 1994), factorial K-means (Vichi and Kiers, 2001) as well as a compromise version of these two methods. For categorical data, cluster correspondence analysis (van de Velden et al, 2017), which, for the analysis of categorical data, is equivalent to GROUPALS (van Buuren and Heiser, 1989), multiple correspondence analysis and K-means (MCA K-means; Hwang et al, 2006), and iterative factorial clustering of binary variables (i-FCB; Iodice D'Enza and Palumbo, 2013) are considered.

2 Clustering mixed data

Data sets with observations being described by both continuous and categorical variables are common in real applications. Since most clustering procedures are designed to deal with variables on a same scale, a simple strategy is to homogenize the variables in a pre-processing phase. That is, either re-coding the continuous variables or the categorical ones. Recoding of continuous variables is achieved via discretization, the range of each continuous variable is split into a set of intervals, and all the values falling in a same interval are labeled with a same category. Of course, such kind of discretization leads to a loss of the original information. To overcome this issue, an alternative transformation through discretization is to code the original values of a continuous variable into a pre-specified number of fuzzy categories, i.e. to a set of k nonnegative values that sum to 1, quantifying the *possibility* of the variable to be in each category. These *pseudo-categorical* values represent each value of a continuous variable uniquely and exactly, i.e. the numerical information of the original variable is preserved (Aşan and Greenacre, 2011). Recoding of categorical data variables aims to put them on the same scale as the continuous. A rather general pre-processing and standardization approach is described in Mirkin (2012). Another approach, described by Everitt et al (2011), consists of clustering the data by type of variable, and then merge the obtained clustering solutions; the obvious drawback of this approach is that any relation between the two sets of variables is ignored.

A more direct approach is to use a dissimilarity measure designed for mixed data. The most popular one is Gower's dissimilarity coefficient, which takes into account the different nature of the variables (e.g., see Everitt et al (2011)). Once pairwise distances are obtained, a clustering procedure such as partitioning around medoids (PAM) can be applied on the distance matrix. Further partitioning methods for mixed data consist of extensions of K-means: examples are the K-prototypes (Huang, 1998), and the K-means Modha-Spangler weighting (Modha and Spangler, 2003), among others. In all of these approaches, assigning weights to variables is a sensitive task.

Probabilistic or model-based clustering is also a very popular way of clustering mixed-type data. Such methods typically assume the observations to follow a normal-multinomial finite mixture model and proved to be effective when the parametric assumptions are met. In this paper, however, we focus on distance-based clustering approaches.

3 JDR for mixed data

Let $\mathbf{X}$ denote a centered and standardized $n \times Q$ data matrix,, $\mathbf{B}$ is a $Q \times d$ columnwise orthonormal loadings matrix, i.e. $\mathbf{B}'\mathbf{B} = \mathbf{I}_d$, where d is the user supplied dimensionality of the reduced space. Furthermore, $\mathbf{Z}_K$ is the $n \times K$ binary matrix indicating cluster memberships of the n observations into the K clusters. Finally, we use $\mathbf{G}$ to

denote the $K \times d$ cluster centroid matrix.
A JDR method for continuous data is reduced K-means clustering (RKM) De Soete and Carroll (1994): both the dimension reduction and cluster analysis aim at maximizing the *between* variance of the clusters in the reduced space. The RKM objective function is

$$\min \phi_{\text{RKM}}\left(\mathbf{B}, \mathbf{Z}_K, \mathbf{G}\right) = \left\|\mathbf{X} - \mathbf{Z}_K \mathbf{G}\mathbf{B}'\right\|^2, \tag{1}$$

where $\|\cdot\|$ denotes the Frobenius norm. It can be shown, that the above minimization problem is equivalent to

$$\max \phi'_{\text{RKM}}(\mathbf{Z}_K, \mathbf{B}) = trace \mathbf{B}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{B} \tag{2}$$

Similar to RKM is cluster CA, a JDR method for categorical data: CCA aim is also to maximize the between cluster variation in reduced space. Let $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$ be the superindicator matrix of dummy coded categorical data and $\mathbf{M}$ the corresponding centering operator. The objective of cluster CA can be expressed as

$$\max \phi_{\text{clusca}}(\mathbf{Z}_K, \mathbf{B}) = trace \mathbf{B}'\mathbf{Z}'\mathbf{M}\mathbf{P}\mathbf{M}\mathbf{Z}\mathbf{B} \quad \text{s.t.} \quad \frac{1}{np}\mathbf{B}^{'}\mathbf{D}_z\mathbf{B} = \mathbf{I}_k. \tag{3}$$

subject to

Comparing this equation to (2), we see that the methods are closely related. Furthermore, letting $\mathbf{B}^* = \frac{1}{\sqrt{np}}\mathbf{D}_z^{1/2}\mathbf{B}$ it is possible to re-write the clusterCA optimization problem as

$$\min \phi_{\text{CCA}}\left(\mathbf{B}^*, \mathbf{Z}_K, \mathbf{G}\right) = \left\|\mathbf{D}_z^{-1/2}\mathbf{M}\mathbf{Z} - \mathbf{Z}_K\mathbf{G}\mathbf{B}^{*\prime}\right\|^2, \tag{4}$$

subject to

$$\mathbf{B}^{*\prime}\mathbf{B}^* = \mathbf{I}_k$$

which is closely related to the RKM problem in Equation 1, and therefore the two equations can be combined to define the problem for mixed data. In particular from Equations (1) and (4) we can formulate as objective for a joint analysis of mixed data:

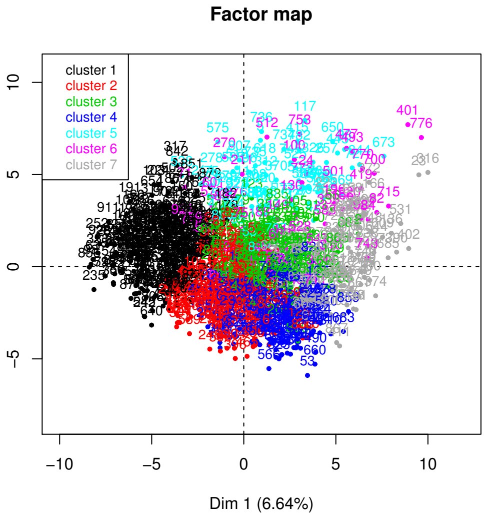
$$\min \phi_{\text{mixed RKM}}\left(\tilde{\mathbf{B}}, \mathbf{Z}_K, \mathbf{G}\right) = \left\|\left(\mathbf{X} \quad \mathbf{D}_z^{-1/2}\mathbf{M}\mathbf{Z}\right) - \mathbf{Z}_K\mathbf{G}\tilde{\mathbf{B}}'\right\|^2, \tag{5}$$

where $\tilde{\mathbf{B}}' = [\mathbf{B}_1' \quad \mathbf{B}_2']$ and $\tilde{\mathbf{B}}'\tilde{\mathbf{B}} = \mathbf{I}_d$.
In a similar way, it is possible to combine FKM with clusterCA and define a further JDR method for mixed data.

4 Clustering patients with low back pain

In this section, we illustrate Mixed RKM on a real dataset. The dataset was part of a clustering challenge connected with the International Federation of Classification

Fig. 1 Mixed RKM factorial map of the patients on the first and second dimension. Clusters are indicated with different colours.



Societies (IFCS) 2017 conference. It contains baseline and outcome assessment of low back pain in 928 adult patients who were consulting chiropractors in Denmark. The clustering aim is to find a (semi-) automatic classification of the patients based on 112 pain history and work-related variables, in order to find clinically applicable and useful groups. 38 of the variables were treated as continuous and 84 as categorical. The data set along with associated meta-data and variable descriptions, can be downloaded at http://ifcs.boku.ac.at/repository/challenge2/.

Missing data were imputed using the regularised iterative FAMD algorithm (Audigier et al, 2016). Mixed RKM was then applied on the imputed dataset. The number of dimensions, 5, was determined on the basis of both empirical and statistical criteria. Solutions between 3 and 12 clusters were investigated. Average Silhouette Width and the Calinski-Harabasz index were used to assess cluster separation. A solution with 7 clusters was selected. Figure 1 depicts the object (patient) scores on the first and second dimension. The seven clusters are indicated with different colours. The description of each cluster in terms of the variables involved in clustering as well as a set of external (outcome) variables was perfomed using the function `catdes()` (package `FactoMineR` Lê et al (2008)). All clusters were significantly associated with external (outcome) variables, which provides evidence for external validity.

Cluster 1 (8.7%): acute suffering LBP, psychological effects of pain, mild improvement after 12m.

Cluster 2 (13.5%): acute suffering LBP, reduced activity, little time in pain, major improvement after 12m.

Cluster 3 (5.2%): leg pain only, showed mild improvement after 12m.
Cluster 4 (9.2%): acute suffering LBP, pain attributed to work, higher bmi, mild improvement after 12m.
Cluster 5 (16.4%): acute suffering LBP, age higher than average, reduced physical activity, little improvement after 12m.
Cluster 6 (12.2%): mild LBP intensity, age less than average, no psychological effects, major improvement after 12m.
Cluster 7 (34.6%): low LBP intensity, major improvement.
Finally, the application of mixed RKM to simulated mixed data, with and without presence of outliers, showed promising results both in terms of effectiveness in cluster structure identification and robustness to outliers.

References

Aşan Z, Greenacre M (2011) Biplots of fuzzy coded data. Fuzzy sets and Systems 183(1):57–71

Audigier V, Husson F, Josse J (2016) A principal component method to impute missing values for mixed data. Advances in Data Analysis and Classification 10(1):5–26

van Buuren S, Heiser W (1989) Clustering n objects into k groups under optimal scaling of variables. Psychometrika 54:699–706

De Soete G, Carroll JD (1994) K-means clustering in a low-dimensional euclidean space. In: Diday E, Lechevallier Y, Schader M, Bertrand P, Burtschy B (eds) New Approaches in Classification and Data Analysis, Springer-Verlag, pp 212–219

Everitt BS, Stahl D, Leese M, Landau S (2011) Cluster analysis. John Wiley & Sons

Huang Z (1998) Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery 2(3):283–304

Hwang H, Dillon WR, Takane Y (2006) An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. Psychometrika 71:161–171

Iodice D'Enza A, Palumbo F (2013) Iterative factor clustering of binary data. Computational Statistics 28(2):789–807

Lê S, Josse J, Husson F, et al (2008) `FactoMineR`: an `R` package for multivariate analysis. Journal of statistical software 25(1):1–18

Mirkin B (2012) Clustering: a data recovery approach. CRC Press

Modha DS, Spangler WS (2003) Feature weighting in *k*-means clustering. Machine learning 52(3):217–237

van de Velden M, DEnza AI, Palumbo F (2017) Cluster correspondence analysis. Psychometrika 82(1):158–185

Vichi M, Kiers HAL (2001) Factorial k-means analysis for two-way data. Computational Statistics and Data Analysis 37:49–64

New Sources, Data Integration and Measurement Challenges for Estimates on Labour Market Dynamics

The development of the Italian Labour register: principles, issues and perspectives

Il registro del Lavoro in Italia: caratteristiche principali e prospettive

C. Baldi, C. Ceccarelli, S. Gigante, S. Pacini

Abstract The ongoing construction of a Labour Register is greatly enhancing the labour statistics either due to the production of output directly or by constituting a coordination framework able to provide a much greater coherence to the entire system for labour statistics. The setting up of such a register in Italy is extremely challenging since it is based on a plurality of administrative sources.
Some pillars of the register: the construction of a set of classifications and coding rules to convert administrative standards into statistical classifications; the definition of the main statistical unit of the register, the job position; the set of measures on labour input to fulfil either directly or indirectly a variety of statistical needs.
Abstract *Il registro tematico del Lavoro, nell'ambito del nuovo sistema integrato dei Registri Istat, sarà il perno delle statistiche sul mercato del lavoro. Una delle sfide principali è legata alla molteplicità di fonti che devono essere integrate per garantire la copertura desiderata in termini di unità e variabili. I pilastri fondamentali del registro di seguito descritti sono: la costruzione di un sistema di metadati per passare da informazioni amministrative a statistiche; la definizione dell'unità base del registro, la posizione lavorativa; un set di misure dell'input di lavoro per soddisfare, direttamente o indirettamente, una serie di output statistici.*

Key words: register, job position, labour statistics, labour input measures, metadata

[1] C. Baldi, Istat; baldi@istat.it
C. Ceccarelli, Istat; clceccar@istat.it
S. Gigante, Istat; gigante@istat.it
S. Pacini, Istat; pacini@istat.it

1 The centrality of labour statistics in the system

Labour indicators occupy a very central place in the statistical system. At the intersection between social and business statistics, employment figures on one side represent the level of participation of individuals in the productive system of a society and the basis for the wellbeing of individuals, households and society as a whole; on the other side they represent the labour input for the production processes of the system of economic units and a factor that drives economic growth. In the same way, given their dual nature (ILO, (1973)), wages represent at the same time the main component of labour cost for the employers and the earning for the employee, the main, or sole, component of personal income for a large share of the population.

Given their importance, in the European statistical system a number of regulations and agreements deals with labour indicators[1]. Employment related figures are also fundamental for the Population Census and some social surveys (LFS, Eu-Silc).

Up to very recently, most national statistical systems organized the production of the above indicators according the classical stove-pipe model, with independent processes, mostly based on traditional surveys with an increasing role of administrative registers. This model has shown several limitations in term of production costs, a large and growing burden on respondents, and various levels and types of inconsistencies among similar indicators. The very attempts of reconciliation have been a post-production exercise usually directed to explain the coarsest discrepancies, still leaving a fragmentation of information in which the users can get easily lost.

The need of releasing integrated instead of disjoint indicators and improving the efficiency of the production processes, in a context of rapid growth of demand of statistical information, is as key points of the strategy of the European Statistical System for the next decade (ESS, (2014)).

Italy is facing this challenge placing at the basis of the production process a system of registers, micro level databases, originated mainly from administrative sources that cover the whole population of statistical units (UNECE, (2017)). In the context of labour statistics, the foundation of the measurement process will be the Labour Register (hereinafter LR). With a complex and rich structure of information, it aims to be the basis for most

[1] Among the labour statistics, beside the household based quarterly Labour Force Survey (LFS), on the business side there are the four yearly Labour Cost and Structure of Earning Surveys, the quarterly Labour Cost Index and Job Vacancy statistics. Moreover, regulations on the business register, structural and short-term business statistics, contain indicators of employment, hours worked, wages, social contributions and labour costs. In the same vein, these variables are central in the National Accounts.

labour market indicators and analysis, requested by international regulations and national needs, either directly using its information for the output, or indirectly, coupled with surveys, to estimate variables not present in it or that do not perfectly fit the statistical definitions.

To be able to accomplish the task of coordinating the entire system of labour statistics and, at the same time, fulfilling the standards required by different regulations and needs, the register, whose main characteristics are describes in § 2, requires at least three main foundations explored in what follows. First a system of metadata, regularly maintained, able to map the administrative information into a variety of statistical classifications (§ 3). Second a multiplicity of statistical units in order to fulfil either the Business or the Social statistics together with precise rules to pass from one to another maintaining the coherence of information (§ 4). Finally a system of measures for the main variables to accommodate different needs (§ 5).

2 The Italian Labour Register: main characteristics

The LR, in the new Integrated System of Registers, is a thematic framework on jobs and their measures in term of labour input, labour cost and income due to work.

The architecture of the LR is based on the main following principles.

Comprehensiveness and completeness. The scope of the register is to cover all regular jobs active in the national territory in all sectors of economic activity, either private or public. Both dependent and independent paid jobs are included. It aims to cover also unpaid work position such as volunteer positions, and positions of persons in training. As for the variables, considering the available sources of information, it contains a number of measures of labour input, working and paid time, wages, social contributions, gross and net income received by each person.

Granularity. Since the register aims to be as flexible as possible to satisfy a variety of needs, it contains information at the maximum level of detail available in the input sources in terms of units, variables and time references. Units and variables will be explored more in detail in § 4 and 3. Here it is important to note that, depending on the variables, there are measures available referred to each day, week, month or year.

Longitudinality. The different statistical units have to be followed over time with rules that defines the continuity. The unit defined in §4 permits to track the relationship between an employer and an employee since its inception and, over time, to construct the working career of each worker within and across employers and status of employment. The longitudinal structure will allow to disentangle, in econometric studies, specific employee from specific employer effects.

Integrability. The register is strictly linked on one hand to the Population Register, through the person ID, and on the other to the Economic Units Registers through the unit ID, with the obvious advantage of matching a lot more information. Moreover, also the Income Register is strictly related. More precisely, they have to be jointly built in order to have measures of labour income, and its components, usable in different conceptual frameworks (from the Camberra group document on statistics on income to the LCS and SNA regulations). The system will also permit to integrate information on samples from households and business surveys.

Extended usability. This infrastructure will be used for a number of scopes: from the construction of standard macro statistical indicators, as mentioned in § 5, to the evaluation of labour policies and the releases of longitudinal microdata standard files for the researchers.

In order to fulfil the previous principles, and in particular to guarantee the coverage of the units and the availability of target variables, a number of different administrative sources have to be integrated. The backbone of the system are the Social Security Institute (INPS) sources, among them the UniEmens declaration[1] is one of the most important. Moreover, the tax declarations, especially those submitted by the economic units for their workers (the new CU declaration) are necessary to guarantee the coverage of both units and variables. In particular the CU, besides checks, is needed to ensure the coverage of not dependent jobs under a threshold of income, and to ensure for variables related to taxation and net income.

One of the main task is to check all this sources and establish the rules to integrate them.

3 A metadata system to bridge administrative and statistical reality

Different regulations requires different classifications and different variables. For instance, although related concepts, wages and salaries as required by LCS and ESA regulations and earnings from the SES regulation have different definitions. One of these differences is related to the payments from the employer to the employee in the face of sickness, maternity,

[1] The UniEmens declaration is referred to the private and public dependent jobs and different information for different subpopulations of workers may be available. Moreover, other sources have to be integrated to cover in particular self-employed workers or worker assimilated to employees such as collaborators.

accidents etc.., which are excluded from the wages and salaries because they are defined imputed social contributions while are included in earnings. Another example regards the classification of work relationships and persons employed in status of employment types of classification. The most representative case is that of project workers that are classified as self-employed in the context of NA, outworkers with respect to the economic units they work for in the BR and dependent contractors, within the more general category of employees, in the incoming ICSE 2018.

This implies that in order the Register may satisfy the different statistical needs a complex system of metadata that maps the administrative information to the different statistical standards has to be built and maintained. This operations requires three main steps: a) a recognition of all the statistical requirements to produce an output metadata system; b) the collection of the metadata of various input sources and their integration in a input metadata system that represent the fiscal-social security reality in a unified way; c) the mapping of the input metadata into the output one. The ideal solution should be to pinpoint the atomic statistical items that can be combined in different ways to get to different definitions. After that an evaluation of which of these tiles are available in the administrative data or may be calculated or estimated, also through other administrative and statistical sources, is needed to map the two information. With reference to the imputed social contribution represented by the payments for sickness, maternity and so at the expense of the employer the Italian social security data have no sufficient details to easily estimate them: right now they are imputed according to the information collected by register-based surveys (in particular the Labour Cost Survey).

4 The statistical units

In order to satisfy the variety of needs stemming from the social and economic studies, the LR will allow analysis based on three main statistical units: the individual, the economic unit, and the job position. This last one, very specific of this register, is the position involving one or two subjects to perform a labour activity.

The job position is defined as the work relationship established between an economic unit and an individual and defined by a starting date. An important attribute of the job position is the ending date, which will be lacking up to the moment of the cessation of the relationship.

In the context of dependent (or assimilated) employment, this statistical unit corresponds to the employment contract between an employer and an employee where the starting date is the date of activation of the contract and the ending date its cessation. Moreover, each job position is characterized by a wealth of other attributes such as the type of contract, the working time, the occupational qualification, the workplace, etc..[1]

The rationale for defining this statistical unit in this way is multiple.

First, it is the basis for analysis of all labour market policies that have an impact on the definition of contracts or formation of new ones. Second, being the natural link with the Compulsory communication (Cc) unit, it is the basis of the extension of the register to the information contained in this source especially useful for short-term estimates of employment dynamic (Rapiti et al,(2018))). Third, the duration of the job position is an essential policy parameter as it makes possible to better evaluate the fragmentation of the employment careers with all its implications. All in all, this statistical unit permits a very rich stock-flow accounting connecting on one side levels of employment, gross and net changes and on the other workers and jobs. Fourth, the business based labour statistics measure parameters referred to job position.

The passage from a statistical unit to another is governed by a set of rules. For instance to characterize, in a given period - say a year, an employee in terms of the type of contract or type of working time it is needed to decide which statistical rules to apply when an employee has had more than one position in the year.

The multiplicity of statistical units and these rules allows to provide alternative representations of the same or related phenomena projected on populations of different units. For instance, in conjunction with the metadata, it permits to analyse the relationships between the inequality of price of labour, defined at the level of job position, to the inequality of earnings among individuals. Moreover, it allows to switch from measures that make sense only for some units (the wages can be defined both at enterprises and at job level) to other that makes sense only for another unit (the income which is related to the individual level).

In the following, some figures are provided to assess the cardinality of the LR in terms of different statistical units. In the year 2015 (Table 1), the segment of the LR referred to the dependent employment in the private sector, is composed of 1.6 millions of enterprises, 14 millions of employees that generate 17.5 millions of job positions.

Table 1: Universe of the dependent jobs in the private sector. Year 2015

[1] Since some of these attributes can change within the duration of a job, the labour register will also allow analysis on another statistical unit, the sub-positions, that is the temporal portion of the job homogeneous for a number of these attributes, type of working time, type of contract, and occupational qualification. This might be very important to evaluate the contract transformations e.g. from fixed term to permanent, that do not constitute an interruption of the same position.

UNITS	*MEASURES*
Employers	1,573,529
Employees	14,004,534
Couples employers / employees	16,075,725
Job positions	17,495,470
Sub-positions	18,752,967

5 The measures of labour input

Different metrics of labour input are necessary to serve different statistical scopes (Table 2). As for dependent jobs, a set of measures are related to the formal activity of each job. The number of jobs formally active in specific dates, along with those activated and closed in an interval, provides the necessary data for the stock-flow accounting that is the number of days each contract has been formally active during a period (e.g. a year). A natural extension of these measures is the average number of formally active job positions in an interval where the average is calculated using as weights the number of days of formal duration in the interval. Those measures are important also because they are the the same used within the Cc system, with the important value added that the LR provides not only the magnitude of the gross and net job flows but also the level of the job stocks.

However, the formal duration of the job may provide wrong signals if one is interested in the effective work input especially for some types of contracts (e.g. job on call). In fact in traditional business statistics and the NA the measure used to calculate the average number of jobs is one where each job is weighted with a quantity that expresses the effective activity (such as the number of days for which the job has been covered by social contributions). Compared with the former measure it reduces dramatically the weights of the contracts such as the jobs on call and, at the same time, corrects the importance of those cases, that might be affected by errors in the dates of activations and cessations.

Besides the measures of jobs stocks and flows, either pointwise or over intervals, the register provides variables to evaluate the time paid by the employer and time worked. The number of hours paid, used primarily as denominator of wages to capture the effective price of labour (e.g. the wage rates), reduces the weight of part-time workers and excludes from the computation the time paid by the social security for events of sicknesses, maternity, accidents, short time allowances etc… Moreover, the time for paid overtime is accounted for. Finally, the number of hours worked, excluding the time spent on holidays and time in the above events paid by the employer, provides the foremost metric to measure the labour input, and as such, is used as denominator in important indicators (e.g. hourly labour cost, labour productivity).

Table 2: Different concepts and measures in the universe of the dependent jobs in the private sector. Year 2015

Concepts	*Measures*
Number of distinct job position active for at least one day in the period	17,495,470
Average number of effectively active Job positions (based on days covered by contributions)	11,381,696
Average number of formally active Job positions (based on days covered by contract)	12,018,315
Number of job position formally active in a determined date e.g. January 1st	11,231,636
Number of jobs activated (activations) in an interval e.g. the first 6 months of the year	3,281,377
Number of jobs terminated (cessations) in an interval e.g. the first 6 months of the year	1,107,739

References

1. ESS vision 2020, Maggio, 2014 http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020
2. ILO Resolution Concerning an integrated system of wages statistics, in Report of the Twelfth International Labour Conference of Labour Statisticians. Geneva, 1973
3. Rapiti, F., Baldi, C., Ichim, D., Pintaldi, F., Pontecorvo, M.E., Rizzi, R., Digging into labour market dynamics: toward a reconciliation of stock and flows short term indicators, SIS, Giugno 2018
4. UNECE, Register-based statistics in the Nordic countries, New York and Geneva, 2007.

DIGGING INTO LABOUR MARKET DYNAMICS: TOWARD A RECONCILIATION OF STOCK AND FLOWS SHORT TERM INDICATORS[*]

F. Rapiti, C. Baldi, D. Ichim, F. Pintaldi, M. E. Pontecorvo, R. Rizzi

Abstract

The recent availability of short-term job flows from Compulsory Communication (Cc) administrative database has increased the publicly disseminated information. While creating opportunities for a deeper labour market understanding, job flows measures might also generate misleading interpretations, especially when compared with stocks statistics. To address this issue and to fill data gaps, the Italian National Institute of Statistics (Istat) promoted an Agreement among Istat, Inps, Ministry of Labour and social policy, Inail and Anpal to produce harmonised, complementary and coherent statistics with three specific outputs: a quarterly joint press release, an annual report and a labour market statistical information system. The main challenge is to reconcile flows and stock data deriving from administrative and statistical sources. This paper illustrates the statistical treatment applied to the Cc database to achieve a better comparability with official statistics used in the quarterly joint press release disseminated since December 2016.

Sintesi

APPROFONDENDO LA DINAMICA DEL MERCATO DEL LAVORO: VERSO UNA RICONCILIAZIONE DEGLI INDICATORI CONGIUNTURALI DI STOCK E FLUSSI

La recente disponibilità di dati congiunturali sui flussi di posizioni lavorative da fonte amministrativa (Comunicazioni Obbligatorie-Co) ha aumentato le informazioni diffuse al pubblico. Pur creando opportunità per una più approfondita comprensione del mercato del lavoro, le misure relative ai flussi potrebbero anche generare interpretazioni fuorvianti, soprattutto se confrontate con le statistiche sugli stock. Per affrontare questo problema e colmare dei data gap nelle statistiche sul mercato del lavoro, l'Istituto Nazionale di Statistica (Istat) ha promosso un accordo con Ministero del lavoro e politica sociale, Inps, Inail e Anpal per produrre statistiche armonizzate, complementari e coerenti. La sfida principale è riconciliare i dati sui flussi e sugli stock derivanti da fonti amministrative e statistiche. Questo paper illustra il trattamento statistico applicato ai dati delle Co, nel contesto del comunicato trimestrale congiunto pubblicato a partire da dicembre 2016, per ottenere una migliore comparabilità con le statistiche ufficiali.

Key words: employment, labour market, stock and flows indicators

1 Introduction

In recent years in Italy there has been an increase in availability of short-term indicators from administrative sources aiming at monitoring the employment trends. Aspiring to favour a better understanding of the labour market, the Ministry of labour and social policy (Mlps) and the National social security institute (Inps) have been disseminating new short-term statistics based on jobs flows. The brand new and interesting job flows indicators have surely enriched the public debate on employment

[*] The authors are very grateful to the members of the Inter-Institutional Technical Working Group of the Mlps, Istat, Inps, Inail and Anpal Agreement who contributed to the development and implementation of the joint quarterly press release and to the building and sharing of the background knowledge behind this paper.

. Sometimes they have also contributed to generate misleading interpretations, especially when compared with Istat traditional Labour force stock statistics (Anastasia, 2017). Different figures provided for (almost) the same phenomenon by various institutions turned out in an overload of information. The natural multidimensionality of the labour market and the consequent use of different indicators only partly explains the users' impression of experiencing an information “excess”. The diffusion of data which are not completely coherent with the concepts, definitions and classifications of official statistics (Istat, 2015) has given rise to misinterpretation, exaggeration and inappropriate interpretation of technically correct numbers. Some media talked about a “war of numbers on employment” focused on the numerical assessment of the recent labour market reform (Giovannini, 2015).

Data, information and knowledge are three different concepts and moving from the first to the latter is a complex process in which official statistic have a responsibility. National statistical authorities have to promote whatever action is necessary to empower public by improving their capacity of extraction of useful knowledge from data. In turn, such actions will increase official statistics credibility. Istat, aware of the key importance of the labour market issue, decided to act for the sake of an informed, transparent and impartial public debate, and promoted a Coordination Agreement among five Institutions Istat, Inps, Mlps, the National Institute for Insurance against Accidents at Work (Inail) and subsequently National Agency for Active Labour Policies (Anpal). The scope is to cooperate to exploit the full potential of the available information assets, to fill information gaps and to disseminate integrated, harmonized, non-redundant and high quality data and analyses. The Agreement, started to be implemented in December 2016, relies on sharing different data sources, metadata and methodology to make, together, a step forward to understand the labour market functioning. The five institutions should jointly release information to empower public to comprehend the real state of employment. There are three specific outputs: a quarterly joint press release with short-term indicators, an annual joint structural report and an underlying statistical information system to be jointly used by the five institutions. By now, the first two outputs are released on a regular basis, while the third is under construction.

The paper focuses on the data integration process applied for the publication of the quarterly joint press release. In particular, the statistical treatment applied to the Compulsory Communication-Cc database (flows data) to achieve comparability with stock based official statistics (Istat) is illustrated. The main challenge is to reconcile flows and stock data derived from administrative and statistical sources.

The structure of the paper is the following. Section 2 synthetizes the recent debate on labour market trends and the relevance of stock and flows data. In Section 3, after a brief description of the Cc database, the statistical treatment of flows data is detailed. Several comparisons with other data sources are presented in Section 4. Some concluding remarks and future work are sketched in Section 5.

2 The recent debate on labour market trends and the importance of stock and flows information

In 2015-16 the public debate on labour market focused on the effects of the policy reform called “Jobs Act” and in particularly on the social security hiring incentive (so called “decontribuzione”) and the new open-ended contract ‘‘Contratto a tutele crescenti’’. In the discussions, stock have been opposed to flow indicators, statistical surveys to administrative data and jobs against persons employed. Some national newspapers even misinterpreted quarter-over-quarter activations changes with quarter-over-quarter employment changes (Anastasia, 2017).

It is useful to remind here that a stock quantity is measured at a specific point in time (or averaged over a period), whereas a flow quantity is measured over an interval. As reported in the next section, the level of employment in a given moment in time is a stock measure, while its change between two moments is a flow measure. More precisely, it is a net flow which can be obtained also from the balance between the gross flows in and out of employment. The analysis of the change starting from the gross flows provides further insights into the labour dynamics revealing the massive gross flows behind the much smaller net flows (highlighting the underlying heterogeneity).

Stock and flows labour indicators are currently produced from two different sources (and points of view): mainly household surveys and data collected from the side of employer. In the first case, they refer to persons (headcount), in the latter to jobs. The most important and used short-term indicators to monitor labour market are Istat Lfs's quarterly average stocks (employment, unemployment, etc.). Based on Oros survey, Istat also releases quarterly indicators of average stocks of jobs. It should be stressed that the Mlps and Inps releases both refer to flows (activations, cessations, transformations). However, the most extensive and long term experience in regularly releasing short- term gross and net flows of jobs is that of Veneto Lavoro-Labour market observatory[1]. As for other international experiences, only Statistics New Zealand and Bureau of labor statistics in USA release short-term quarterly indicators. The first institute releases results with a delay of 12 months and employs administrative data while the latter uses a traditional business survey[2].

The main objective of the quarterly joint press release (hereafter Note) is to present information from different sources in order to provide an integrated overview of employment trends. To reach this aim, a particular attention has been given to metadata information with a systematic review of the concepts, definitions, classifications, differences between data sources and implications of different survey techniques. The analyzed data are those from the Istat's Lfs and Oros processes, the Inps's Observatory on precariousness, the Inail's work injuries and the Mlps's Cc. The latter has been processed in a way to make these indicators as much comparable as possible with stocks statistics (see section 3).

The Note was welcomed by the media and succeeded in clarifying many issues related to the interpretation and comparison of different indicators. It has apparently achieved the result to calm the stormy waters of the public debate about employment trends. Notwithstanding, the complexity of the phenomena still leaded some analysts to require further simplification and invoke a single indicator, possibly "carved in stone". However, in a data driven society in which any private and public decision is based on information, more data and indicators, produced by different providers, means also more freedom to choose and to evaluate, which, in turn, strengthen a pluralistic and democratic society (Alleva, 2017).

3 Jobs, Activations and Cessations. A bit of stock-flow accounting

In this section, after a brief description of the Cc database, the definitions and computations needed to relate Cc flow statistics to stock statistics are illustrated. Since March 2008, the Compulsory Communications System (Cc) encloses all the

[1] http://www.venetolavoro.it.

[2] See https://stats.govt.nz/ and https://www.bls.gov/bdm/.

communications sent by public and private employers to notify activation, extension, transformation and cessation of jobs, as well as business transfers and change of company name. The Cc encloses different forms. The main one is Unilav, through which all private and public employers (excluding the temporary work agencies for agency workers) fulfil the obligation to communicate any job-related event, i.e. activations, extensions, transformations and cessations of jobs. The events must be communicated within the day before their entry into force.

Every form has a Linked Employer Employee Database information structure with individual data regarding both the employer (fiscal code, name, economic activity sector, flag public administration, registered office, place of work) and the worker (fiscal code, surname, name, sex, date and place of birth, citizenship, permit to stay and motivation, place of residence, education level). Jobs are characterized by a common set of information, i.e. contract type, working time, professional qualification, seasonal work, etc. Moreover, specific dates, depending on the kind of event (activation, cessation, training period cessation, extension, transformation), are useful to determine the duration of job. The coverage of the Cc source is really wide and for this reason very useful for statistical purposes. In terms of employers, it includes all public and private employers belonging to all economic sectors with the exception of armed forces. Referring to the contract type it includes: all regular employees; part of self-employed workers (project workers, occasional contractors, self-employed in entertainment sector, business agents, association in participation; stages and socially useful work. Cc excludes also some managerial contracts of public and private corporations. For further details on the Cc see Baldi *et al.* 2011.

In order to derive Cc-based statistical indicators comparable with the more traditional ones, the link between stocks and flows (Davis *et al.*, 2006) is exploited. The Cc's implicit statistical unit is the employer-employee work relationship, here called job and denoted by $\mathbb{J}_j$. A job is characterized by an activation date, $t_a(j)$, and, possibly, a cessation date, $t_c(j)$. Let us consider a fixed period of time *q*, for example a given quarter, identified by beginning, *b(q)*, and ending dates, *e(q)*. Considering a pair q - $\mathbb{J}_j$, in a given date *t*, a job may be active or not; in the period *q*, $\mathbb{J}_j$ may be active at least for one day or not active at all. In notation, the period-indicator function

$$I_j(q) = 0, if\ t_c(j) < b(q)\ or\ e(q) < t_a(j);\ 1, otherwise \qquad [1]$$

signals whether a given job is active during *q*. Similarly, the date-indicator function

$$I_j(t) = 1, if\ t_a(j) \le t \le t_c(j);\ 0, otherwise \qquad [2]$$

signals whether a given job is active at date *t*. It is straightforward to observe that the following identity holds:

$$I_j(t) = I_j(q) * I_j(t)\ \forall j, q \text{ and } \forall\ b(q) \le t \le e(q)\ .$$

It follows that J_t, i.e. the number of active jobs at any date *t*, is given by $J_t = \sum_j I_j(t)$. Consequently, the average number of jobs over the period *q* may be expressed as

$$\begin{aligned} \bar{J}_q &= \frac{\sum_{t=b(q)}^{e(q)} J_t}{e(q) - b(q)} = \frac{\sum_{t=b(q)}^{e(q)} \sum_j I_j(t)}{e(q) - b(q)} = \frac{\sum_j \sum_{t=b(q)}^{e(q)} I_j(t)}{e(q) - b(q)} \qquad [3] \\ &= \frac{\sum_j \sum_{t=b(q)}^{e(q)} I_j(t) * I_j(q)}{e(q) - b(q)} = \Sigma_j \frac{\sum_{t=b(q)}^{e(q)} I_j(t)}{e(q) - b(q)} * I_j(q) = \Sigma_j D_{qj}^s * I_j(q) \end{aligned}$$

where D_{qj}^s denotes the standard duration of a job $\mathbb{J}_j$ in a period *q*, i.e. $D_{qj}^s = \frac{\sum_{t=b(q)}^{e(q)} I_j(t)}{e(q)-b(q)}$ (the standard duration is defined as the number of days the job has been active during

q standardized by the length of (q). Equation [3] states that $\bar{J}_q$ equals the weighted sum of the active jobs where the weights equal the standard durations.

The stock-flow relationship relates the number of jobs, activations and cessations: the difference between the number of jobs at times *t* and *t-1* equals the difference between the number of activations at time *t* and the number of cessations at time *t-1*:

$$J_t = J_{t-1} + A_t - C_{t-1} \qquad [4]$$

where A_t and C_t represent the number of job activations and cessations occurring at a given date *t*, respectively (Figure 1). The asymmetry between A_t and C_t is due to equation [2]. Indeed, a job $\mathbb{J}_j$ has to be considered "active" even in its end date $t_c(j)$. Consequently, a job will be excluded from the jobs stock only the day after $t_c(j)$.

It is easy to see that both J_t and $\bar{J}_q$ may be expressed in terms of a stock of jobs at a time 0 and a difference between activations and cessations. In fact, putting *t* = 0 in equation [4] and then iterating over time *t*, the number of jobs J_t may be expressed as:

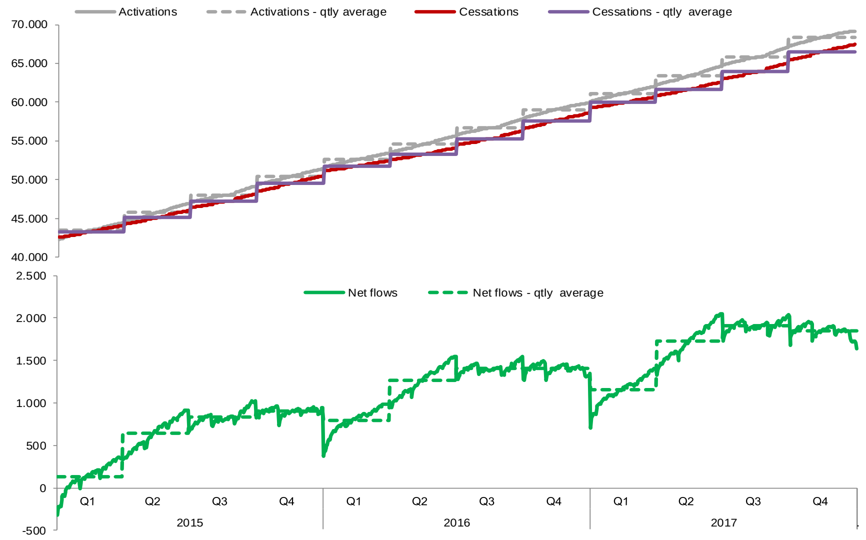
$$J_t = J_0 + \sum_{s=1}^{t} A_s - \sum_{s=0}^{t-1} C_s = J_0 + A_t^u - C_t^u \qquad [5]$$

Averaging over the period *q*, the average number of jobs over *q* may be obtained:

$$\bar{J}_q = \frac{\sum_{t=b(q)}^{e(q)} J_t}{e(q)-b(q)} = J_0 + \frac{\sum_{t=b(q)}^{e(q)} \sum_{s=1}^{t} A_s}{e(q)-b(q)} - \frac{\sum_{t=b(q)}^{e(q)} \sum_{s=0}^{t-1} C_s}{e(q)-b(q)} = J_0 + \overline{A_q^u} - \overline{C_q^u} \qquad [6]$$

where $\overline{A_q^u}$ and $\overline{C_q^u}$ represent, respectively for the activations and cessations, the average over the period *q* of the flows cumulated from the starting point.

Figure 1: Cumulated daily activations, cessations, net flows and quarterly averages Q1.2015 –Q4.2017



The Cc system registers only job activations and cessations. Thus, only the terms A_t^u and C_t^u in [5] and $\overline{A_q^u}$ and $\overline{C_q^u}$ in [6] may be derived using the Cc data. It follows that the number of jobs cannot be calculated within the Cc system. However, the changes of this level may be calculated. Indeed, using [5], the *change between pointwise number of jobs* (CPJ) at the end dates of two consecutive quarters becomes:

$$J_{e(q)} - J_{e(q-1)} = (J_0 + \Sigma_{s=1}^{e(q)} A_s - \Sigma_{s=0}^{e(q)-1} C_s) - (J_0 + \Sigma_{s=1}^{e(q-1)} A_s - \Sigma_{s=0}^{e(q-1)-1} C_s) \qquad [7]$$
$$= \Sigma_{s=b(q)}^{e(q)} A_s - \Sigma_{s=b(q)-1}^{e(q)-1} C_s$$

Furthermore, using [6], the *change of the average number of jobs* (CAJ), becomes:

$$\bar{J}_q - \bar{J}_{q-1} = \left(J_0 + \overline{A_q^u} - \overline{C_q^u}\right) - \left(J_0 + \overline{A_{q-1}^u} - \overline{C_{q-1}^u}\right) = \left(\overline{A_q^u} - \overline{C_q^u}\right) - \left(\overline{A_{q-1}^u} - \overline{C_{q-1}^u}\right) \qquad [8]$$

Notice that CAJ in equation [8] takes into consideration the standard duration of each job in the two quarters, while CPJ in equation [7] does not. Indeed, in equation [7], only pointwise periods, i.e. the end dates of different quarters, are considered. On the contrary, in equation [8], the average number of jobs over a period is involved and $\bar{J}_q$ implicitly takes into account the standard durations, see [3]. It means that CAJ better measures the change of total amount of work between periods.

The two indicators CAJ and CPJ have both weaknesses and strengths. Being based on the average number of active jobs over a period, the CAJ allows for a better comparability to many labour market indicators, for instance, the Labour Force Survey and Oros. From the timeliness point of view, CPJ clearly outperforms CAJ. An unexpected surge of activations at the end of a quarter should be signaled much more evidently by CPJ than by CAJ. However, the meaningfulness of the CPJ signal depends on the duration of the created jobs. If the duration of the activated jobs is considerable, the CPJ could signal such labour input increase soon while CAJ would signal it only the following quarter. If, instead, the created jobs are very short lived, CPJ could give "too strong" and erratic signals that might be misinterpreted as a sign of durable increase in employment. This feature is crucial in a system with a very high share of short-term contracts (Istat et al. 2017). In other terms, average-based indicators, like CAJ, would provide a more accurate measure of the "long-term" impact of such events, but at the cost of a possible delay. The two ways of calculating changes thus prove complementary to each other and their joint use improves the interpretation of the labour market short-term evolution.

4 Comparison of different indicators

This section compares three of the sources analyzed in the quarterly joint Note, Cc (calculated as CAJ), Oros and Lfs, which are more comparable thanks to the length of their time series. These sources differ in several aspects: administrative vs. survey data; stock vs. flows measures, demand side and jobs (Cc and Oros) vs. supply side and persons (Lfs). Thus, different definitions of employment arise: Cc and Oros are based on the relationship between employer and worker established through a formal work contract; Lfs is based on individuals who claim to have done at least one hour of paid work in the reference week. Therefore, the difference between the coverage of the sources must be considered: Lfs includes dependent and independent, regular and irregular employment and all the economic activity sectors (A-U sections of Nace 2007); Cc includes regular employees and some dependent self-employment in the A-U sectors, while Oros includes regular employees in the industry and services sectors (from B to S, excluding O). Thus, in order to correctly compare the sources, the analysis was limited only to the employees in industry and private services (B to N) sectors, excluding temporary agency worker contracts (which are absent from Cc) and job-on-call work because this type of the contract do not imply an actual labor input. The Cc, Lfs and Oros labour market trends are very close. They unequivocally depict the decline in employment during the economic crisis and the troughs of the business cycle at the end of 2013. The subsequent employment recovery appears stronger in the administrative source (Cc). In the year-on-year quarterly changes comparison, the dynamics of jobs in Oros and Cc exhibit a more similar trend. The Lfs trend is more irregular, showing opposite patterns with respect to Cc trend in few quarters and a forward shift in the 2016 employment growth peak (Figure 2). In the absolute cumulative changes, the Cc series shows a divergence with respect to the other two; the drift increases over time especially from the end of 2015 (Fig. 3).

Figure 2: Lfs, Cc (CAJ) and Oros absolute year-on-year absolute quarterly changes in industry and private service (B-N). Q1.2011-Q4.2017. Raw data

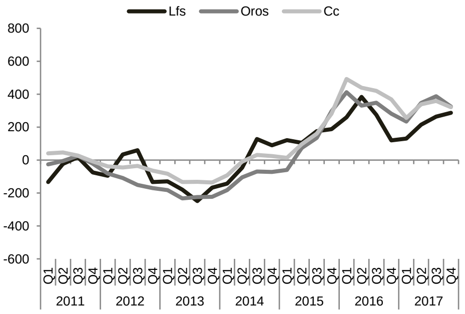
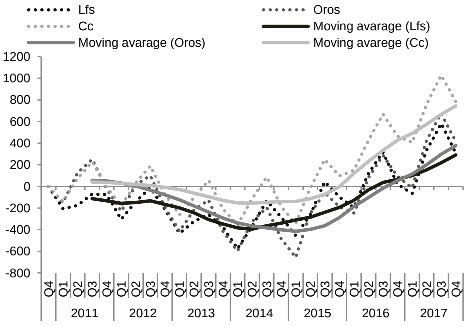


Figure 3: Lfs, Cc (CAJ) and Oros absolute cumulative changes in industry and private service (B-N). Q1.2011-Q4.2017 (Q4.2010=0). Raw data and moving averages



On the other hand, the series of cumulative changes of Lfs and Oros are closer in the last period and slightly compensate the diverging trends in previous years. The differences may be due to the previous definition and measurement issues, as well as to over-coverage and under-coverage errors (De Gregorio *et al.*, 2014). Indeed, administrative data generally tend to have over-coverage errors, e.g. false activations, while survey data more often have under-coverage errors, e.g. sampling and non-sampling errors, (Statistics Canada 2009).

To further investigate the determinants of such differences between trends, a microdata level record-linkage was performed. The Cc and Lfs 2011-2015 microdata were integrated by linking the records corresponding to the individuals having at least one activation in the Lfs reference week. The very preliminary results show that there are cases of employed persons without an employment relationship (persons absent in the Cc data; probably irregular workers) and cases of contracts without work (i.e. unemployed or inactive persons in Lfs with formal contract in Cc[3]). Differences were also found with respect to job duration. There are fixed-term contracts from the Lfs side which are open-ended in the Cc data. On the contrary, a continuous and uninterrupted succession of fixed-term contracts with the same employer can be considered "de-facto open-ended contracts" (Istat-Cnel, 2017). In general, individuals tend to declare their own employment status rather than the formal one. These first results suggest that the differences between the de-facto (from Lfs) and the formal (Cc) conditions should be considered. In some cases, such differences may indicate different and more complex phenomena rather than simple inconsistencies.

5 Conclusions and future development

The Agreement signed in December 2015 represents a real step forward in the collaboration among the five Italian main producer of labour market statistical

[3] This may be caused by temporal misalignment between work and contract, especially for short-term jobs and different irregular practices (fictitious work i.e. false assumptions only to get social security benefits).

information. In the quarterly joint Note the combination of comparable variables on clear and unambiguous domains and use of precise definitions and transparent methodology enhance the release of more coherent statistics. The statistical treatment of the Cc data source allows the estimation of the change of average number of active jobs over a period (CAJ), an indicator more comparable and consistent with traditional statistics on stocks. However, comparing the sources, non-negligible deviations in trends remain. Such deviations are due to definitional and conceptual differences and to many other factors that must be further investigated. A more systematic micro integration of the sources is going to be carried on. This is one of the tasks of the inter-institutional technical working group created to implement the Agreement.

The integration of sources at micro level involves a new phase of literacy to statistical data, oriented to the complexity of the modern labour market phenomena. Sometimes, it is not much a question of distinguishing between “right” and “wrong” data but rather the correct interpretation of its meaning. The differences between the sources must be carefully analyzed, being able to discern between: i) unwanted and (theoretically) avoidable errors (e.g. non-sample survey errors, wrong reporting in questionnaire or in administrative form); ii) irregular or false declarations for interest of the individual or firm; iii) diversity of meaning of the phenomenon. The non-complete comparability between data from different sources has always been one of the main obstacles to their integrated use. This issue becomes even more relevant when developing an integrated statistical system based on microdata. In this context, consistency at micro level becomes crucial. Furthermore, the ability to distinguish the reasons causing the differences and the importance of define hierarchies among sources are fundamental. The final objective is to derive an integrated and validated micro dataset which enables deeper labour market analyses founded on the full exploitation of the richness of integrated variables (profession, education level, etc.).

References

Alleva G., “Troppe statistiche? Inevitabile”, Il Messaggero, 14 March 2017.

Anastasia B., (2017), “Le statistiche aiutano a capire il mercato del lavoro o servono a confondere?” in Economia e Società Regionale n. 3.

Baldi C., De Blasio G., Di Bella G., Lucarelli A., Rizzi R.: *Turning the Compulsory Communication data into a statistical system,* Springer International Publishing Switzerland 2014 “Studies in Theoretical and Applied Statistics” (p.215-224).

Davis S. J., Faberman R. J., Haltiwanger J., The Flow Approach to Labor Markets: New Data Sources and Micro–Macro Links, Journal of Economic Perspectives, 2006, 20 (3), pp 3–26.

De Gregorio C., Filipponi D., Martini A., Rocchetti I., (2014), *A comparison of sample and register based survey: the case of labour market data. Proceedings of the European conference on quality in official statistics* (Quality 2014), Vienna.

Giovannini R., “La guerra dei numeri sul Jobs Act è appena cominciata”, in Internazionale, 23 April 2015.

Istat 2015, “Audizione del Presidente dell’Istat G. Alleva all’Ufficio di Presidenza della Commissione Lavoro del Senato”, 30 September.

Istat, Ministry of labour and social policy, Inps, Inail, Anpal, 2017, “Il mercato del lavoro. Verso una lettura integrata”, https://www.istat.it/it/archivio/206846.

Istat-Cnel (2017), Bes 2017. *Il benessere equo e sostenibile in Italia*, Rome, Istat.

Paolini O., Fortis M. (2017) “Una proposta per fermare l’overdose di statistiche”, Il Messaggero, 4 March.

Statistics Canada 2009 “Statistics Canada Quality Guidelines”. Third Edition - October 1998.

How effective are the regional policies in Europe? The role of European Funds

Efficacia delle politiche regionali in Europa. Il ruolo dei Fondi Europei

Gennaro Punzo, Mariateresa Ciommi, and Gaetano Musella

Abstract The European Structural and Investment Funds are the leading policy instrument through which the EU encourages growth-enhancing conditions for less well-off regions in order to make territorial cohesion within countries. In this work, we perform the Difference in Differences technique to assess the effectiveness of the 2007-2013 EU funding in achieving the convergence in employment levels across NUTS2 regions. Controlling for the socio-economic background at the regional level, a special focus is devoted to Italy. Some empirical results suggest that the EU funding was not effective enough to help the convergence for most countries.

Abstract *I Fondi Strutturali e di Investimento Europei rappresentano il principale strumento attraverso cui l'UE promuove la crescita delle regioni meno sviluppate e indirizza il processo di coesione territoriale nei singoli paesi. Con l'ausilio del metodo delle differenze nelle differenze, il lavoro valuta il contributo dei fondi europei, relativi al periodo di programmazione 2007-2013, alla convergenza dei livelli di occupazione tra le regioni NUTS2. Particolare attenzione è dedicata al caso Italia tenendo conto del background socio-economico delle regioni. Si rileva un ruolo non sempre decisivo dei fondi europei nella realizzazione del processo di convergenza territoriale interno ai diversi paesi dell'Unione.*

Key words: Employment, EU funding, Diff-in-Diff, Regional convergence

[1] Gennaro Punzo, University of Naples "Parthenope", Department of Economic and Legal Studies, email: gennaro.punzo@uniparthenope.it

Mariateresa Ciommi, Università Politecnica delle Marche, Department of Economic and Social Sciences, e-mail: m.ciommi@univpm.it

Gaetano Musella, University of Naples "Parthenope", Department of Management and Quantitative Studies, email: gaetano.musella@uiparthenope.it

1 Introduction

Narrowing socio-economic disparities between richer and poorer regions is one of the key principles on which the European Union (EU) has been based ever since its inception (Art. 158 of the founding Treaty, art. 130a of the Maastricht Treaty). From this perspective, the EU policies promote growth-enhancing conditions for the least-developed regions and the European Structural and Investment Funds (ESI Funds) are the most intensively instruments used by the EU institutions for encouraging the territorial convergence. During the last programming cycles (2000-2006, 2007-2013, 2014-2020), most of the ESI Funds is channelled through the European Fund for Regional Development (EFRD), the European Social Fund (ESF) and the Cohesion Fund (CF) with the aim of leading the less well-off regions to a real convergence within countries [10] and making the EU's system of market integration viable [5].

As regards employment, territorial convergence consists in lowering disparities in employment rates between regions through a spiral of economic growth. In the current programming period (2014-2020), a new legislative framework for these funds has been brought forward and a new set of rules clearly links the ESI Funds with the EU 2020 Strategy for smart, sustainable and inclusive growth [10]. Indeed, at least three of the five headline goals of the EU 2020 Strategy relate directly to employment and productivity, with a focus on the target of "new skills for new jobs". In particular, within the Strategy's objective of sustainable growth, EU 2020 fosters a high-employment economy that delivers social and territorial cohesion. In this respect, the European Commission sets the first target that 75% of the population aged 20-64 should be employed by 2020, and within the seven flagship initiatives, the Commission puts forward "an agenda for new skills and jobs" [9] to empower people by developing skills throughout the lifecycle with a view to better match labour supply and demand.

In this paper, having regard to the close links between the different programming periods of the ESI Funds (the progress reached in a period lays the basis for the time after), we investigate the impact of the EU funding on employment at regional level within each EU country over the period 2007-2013. Some researchers have analysed the relationship between the structural funds of the previous programming cycles and the economic convergence at national and regional levels with conflicting results (see, among others, [4], [13], [5]), which often also depend on the empirical strategy used [2]. By performing the Difference-in-Differences (DiD) strategy, we assess the extent to which the EU funding has supported the convergence for employment between 2007 and 2013, paving the way for the first headline goal of the EU 2020 Strategy. Later, controlling for the socio-economic background at the regional level, a special concern is devoted to Italy with the aim of exploring the main reasons behind the controversial impact of EU funds on regional employment.

With specific reference to the 2007-2013 funding period, NUTS2 regions with a per capita GDP lower than 75% of the EU average were eligible under the convergence objective[2]. Because of the enlargement to EU-25, a phasing-out support was assured to regions[3] that were below the threshold of 75% of GDP for the EU-15 while they were above the same threshold for the EU-25. Based on the Council Regulations (No 1080/2006 and 1083/2006), the convergence objective concerned 84 regions within 18 Member States of EU-27 with a total of 154 million inhabitants, and on the "phasing-out" basis, other 16 regions within 8 countries with a population of 16.4 million. The amount allocated for ESI Funds was assigned to the convergence regions (70.47%) and "phasing-out" regions (4.95%), while the remaining part concerned the Cohesion Fund (24.58%).

2 Methods and data

The effect of the regional funds on local employment and their contribution to convergence across EU NUTS2 regions is exploited by the DiD strategy [3]. This method evaluates the differential effect of a treatment by comparing the change over time in the outcome for the treated group with the change that has occurred in the same timespan for the control group.

The DiD strategy allows us to compare the NUTS2 regions that received the EU funding (treated group) between 2007 and 2013 with those regions that did not (control group). Both groups of regions are analysed before (2000-2006) and after (2014-2016) the 2007-2013 funding period. The DiD model is drawn as follows:

$$y = \alpha_0 + \alpha_1 S + \delta_0 T + \delta_1 T \cdot S + X\beta + u \qquad (1)$$

where y is the *Nx1* vector of the outcome variable. S is a dummy variable that takes on value 1 for the recipient regions and 0 otherwise; α_1 accounts for the difference between the two groups prior to treatment; α_0 is the intercept. T is a dummy whose value is 1 for the period of treatment and 0 otherwise; δ_0 captures the time effects, and thus the changes in the outcome variable in the absence of the treatment. The interaction term $(T \cdot S)$ is the same as a dummy equal to 1 for those treated units in the treatment period; δ_1 represents the treatment effect. X is the *NxK* matrix of

[2] The entire national territories of Bulgaria, Estonia, Latvia, Lithuania, Malta, Poland, Romania and Slovenia were eligible for funding. Instead, a selection of NUTS2 regions was eligible for the following EU countries: Czech Republic (Střední Čechy, Jihozápad, Severozápad, Severovýchod, Jihovýchod, Střední Morava, Moravskoslezsko); Germany (Brandenburg-Nordost, Mecklenburg-Vorpommern, Chemnitz, Dresden, Dessau, Magdeburg, Thüringen); Greece (Anatoliki Makedonia, Thraki, Thessalia, Ipeiros, Ionia Nisia, Dytiki Ellada, Peloponnisos, Voreio Aigaio, Kriti); Spain (Andalucía, Castilla-La Mancha, Extremadura, Galicia); France (Guadeloupe, Guyane, Martinique, Réunion); Hungary: Közép-Dunántúl, Nyugat-Dunántúl, Dél-Dunántúl, Észak-Magyarország, Észak-Alföld, Dél-Alföld); Italy (Calabria, Campania, Puglia, Sicilia); Portugal (Norte, Centro, Alentejo, Região Autónoma dos Açores); Slovakia (Západné Slovensko, Stredné Slovensko, Východné Slovensko); the United Kingdom (Cornwall and Isles of Scilly, West Wales and the Valleys).

[3] Belgium (Province du Hainaut); Germany (Brandenburg-Südwest, Lüneburg, Leipzig, Halle); Greece (Kentriki Makedonia, Dytiki Makedonia, Attiki); Spain (Ciudad Autónoma de Ceuta, Ciudad Autónoma de Melilla, Principado de Asturias, Región de Murcia); Austria (Burgenland); Portugal (Algarve). Italy Basilicata; the United Kingdom (Highlands and Islands).

covariates and β the vector of coefficients. Finally, $u \sim N(0, \sigma_u^2)$ is the *Nx1* vector of uncorrelated errors.

Formally, δ_1 is given by the following difference in differences:

$$\delta_1 = [E(y|x, S = 1, T = 0) - E(y|x, S = 0, T = 0)] - [E(y|x, S = 1, T = 1) - E(y|x, S = 0, T = 1)] \tag{2}$$

Let $\bar{y}_{S,T}$ be the average of the outcome variable in the group of regions S at time T, the estimation of the treatment effect (diff-in-diff) is:

$$\hat{\delta}_1 = (\bar{y}_{1,0} - \bar{y}_{0,0}) - (\bar{y}_{1,1} - \bar{y}_{0,1}) = \hat{\Delta}_0^{\bar{y}} + \hat{\Delta}_1^{\bar{y}} \tag{3}$$

Thus, $\hat{\Delta}_0^{\bar{y}}$ is the difference of the averages of the outcome variable *before* the EU funding between recipient and control regions, and $\hat{\Delta}_1^{\bar{y}}$ is the difference of the averages *after* the EU funding between recipient and control regions.

The DiD model requires that the assumption of parallel paths is satisfied. It postulates that the average change in the control group represents the counterfactual change in the treated group if there was no treatment. In other words, without treatment, the average change for the treated would have been equal to the observed average change for the controls (for details, see [14]).

Our analysis draws upon official data from Eurostat, Istat, and SIEPI[4]. The data, which are related to NUTS2 regions, cover a seventeen-year period, spanning from 2000 to 2016. The outcome variable, the Total Employment Rate (EMRT), is the percentage of employed persons aged 20-64 in relation to the working-age population. The covariates are: *i*) Total Population by Educational Attainment Level (PEALT), which represents the share of the total population with tertiary education (ISCED 5-8); *ii*) Active Labour Market Policies (ALMP), expressed as the average annual number of beneficiaries of active policies; *iii*) per capita Gross Domestic Product (per capita GDP); *iv*) Total Early Leavers from Education and Training (ELETT), which is the share of 18 to 24 year olds having attained ISCED 0-2 level and not receiving any formal or non-formal education or training; *v*) People at-Risk-of-Poverty or Social Exclusion (PRPSE), which is the percentage of the total population below either the poverty threshold (60% of the national median equivalised disposable income after social transfers) or severely materially deprived or living in a household with a very low work intensity; *vi*) five dimensions of the Institutional Quality Index (IQI), which are Corruption, Government effectiveness, Regulatory quality, Rule of law, Voice and accountability. More specifically, Corruption measures the degree of corruption of those performing public functions and crimes against the public administration; Government effectiveness evaluates the quality of public service and the policies of local governments; Regulatory Quality measures the ability of government to promote and formulate effective regulatory interventions; Rule of law quantifies the crime levels, shadow economy, police force, and magistrate productivity; Voice and accountability assesses the degree of freedom of the press and association [15].

[4] Società Italiana di Economia e Politica Industriale.

3 Empirical findings

We analyse the role of the EU funding in fostering homogeneity in employment across NUTS2 regions within countries. Once the treated group (regions exposed to convergence objective) and the control group (regions not included in the objective) are defined, we perform the DiD estimations. In doing so, we compare the average employment rates across the two groups of regions before and after the provision of EU funds. First, we test the effectiveness of the EU funding in 10 EU countries[5] for a total of 256 NUTS2 regions (84 treated and 172 untreated) through DiD null models (Section 4.1). Second, we estimate the DiD model for Italian regions (5 treated and 16 untreated) by controlling for a set of covariates (Section 4.2).

3.1 The European context

Before assessing the effect of EU funds on the convergence objective, we verify the parallel-path assumption by using the Mora&Reggio's approach [14]. It tests whether, in the absence of the EU policy, the averaged EMRTs of the two groups of regions follow the same trend. Based on our results, the parallel-trend assumption is met for each country covered (Tables 1-2), which means that the group of control regions may be considered as a suitable counterfactual for the group of treated regions[6].

Tables 1-2 show the estimations of DiD models by country without controlling for covariates. They allow us to evaluate exclusively the differential effects of EU funds on EMRTs between the treated and untreated regions. Our evidence suggests that the EU funding did not influence significantly the convergence process for employment in most countries. In fact, except for Germany and Italy, the insignificance of $\hat{\delta}_1$ coefficients shows that, on average, the differences between the employment rates of the control and treated groups look much the same before and after the provision of EU funds. It could be assumed that in some countries (Austria, Czech and Slovak Republics, and the United Kingdom) convergence processes were already underway, though at different timing and extent, before the entry into force of the EU policy. However, between 2007 and 2013, the average employment levels have even raised faster in the untreated regions of Belgium and Hungary while they have decreased for both groups of regions in Spain and Portugal, widening in either case the territorial divergence within country.

[5] We exclude France and Greece from the analysis for reasons of data availability.

[6] This is a quite remarkable when one considers that the economic recession reached its highest intensity before the end of 2009 in most EU countries in the middle of the 2007-2013 programming cycle and that these EU funds were not designed to offset the adverse effects of the crisis. Therefore, although some Member States have been more vulnerable than others (i.e., the countries of southern Europe), the parallel-trend test claims that the averages of employment rates of the treated regions would follow the same trend of those of the control regions even during the years of the crisis. Probably, the crisis has had a quite pervasive impact on employment levels within each country without significant differences in the averages of employment rates between the less- and more-developed regions.

Table 1: Diff-in-Diff estimates by country. Model without covariates, 2000-2016

	Austria	***Belgium***	***Czech Republic***	***Germany***	***Spain***
Before					
Control	70.80	66.29	77.26	69.80	67.38
Treated	70.84	57.60	70.04	66.57	58.58
Diff (T – C)	0.05	-8.69***	-7.17***	-3.24***	-8.81***
	(1.12)	(1.79)	(1.18)	(.59)	(.87)
After					
Control	74.67	68.20	77.35	76.38	66.19
Treated	73.83	58.62	71.82	75.63	57.20
Diff (T – C)	-0.84	-9.58***	-5.53***	-0.73	-8.99***
	(.94)	(1.50)	(.99)	(.45)	(.73)
Diff-in-Diff	-0.89	-0.89	1.65	2.51***	-0.18
	(1.46)	(2.34)	(1.54)	(.74)	(1.14)
Parallel Assumption	3.52	4.86	0.74	1.89	0.18

**Significant at 10%; **Significant at 5%; ***Significant at 1%; Standard errors in brackets.*

Table 2: Diff-in-Diff estimates by country. Model without covariates, 2000-2016

	Hungary	***Italy***	***Portugal***	***Slovak Republic***	***United Kingdom***
Before					
Control	66.64	64.34	72.20	75.94	75.45
Treated	60.33	48.18	72.38	62.18	72.34
Diff (T – C)	-6.31***	-16.17***	0.18	-13.76***	-3.11***
	(2.18)	(1.03)	(1.08)	(1.30)	(.89)
After					
Control	68.33	66.71	68.58	76.10	75.74
Treated	61.87	47.10	69.37	65.01	74.31
Diff (T – C)	-6.46***	-19.61***	0.79	-11.09***	-1.42*
	(1.82)	(.85)	(.90)	(1.08)	(.74)
Diff-in-Diff	-0.15	-3.45***	0.62	2.66	1.68
	(2.84)	(1.33)	(1.40)	(1.69)	(1.16)
Parallel Assumption	1.51	0.41	0.50	6.94	3.21

**Significant at 10%; **Significant at 5%; ***Significant at 1%; Standard errors in brackets.*

Italy and Germany show opposite effects of the 2007-2013 funding. In Germany, the employment divide between the treated and untreated regions is practically disappeared after the treatment, and more importantly, the EU target of 75% of the population employed has been reached, on average, in both groups of regions long before the date appointed by the EU 2020 Strategy. Instead, in Italy, the two groups of regions show an increased gap in average rates of employment such that the diff-in-diff estimate is even negative. Italian regions exposed to convergence retain, on average, significant lower levels of employment than those of the control group: they still have a long way to go before converging in employment.

3.2 The Italian context

With the aim of exploring the main reasons behind the potential failure of the EU funding policies in Italy and how much of the lack of regional convergence can be attributed to the incapacity of the Country and its recipient regions to generate economic growth, we perform a DiD model by including a set of control variables (see Section 2). In fact, the impact of EU funds on economic development cannot avoid some institutional, political, and socio-economic factors.

The analysis is carried out by using a log-log model in which each coefficient (except for those of the IQI dimensions) represents the percentage change in the employment rate for a percentage change in the given covariate (i.e., elasticity of the dependent variable with respect to covariates). In the preliminary step, we resort to the instrumental variables (IV) method to overcome the endogenous relationships between the employment rates and per capita GDP and some specific dimensions of IQI. In detail, we use exogenous covariates and their lagged versions as instruments. The Sargan-Hansen's *J* test leads to conclude with the exogeneity, and thus, the validity of these instruments.

Table 3: Diff-in-Diff estimates, Italy. Model with covariates, 2004-2012

	Coeff	*S.E.*
per capita GDP	0.280***	.023
PEALT	-0.041	.026
ALMP	-0.005*	.003
ELETT	-0.036**	.016
PRPSE	-0.016***	.005
Corruption	0.135***	.021
Government Effectiveness	0.162***	.026
Regulatory Quality	0.115***	.032
Rule of law	-0.012	.016
Voice and Accountability	-0.066*	.037

**Significant at 10%; **Significant at 5%; ***Significant at 1%;*

The employment growth is positively related to the GDP level (Table 3) in accordance with the literature that investigated this topic for the most developed countries (see, among others, [6; 16; 18]). It is definitely clear the negative effect on regional performance of low human capital and poor/deprived/excluded people. However, the active labour market policies, which could be regarded as a proxy of the degree of unionisation of a country [12], seem not to help the unemployed in finding jobs. A significant share of the observed gap in employment between regions is also attributable to differences in the quality of institutions [1; 17; 8]. In particular, three out of the five dimensions of institutional quality (i.e. corruption, government effectiveness and regulatory quality) suggest that a more efficient legal system and a lower propensity to corruption improve the employment performance of regions.

4 Conclusions

Except for Germany, the convergence process seems not to be closely linked to the EU funding. Countries whose labour markets are characterised by hybrid patterns (either polarised or upgraded) with a joint contraction of low-, middle- and high-skill jobs [7; 11] even show wider divergences at the end of 2007-2013 period. Italy experienced a reduction in employment rates throughout the entire territory and an increasing territorial divergence that states the failure of EU funding. Inefficiencies of national policies and the poor quality of institutions reflected in bureaucracy and mismanagement of funds, which enable synergies between different funding sources with restriction in national funds when a larger availability of EU funds existed [5].

References

1. Acemoglu, D., Robinson, J.: The role of institutions in growth and development. World Bank, Washington DC. (2008)
2. Arbia G., Le Gallo J., Piras, G.: Does evidence on regional economic convergence depend on the estimation strategy? Outcomes from analysis of a set of NUTS2 EU regions. Spatial economic analysis 3(2), 209-224 (2008)
3. Ashenfelter, O.: Estimating the Effect of Training Programs on Earnings. Review of Economics and Statistics 60, 47-57 (1978)
4. Bähr C.: How does sub-national autonomy affect the effectiveness of structural funds?. Kyklos 61(1), 3-18 (2008)
5. Becker S.O., Egger P.H., Von Ehrlich M.: Effects of EU Regional Policy: 1989-2013. Reginal Science and Urban Economics 69, 143-152 (2018).
6. Boltho, A., Glyn, A.: Can macroeconomic policies raise employment. International Labour Review 134, 451-470 (1995)
7. Castellano R., Musella G., Punzo G.: Structure of the labour market and wage inequality: evidence from European countries. Quality & Quantity 51(5), 2191-2218 (2017)
8. Di Liberto, A., Sideri, M.: Past dominations, current institutions and the Italian regional economic performance. European Journal of Political Economy 38, 12-41 (2015)
9. European Commission: An agenda for new skills and jobs: A European contribution towards full employment (2010)
10. European Commission: European Structural and Investments Funds 2014-2020: Official text and commentaries. Luxembourg: Publication Office of the European Union (2015)
11. Garofalo A., Castellano R., Punzo G., Musella G.: Skills and labour incomes: how unequal is Italy as part of the Southern European countries? Quality & Quantity 1-30 (2017)
12. Huo J., Nelson M., Stephens J.: Decommodification and activation in social democratic policy: resolving the paradox. Journal of European Social policy 18, 5-20 (2008)
13. Mohl P., Hagen, T.: Do EU structural funds promote regional growth? New evidence from various panel data approaches. Regional Science and Urban Economics 40(5), 353-365 (2010)
14. Mora, R., Reggio, I.: didq: A command for treatment-effect estimation under alternative assumptions. Stata Journal 15(3), 796-808 (2015)
15. Nifo, A., Vecchione, G.: Do institutions play a role in skilled migration? The case of Italy. Regional Studies 48(10), 1628-1649 (2014).
16. Padalino, S., Vivarelli, M.: The employment intensity of economic growth in the G-7 countries. International Labour Review 136, 191-213 (1997)
17. Rodríguez-Pose, A.: Do institutions matter for regional development? Regional Studies 47(7), 1034-1047 (2013)
18. Seyfried, W.: Examining the relationship between employment and economic growth in the ten largest states. Southwestern Economic Review 32, 13-24 (2011)

Labour market condition in Italy during and after the financial crises: a segmented regression analysis approach of interrupted time series

Le condizioni del mercato del lavoro in Italia nel periodo durante e successivo alle crisi finanziarie: un'analisi basata sulle serie temporali interrotte

Lucio Masserini and Matilde Bini

Abstract
One of the most widely recognized indicators of the labour market condition is a rising unemployment rate. In Italy, after the 2008 global financial crisis and the 2012 European sovereign debt crisis, this indicator continuously increased over time until late 2014, after which it seems to happen a trend reversal. The aim of this paper is to assess the existence a significant trend reversal in the unemployment rate after 2014, by analysing quarterly data collected from the Italian National Institute of Statistics using a segmented regression analysis approach of interrupted time series. In particular, the analysis is carried out considering some subpopulations of interest, by stratifying unemployment rate for age groups, in order to examine youth unemployment, gender and macro-regions. Moreover, a focus is given to the analysis of the percentage of people Not Engaged in Education, Employment or Training, to provide a more in-depth analysis of the labour market.

Abstract *Uno degli indicatori più ampiamente riconosciuti come misura della recessione è il tasso di disoccupazione. A partire dalle due recenti crisi mondiali, quella finanziaria globale del 2008 e quella del debito sovrano europeo del 2012, questo indicatore in Italia è costantemente aumentato fino a quando al termine del*

[1] Lucio Masserini, Statistical Observatory – University of Pisa; email: lucio.masserini@unipi.it
Matilde Bini, European University of Rome; email: matilde.bini@unier.it

2014 subisce un'inversione di tendenza, ciò forse dovuto ai recenti interventi locali introdotti nel mercato del lavoro. Lo scopo di questo lavoro è di valutare l'esistenza di un'inversione di tendenza significativa dell'indicatore dopo il 2014, analizzando i dati trimestrali raccolti dall'Istituto Nazionale di Statistica Italiano, utilizzando un approccio di analisi di regressione segmentato delle serie temporali interrotte. In particolare, l'analisi è stata realizzata considerando alcune sottopopolazioni di interesse, stratificando il tasso di disoccupazione per fasce di età, al fine di esaminare la disoccupazione giovanile, sul genere e sulle macro-regioni. Inoltre, un'attenzione particolare è stata data all'analisi della percentuale di persone non impegnate nell'istruzione, nell'occupazione o nella formazione per fornire un'analisi più approfondita del mercato del lavoro.

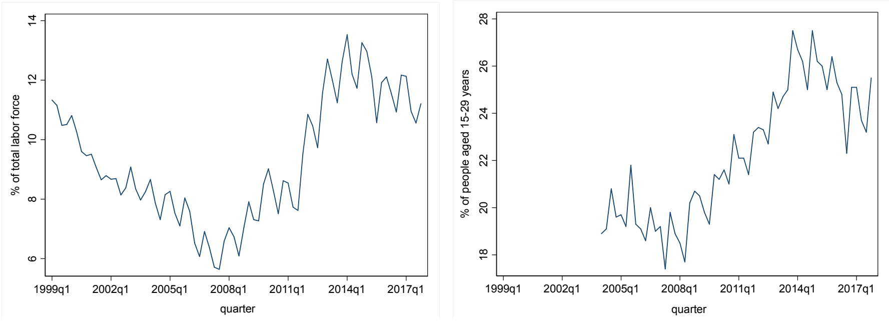
Key words: unemployment rate, interrupted time series analysis, segmented regression, NEET

1 Introduction

The recent two economic crises occurred around the world during the period from late 2000s to late 2011, produced negative effects on countries' economies, particularly on GDP growth, on labour productivity and on labour market. As revealed by the International Labour Organization (ILO, 2001), already since the 2009 about 22 million people were unemployed worldwide in particular in developed economies and in the European Union. The unemployment rate continued towards a dramatically increase with high and persistent levels of unemployment until 2014 after which, it seems to happen a trend reversal, maybe due to the recent domestic interventions introduced in the labour market. The aim of this paper is to assess the existence of a significant trend reversal in the unemployment rate after 2014, using a segmented regression analysis approach of interrupted time series. Quarterly data were collected from the website of the Italian National Institute of Statistics. the analysis is carried out considering some subpopulations of interest, by stratifying unemployment rate for age groups, in order to examine youth unemployment, gender and macro-regions. Changes in the trend are also evident in the percentage of a particular sub group of young people, defined as neither employed nor in education or training (NEET). This could mean that local interventions produced effects in some extent also to this category.

2 Data and empirical strategy

Data were collected from I.Stat, the warehouse of statistics currently produced by the Italian National Institute of Statistics (ISTAT) which provides an archive of about 1,500 time series (http://dati.istat.it/). Quarterly data on two different kind of indicators were downloaded from the theme ‘Labour and wages’: UR for the period 1993–four quarter of 2017, overall and stratified by gender, age groups and macro-regions; and the percentage of NEET for the period 2004– four quarter of 2017, overall and stratified by gender. Such data are derived from the official estimates obtained in the Labour force survey, carried out on a quarterly basis interviewing a sample of nearly 77,000 households representing 175,000 individuals. According to the Eurostat definition (Eurostat, 2017), UR is given by the number of people unemployed as a percentage of the labour force. The youth unemployment rate (YUR) is the number of unemployed 15–24 years-old expressed as a percentage of the youth labour force, and the NEET refers to the percentage of people aged between 15 and 29 years who currently do not have a job, are not enrolled in training or are not classified as a student. Figure 1a illustrates the trend in the overall UR in Italy from 1999 to the end of 2017. The choice of such a long period allows for a more accurate estimate of the secular trend, and this will be useful for the subsequent analysis. Looking at the graph, UR steadily declines until the third quarter of 2007 (2007q3). Then, starting from the fourth quarter of 2007 (2007q4), period in which the effects of the financial crisis following the bankruptcy of Lehman Brothers begin to appear, UR undergoes a first shock and shows a trend reversal. Afterwards, UR increases even more dramatically up to 13.26% at the end of 2014 (2014q4), after the European sovereign-debt crisis occurred in the late 2011 (2011q4). After this peak, UR seems to show a new trend reversal, showing a possible structural change.



(a) (b)

Figure 1: Total UR (a) and percentage of NEET (b) in Italy over the observed period

Figure 1b illustrates the trend of the overall percentage of NEET in Italy from the first quarter of 2004 (2004q1) to the four quarter of 2017 (2017q4). Here the series is shorter because data from previous years are not available. This trend seems, at least partly, to be similar to that of UR. Indeed, after a slight decrease in the period before the onset of the global financial crisis (2007q3), the percentage of NEET starts a steeply and steady growth that continues unchanged also after the occurrence of the sovereign debt crisis (2011q3). And as in UR, a trend reversal occurred starting from the end of 2014 (2014q4).

In the light of the previous considerations, the analysis period was divided into the following four sub-periods: the period before the so-called 2008 global financial crisis (until 2007q3); the subsequent three-year period known as the Great Recession aftermath of the financial crisis, characterised by a general economic decline observed in world markets (2007q4–2011q3); the period following the European sovereign debt crisis, which resulted in a second economic recession (2011q4–2014q4); and finally, the last three years (2015q1–2017q4), during which it seems to glimpse a slight decrease in both the UR and the percentage of NEET. Consequently, the three breaks in the series were set in 2007q4, 2011q4 and 2015q1. Therefore, the analysis period is limited to the years 1999q1–2017q4 for the UR and to 2004q1–2017q4 for the percentage of NEET. The identified interruptions allow to highlight: the severity of the two financial crises, respectively, the continuation of their effects in the subsequent years of recession, and the more recent trend reversal occurred after the end of 2014.

3 Interrupted time series analysis

In this study, a segmented regression approach of interrupted time series (ITS) analysis was carried out in order to assess and measure, in statistical terms, whether and how much the two financial crises have changed the level and trend in the outcome variables, immediately and over time, and to see if these changes are short- or long-term (Wegner, Soumerai and Zhang, 2002).

ITS analysis (Shadish, Cook and Campbell, 2002) is a simple but powerful tool used in quasi-experimental designs for estimating the impact of population-level or large scale interventions on an outcome variable observed at regular intervals before and after the intervention. In such circumstances, ITS allows to examine any change on the outcome variable in the post-intervention period given the trend in the pre-intervention period (Bernal, Cummins and Gasparrini, 2016). In this respect, the underlying secular trend in the outcome before the intervention is determined and used to estimate the counterfactual scenario, which represents what would have happened if the intervention had not taken place and serves as the basis for comparison. For the purposes of our study the interventions are given by two unplanned and real-world events, the aforementioned and well recognized financial crises, and by the subsequent already mentioned trend reversal. In segmented regression of ITS, each sub-period of the series is allowed to exhibit its own level and trend, which can be represented by the intercept and slope of a regression model, respectively. The intercept indicates the value of the series at the beginning of an observation sub-period; and the slope is the rate of change during a segment (or sub-period). Therefore, by following this approach it is possible to compare the pre-crisis level and trend with their post-crisis counterpart, in order to estimate the magnitude and statistical significance of any differences.

The ITS regression model with a single group under study (here, the Italian population), two interventions, which in this study are given by the two economic recessions in 2007q4 and 2011q4, and a possible UR trend reversal in 2015q1, can be represented as it follows (Linden and Adams, 2011; Bernal, Cummins and Gasparrini, 2016):

$$y_t = \beta_0 + \beta_1 T_t + \beta_2 x_{t2007q4} + \beta_3 T_{t2007q4} x_{t2007q4} + \beta_4 x_{t2011q4} + \beta_5 T_{t2011q4} x_{t2011q4} + \\ + \beta_6 x_{t2015q1} + \beta_7 T_{t2015q1} x_{t2015q1} + \varepsilon_t .$$

In particular, y_t is the aggregated outcome variable at each equally-spaced time-points t, here represented by quarters; T_t is the time elapsed since the start of the study, where t varies between 1999q1 to 2017q4 for UR and between 2004q1 to 2017q4 for NEET, respectively; $x_{t2007q4}$ is a dummy variable indicating the onset of the global financial crisis in fourth quarter of 2007, coded as 0 (pre-crisis period) and 1 (post-crisis period); $T_{t2007q4}x_{t2007q4}$ is the interaction term between time and the 2007q4 global financial crisis; $x_{t2011q4}$ is a dummy variable indicating the onset of the 2011q4 European sovereign debt crisis, coded as 0 (pre-crisis period) and 1 (post-crisis period); and $T_{t2011q4}x_{t2011q4}$ is the interaction term between time and 2011q4 European sovereign debt crisis. Finally, $x_{t2015q1}$ is a dummy variable indicating the time in which a new trend reversal occurred, coded as 0 (before the trend reversal) and 1 (after the trend reversal); and $T_{t2015q1}x_{t2015q1}$ is the usual interaction term. Accordingly, β_0 is the intercept and represents the starting level of the outcome variable at T = 1999q1 for UR and T = 2004q1 for NEET, respectively; β_1 is the slope and represents the trajectory (or secular trend) of the outcome variable until the 2007q4 global financial crisis; β_2 is the level change that occurs immediately following the 2007q4 global financial crisis (compared to the counterfactual); β_3 is the difference between the slope pre and post the global financial crisis; β_4 is the level change that occurs immediately following the 2011q4 European sovereign debt crisis; β_5 is the difference between the slope pre and post the European sovereign debt crisis; β_6 is the level change that occurs immediately following the 2014q4 (compared to the counterfactual); β_6 is the difference between the slope pre and post the trend reversal; and ε_t represents the random error term which is assumed to follow a first auto-regressive (AR1) process. The regression coefficients are estimated by using Ordinary least-squares (OLS) method with the Newey-West (1987) standard errors.

4 Results

Four periods of linear trend were considered to analyse UR and NEET, with interruptions at 2007q4, 2011q4 and 2015q1, respectively. Separate segmented regression models were then estimated for age groups, gender and macro-regions via ordinary least-squares by using Newy-West standard errors in order to handle one lag autocorrelation. To account for the correct autocorrelation structure, Cumby-Huizinga test (Cumby and Huizinga, 1992) was performed and results confirm that autocorrelation was present at lag 1, but not at higher orders (up to the 9 lags were tested). Results are shown in Table 1 for the UR and in Table 2 for the NEET. As regards the UR, the 1999 base rate showed some variability in the considered subgroups. In fact, starting from 10.754 at national level, its value was particularly higher for the age group 15–24 (26.726) and for the South macro-regions (20.211) but lower for the North East (4.538) and North West (5.831) macro-regions, as well as for the males (8.177) and for the 45–54 age group. Moreover, the trend prior to the 2008 global financial crisis (1999q1–2007q3) showed a significant and general decrease, both at national level (-0.137; $p < 0.001$) and for the different age groups, macro regions and gender. Such reduction was more pronounced for the sub-groups traditionally considered as the most vulnerable ones, namely South macro-regions (-0.266; $p < 0.001$), females (-0.202; $p < 0.001$) and YUR (-0.182; $p < 0.001$). The onset of the global financial crisis (2007q4) caused an immediate and substantial UR increase at national level (+0.788; $p < 0.05$) and in almost all the considered subgroups but no significant change was detected for younger people (age groups 15–24 and 25–34) and the North-East macro region. In particular, the more severe direct consequences were observed among females (+1.061; $p < 0.001$), for people in the central (+1.047; $p < 0.001$) and southern regions (+0.978; $p < 0.05$) and for the intermediate age group 35–44 (+0.997; $p < 0.001$). The aftermath of the financial crisis was quite strong and resulted in the Great recession in the subsequent years during which a substantial and significant trend change was observed compared to the previous period (+0.253 $p < 0.001$). However, in this case, the most serious consequences occurred particularly for YUR (+0.717; $p < 0.001$) and, to a much lesser extent, for the South macro-regions (+0.365; $p < 0.001$). On the other hand, the immediate consequences of the second financial crisis, following the European sovereign debt crisis (2011q4) were even stronger when compared to the previous financial crisis and resulted in a second economic recession, with an UR increase almost double at the national level (+1.540; $p < 0.001$). Such increase was higher for YUR (+3.676; $p < 0.05$) and for the South macro region (+2.611; $p < 0.001$) while there was no significant increase again for North East macro region.

After this second financial shock, the UR seems to further accelerate its increase only in some sub-groups while at national level no significant trend difference was observed. In particular, such acceleration was particularly higher for the South macro-regions (+0.355; $p < 0.001$), age group 25–34 (+0.246; $p < 0.05$) and females (+0.187; $p < 0.001$) while no significant further rate increase was detected for YUR. However, it should be emphasized here that this further increase, although lower than the one highlighted during the Great Recession, where present has to be added to that already existing, thus making particularly critical the situation.

Table 1: Estimates of the impact of the 2007q4 and 2011q4 financial crises on the UR in Italy and trend reversal after 2015q1

…	***Base rate (1999)***	***Trend 1999q1-2007q3***	***Rate change 2007q4***	***Trend change 2007q4***	***Rate change 2011q4***	***Trend change 2011q4***	***Rate change 2015q1***	***Trend change 2015q1***
Overall	10.754***	-0.137***	0.788**	0.253***	1.540***	0.136**	-1.234**	-0.357***
Males	8.177***	-0.095***	0.602*	0.250***	1.426***	0.101	-1.018	-0.396***
Females	14.655***	-0.202***	1.061***	0.261***	1.675***	0.187***	-1.515***	-0.304***
15–24	26.726***	-0.182***	0.955	0.717***	3.676**	0.350	-3.371	-1.607***
25–34	11.074***	-0.068***	0.127	0.290***	1.566**	0.246**	-1.811*	-0.600***
35–44	7.771***	-0.095***	0.997***	0.177***	1.245***	0.166***	-1.225***	-0.285***
45–54	6.401***	-0.099***	0.725***	0.196***	1.111***	0.109**	-0.869*	-0.242***
55–64	7.901***	-0.166***	0.847***	0.219***	1.350***	-0.005	-0.143	-0.031
North West	5.831***	-0.067***	0.800**	0.221***	0.892**	0.008	-0.776	-0.320***
North East	4.538***	-0.037***	0.220	0.166***	0.855	0.007	-0.673	-0.269***
Center	8.540***	-0.097***	1.047***	0.188***	1.364***	0.134**	-1.021	-0.329***
South	20.211***	-0.266***	0.978**	0.365***	2.611***	0.355***	-2.290***	-0.501***

Table 2: Estimates of the impact of the 2007q4 and 2011q4 financial crises on the percentage of NEET in Italy and trend reversal after 2015q1

…	***Base rate (2004)***	***Trend 2004q1-2007q3***	***Rate change 2007q4***	***Trend change 2007q4***	***Rate change 2011q4***	***Trend change 2011q4***	***Rate change 2015q1***	***Trend change 2015q1***
Overall	19.866***	-0.062	-0.250	0.316***	0.103	0.094	-1.510**	-0.529***
Males	15.152***	0.014	-0.398	0.334***	-0.037	0.099	-1.787**	-0.680***
Females	24.584***	-0.125*	-0.221	0.285***	0.269	0.084	-1.135	-0.369***
North West	12.575***	-0.079	0.664	0.382***	-1.242	0.103	-1.632	-0.572***
North East	10.522***	-0.009	-0.558	0.364***	-0.268	-0.069	-1.169	-0.545***
Center	15.406***	-0.090	-0.995	0.404***	0.989	-0.001	-1.053	-0.561***
South	29.798***	-0.072	0.295	0.258***	0.564	0.204***	-1.801***	-0.482***

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$

The consequences of this second financial crisis seem to run out by the end of 2014. Indeed, a trend reversal occurs staring from 2015q1. In particular, the rate change decreases significantly only overall (-1.234; p < 0.05), for females (-1.515; p < 0.001), age group 35–44 (-1.225; p < 0.001) and the South macro-region (-2.290; p < 0.001). On the other hand, the trend change is more evident and is significant for all categories except for age class 55–64 and shows a stronger reduction for young generation 25–34 (-1.607; p < 0.001) and, to a lesser extent, for the South macro-region (-0.501; p < 0.001).

As regards the percentage of NEET, a considerable heterogeneity was found in the 2004 base rate, which was 19.866 at national level. Its value was higher for the South macro-regions (29.798) but lower for the North East (10.522) and North West (12.575) macro-regions; moreover, it was higher for females (24.584) than males (15.152). On the other hand, its trend prior to the 2008 global financial crisis (2004q1–2007q3) was basically constant at national level. The onset of the of the global financial crisis (2007q4) did not cause an immediate impact on the percentage of NEET, overall and in any of the considered sub-groups. However, a significant trend change was found both at national level (+0.316; p < 0.001) and for all the other sub-groups; such change was particularly higher only for the macro-regions of Center (+0.404; p < 0.001). The European sovereign debt crisis (2011q4) does not seem to alter this situation, neither for the rate change nor for the trend change. This means that after this second financial crisis the rise of the percentage of NEET remains steady and equal to the previous period without showing any jump. But for the last trend change (2015q1), results reveal to be significant for all sub groups, particularly for females (-0.680; p < 0.001). This confirms that after the end of the two crises and local interventions, a new period of recovery has finally started again. Also for the percentage of NEET, a trend reversal occurs staring from 2015q1. In particular, analogously to the UR, this reduction is significant for all the considered sub-groups, but it is more pronounced for males (-0.680; p < 0.001) and lower for females (-0.369; p < 0.001) and South marco-region (-0.482; p < 0.001).

References

1. Cumby, R. E., and Huizinga, J.: Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. Econometrica, 60, 185—195 (1992).
2. Linden, A., and Adams, J. L.: Applying a propensity-score based weighting model to interrupted time series data: Improving causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, 17, 1231–1238 (2011).
3. Lopez Bernal, J., Cummins, S., and Gasparrini, A.: Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int. J. Epidemiol.*, 1–8 (2016).
4. Newey, W. K., and West, K. D.: A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703–708 (1987).
5. Shadish, W. R., Cook, T. D., and Campbell, D. T.: Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin (2002).
6. Wagner, A. K., Soumerai, S. B., Zhang, F., and Ross-Degnan, D.: Segmented regression analysis of interrupted time series studies in medication use research. J. *Clin. Pharm. Ther.*, 27(4), 299–309 (2002).
7. Eurostat: *Glossary*. Luxembourg: Eurostat (2017).
8. ILO: *World of work report 2011: Making markets work for jobs*. Geneva: International Labour Office (2011).

Quantile and Generalized Quantile Methods

Multiple quantile regression for risk assessment

Regressione quantile multipla per la valutazione del rischio

Lea Petrella and Valentina Raponi

Abstract This paper proposes a maximum-likelihood approach to jointly estimate conditional marginal quantiles of multivariate response variables using a linear regression framework.We consider a slight reparameterization of the Multivariate Asymmetric Laplace distribution proposed by Kotz et al (2001) and exploit its location-scale mixture representation to implement a new EM algorithm to estimate model parameters. The idea is to extend the link between the Asymmetric Laplace distribution and the well-known univariate quantile regression model to a multivariate context. The approach accounts for association among multiple response variables and study how such association structure and the relationship between responses and explanatory variables can vary across different quantile levels of the conditional distribution of the responses. A penalized version of the EM algorithm is also presented to tackle the problem of variable selection. The validity of our approach is analyzed in a simulation study, where we also provide evidence on the efficiency gain of the proposed method compared to estimation obtained by separate univariate quantile regressions. A real data application is finally proposed to study the main determinants of financial distress in a sample of Italian firms.

Abstract *Questo lavoro propone un approccio di massima verosimiglianza per stimare congiuntamente i quantili marginali condizionati associati a variabili risposta multivariate, in un contesto di modelli di regressione lineare multivariata. Viene considerata una leggera riparametrizzazione della distribuzione Asimmetrica di Laplace Multivariata proposta da Kotz et al. (2001) e viene proposto un nuovo algoritmo EM per la stima dei parametri del modello, sfruttando la particolare rappresentazione a mistura tipica della distribuzione Asimmetrica di Laplace. L'idea é quella di estendere il link esistente tra la distribuzione di Laplace e il modello*

Lea Petrella
MEMOTEF Department, Sapienza University of Rome, Rome, Italy e-mail: lea.petrella@uniroma.it

Valentina Raponi
MEMOTEF Department, Sapienza University of Rome, Rome, Italy e-mail: valentina.raponi@uniroma1.it

di regressione quantile univariato ad un contesto multivariato. L'approccio proposto consente di tenere conto dell'eventuale associazione esistente tra le variabili risposta e intende studiare come tale struttura di associazione vari quando si considerano diversi quantili della distribuzione condizionata della variabile risposta. Viene inoltre proposto un algoritmo EM penalizzato (PEM) per affrontare il problema di selezione delle variabili. La validita' del nostro approccio viene analizzata attraverso uno studio di simulazione, attraverso il quale viene anche mostrato il guadagno in termini di efficienza del modello proposto rispetto alla stima ottenuta da (separate) regressioni quantili univariate. Infine, viene proposta un'applicazione empirica con l'obiettivo di studiare le determinnti principali del distress finanziario in un campione di aziende italiane.

Key words: Multiple quantiles, Quantile Regression, Multivariate Asymmetric Laplace Distribution, EM algorithm, Maximum Likelihood, Multivariate response variables.

References

1. Arslan, O., (2010). An alternative multivariate skew Laplace distribution: properties and estimation, *Statistical Papers* , Vol. 51, pp. 865–887.
2. Bastos, R., Pindado, J., (2013). Trade credit during a financial crisis: A panel data analysis. *Journal of Business Research*, Vol 6, pp. 614–620.
3. Cho, H., Kim, S., Kim, M., (2017). Multiple quantile regression analysis of longitudinal data: Heteroscedasticity and efficient estimation. *Journal of Multivariate Analysis*, Vol. 155, pp. 334–343.
4. Geraci, M., Bottai, M., (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, Vol. 8, pp. 140–154.
5. Koenker, R. (2017): Quantile Regression: 40 Years On. *Annual Review of Economics*. Vol. 9, pp 155–176.
6. Koenker, R., Bassett, G. (1978) Regression quantiles. *Econometrica*. Vol. 46, pp. 33–50.
7. Kotz, S., Kozubowski, T. J., Podgorski, K. (2001). The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance. Boston: Birkhauser.
8. Kozumi, H, Kobayashi, G., (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, Vol. 81, pp. 1565–1578.
9. Pindado, J., Rodrigues L., De la Torre, C., (2008). Estimating financial distress likelihood. *Journal of Business Research*, Vol. 61, pp. 995–1003.
10. Yu, K. and Moyeed, R.A., (2001), Bayesian quantile regression. *Statistics and Probability Letters*, Vol. 54, pp. 437–447.

Parametric Modeling of Quantile Regression Coefficient Functions

Modelli Parametrici per i Coefficienti di una Regressione Quantilica

Paolo Frumento and Matteo Bottai

Abstract In standard quantile regression (QR), quantiles are estimated one at the time. An alternative approach, which is referred to as *quantile regression coefficients modeling* (QRCM), is to describe the functional form of the regression coefficients parametrically. This approach facilitates estimation and inference, simplifies the interpretation of the results, and generates more efficient estimators. Moreover, thanks to the imposed parametric structure, it makes it easier to estimate quantiles in situations involving latent variables, missing or partially observed data, and other complications arising in survival analysis, longitudinal data analysis, and causal inference, where applying standard QR proves difficult and computationally inefficient. We describe the method, discuss applications, and illustrate the R package `qrcm`.

Abstract *I metodi standard di regressione quantilica (*QR*) permettono di stimare un quantile alla volta. Un approccio alternativo, noto come* ***quantile regression coefficients modeling*** *(*QRCM*) è quello di descrivere i coefficienti di regressione con un modello parametrico. Questo metodo semplifica la stima e l'inferenza, migliora l'interpretazione dei risultati, e aumenta l'efficienza degli stimatori. La struttura parametrica permette di stimare i quantili in presenza di variabili latenti, dati mancanti o incompleti, e altre complicazioni tipiche in analisi di sopravvivenza, in presenza di dati longitudinali, e in inferenza causale, dove le tecniche standard sono difficilmente applicabili e computazionalmente inefficienti. In questo lavoro, descriviamo il metodo e le sue applicazioni, e illustriamo il pacchetto R* `qrcm`*.*

Key words: Quantile modeling, censored and truncated quantile regression, longitudinal quantile regression, R package `qrcm`

Paolo Frumento
Karolinska Institute, Institute of Environmental Medicine, Unit of Biostatistics, 17177 Stockholm, Sweden, e-mail: paolo.frumento@ki.se

Matteo Bottai
Karolinska Institute, Institute of Environmental Medicine, Unit of Biostatistics, 17177 Stockholm, Sweden, e-mail: matteo.bottai@ki.se

References

1. Frumento P, Bottai M (2016). Parametric modeling of quantile regression coefficient functions. Biometrics, 72 (1), 74-84, doi: 10.1111/biom.12410.
2. Frumento P, Bottai M (2017). Parametric modeling of quantile regression coefficient functions with censored and truncated data. Biometrics,, 73(4), 1179-1188, doi: 10.1111/biom.12675.
3. Frumento P (2017). qrcm: Quantile Regression Coefficients Modeling. R package version 2.1. https://cran.r-project.org/package=qrcm

Modelling the effect of Traffic and Meteorology on Air Pollution with Finite Mixtures of M-quantile Regression Models

Uno studio dell'effetto del traffico e delle variabili meteo sulla concentrazione di particolato atmosferico attraverso l'uso di modelli M-quantile a mistura finita

Simone Del Sarto, Maria Francesca Marino, Maria Giovanna Ranalli and Nicola Salvati

Abstract Between 2012 and 2015, the PMetro project collected meteorological, aerosol, and gas measurements by using dedicated instruments integrated on one of the cabins of the Minimetro, a public conveyance of the town of Perugia. In this work, the effect of vehicular traffic and meteorological variables on the distribution of particulate matter (PM) is analyzed using a Finite Mixture of M-Quantile regression models. Results show that radon concentration and vehicular traffic have the largest effect on the distribution of PM. In particular, the parameter estimate for radon concentration is always positive and increases along the M-quantiles, while the effect of vehicular traffic is similar both for lower and higher PM concentrations.
Abstract *Fra il 2012 ed il 2015, il progetto PMetro ha raccolto dati meteo e sulla concentrazione di particolato atmosferico e gas usando strumentazione appositamente installata su una delle cabine del Minimetro, una linea di trasporto urbana su rotaia della città di Perugia. In questo lavoro usiamo modelli M-quantile a mistura finita per studiare l'effetto del traffico e delle variabili meteo sulla distribuzione del particolato atmosferico. I risultati principali mostrano che la concentrazione di radon ed il traffico sono le variabili che incidono maggiormente, con andamenti, tuttavia, diversi per diversi valori degli M-quantili.*

Simone Del Sarto
Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione; e-mail: `simone.delsarto@email.com`

Maria Francesca Marino
Dipartimento di Statistica, Informatica, Applicazioni, Università degli studi Firenze; e-mail: `marino@disia.unifi.it`

Maria Giovanna Ranalli
Dipartimento di Scienze Politiche, Università degli studi di Perugia; e-mail: `giovanna.ranalli@unipg.it`

Nicola Salvati
Dipartimento di Economia e Management, Università di Pisa; e-mail: `nicola.salvati@unipi.it`

Key words: Particulate Matter, PMetro, Spatio-temporal data, Random Parameters

1 Introduction

Air quality is determined by the level of the so called "atmospheric aerosol", including concentration of natural and anthropogenic pollutants, such as gases and particulate matter (PM). Atmospheric aerosol represents a critical component of the atmosphere that impacts not only on regional air pollution and human health, but also on global climate [10]. Aerosol formation in the atmosphere involves several and complex processes and particle abundance at a given time depends also on atmospheric dispersion, chemical transformation and particle loss processes. As a consequence, the concentration of PM in the atmosphere is highly variable in time and space, then the choice of the location and the time resolution of the air monitoring stations is crucial and regulated by law (e.g., Directive 2008/50/EC).

PMetro (`www.pmetro.it`) is a project that ran between 2012 and 2015 and whose main purpose was to create an innovative system to monitor urban air-quality in real-time in Perugia (Central Italy). PMetro used highly resolved space-time meteorological, aerosol and gas measurements from instruments integrated on one of the cabins of the Minimetro, a public conveyance of the town. In this paper, we aim at assessing the effect of vehicular traffic on the distribution of PM in the town of Perugia using PMetro data, while controlling for other climatic and environmental factors which may have an impact on the PM concentration. To this end, we consider the Finite Mixture of M-Quantile regression models recently introduced in the literature [3].

The proposed approach allows us to pursue a two-fold aim. First, it allows us to "move beyond mean regression" [8] which can provide a rather incomplete picture of the phenomenon under investigation. Second, it allows us to account for heterogeneity in the data, by means of random parameters having unspecified distribution. This is directly estimated from the observed data and, as a result, allows us to avoid unverifiable assumptions on the random parameter distribution.

The paper is organized as follows. Section 2 provides a close insight into the data and the modeling issues. Section 3 illustrates the finite mixture of M-quantile regression models, while Section 4 shows the results of the application to the PMetro data. Finally, Section 5 summarizes our findings.

2 Data

Aerosol data, in particular size-fraction particle number counts, were collected by an Optical Particle Counter (OPC) placed on one of the cabins of the Minimetro. The OPC provided a measure of particle number concentration in the air (i.e., the count of particles) sampled every six seconds through an aspiration line sited on the roof of the cabin. Furthermore, the cabin provided a signal which allowed its exact positioning along the path [5].

The concentration measure collected by the OPC was divided into 22 dimensional bins according to the particle diameter, expressed in micrometers (μm, 10^{-6} meters), from 0.28 μm to 10 μm. In this paper, we focus on particles with a diameter lower than 1.10 μm, obtained summing the first nine dimensional bins. This kind of particles are often called "fine" particles, even if the classification in fine and coarse is not universal and can be different according to the location (urban/rural) and/or the spatial position of the monitoring stations [11].

OPC data are integrated with two other datasets providing information on traffic volume and meteorological conditions. These three types of data have different temporal resolutions, as they were collected with different time frequencies (six seconds for OPC, five minutes for traffic, and one hour for meteorological conditions). Moreover, the OPC was a mobile station, since it collected data in the transect of the town covered by the Minimetro path, while traffic and meteorological data were obtained from fixed stations, located along the metro path. To exploit the information included in all these datasets, only OPC data collected at a particular crossroad, where traffic and meteorological conditions are available as well, are considered.

In particular, we focus on data collected from March, 2014 to February, 2015: in this period, considering the days in which all the data sources are simultaneously available, we have data for 201 days. Since Minimetro operated between 6.30 and 21.30, only the data collected within this time period are considered. Furthermore, with the aim of tackling different time resolutions, we summarize OPC, traffic, and meteorological data every half-hour. Considering that these latter are collected every hour, we computed the average between two consecutive measurements to obtain half-hour information. Last, as it is typically done when dealing with particle concentration [11], we log-transformed and normalized such a variable as follows

$$y = \log\left[\frac{count+1}{\log 1.10 - \log 0.28}\right].$$

As far as potential explanatory variables are concerned, we consider the following meteorological covariates: radon concentration, temperature, wind speed, total solar radiation and rainfall. Vehicular traffic is also considered; this is measured in terms of number of vehicles passed through the crossroad at each half-hour. Furthermore, the cumulative count of vehicles passed in the previous hours (with respect to the current time point) is also taken into account, with the aim of investigating the best proxy of vehicular traffic. More details on this issue are provided in Section 4. Here, we focus on complete data only, for a total of 5,437 observations referred to 201 days. In the following section we describe the model used.

3 Finite Mixture of M-quantile regression models

Let Y_{ij} be a continuous response variable for unit $j = 1, \dots, n_i$, belonging to cluster $i = 1, \dots, m$, and let $\mathbf{x}_{ij}$ denote a p-dimensional vector of observed covariates. We aim at analyzing the effect of these covariates on the joint distribution of the observed responses y_{ij}. To ensure robustness with respect to possible outliers in the data and to obtain a deeper description of the response variable distribution, we focus on a M-quantile specification of the regression model. Also, as we are dealing with clustered observations, dependence between measures recorded from units in same cluster represents a crucial aspect to be taken into account for deriving unbiased estimates. Typically, such a dependence is modelled by including in the response data model sources of unobserved heterogeneity in the form of cluster-specific random parameters. In this respect, for a given M-quantile $q \in (0,1)$, let $\mathbf{b}_{i,q}$ denote an r-dimensional vector of random parameters associated to the i-th cluster, distributed according to the density $f_{b,q}(\cdot)$; frequently, a Gaussian distribution is considered for such random parameters.

Here, we follow a different path and consider the semi-parametric approach suggested by [3], according to which the mixing distribution $f_{b,q}(\cdot)$ is left unspecified and is directly estimated from the data. In particular, a nonparametric maximum likelihood approach is considered to estimate $f_{b,q}(\cdot)$; this is known to lead to a discrete mixing distribution putting masses $\pi_{k,q} > 0$ on locations $\zeta_{k,q}, k = 1, \dots, K$, with $\sum_{k=1}^{K} \pi_{k,q} = 1$, see also [1, 2].

As it is commonly done when dealing with mixed effect models, we assume that, conditional on the (discrete) random parameters $\mathbf{b}_{i,q}$, responses Y_{ij} are independent each other and, for a given M-quantile $q \in (0,1)$, follow a Generalised Asymmetric Least Informative (GALI) distribution [4, 3] :

$$Y_{ij} \mid \mathbf{b}_{i,q} \sim \frac{1}{B_q(\sigma_q)} \exp\{-\rho_q[y_{ij} - MQ_q(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{b}_{i,q}; \psi)]\}. \tag{1}$$

In the equation above, $\rho_q(\cdot)$ denotes the Huber loss function, $B_q(\cdot)$ is a normalising constant that ensures the density integrates to one, and σ_q is a M-quantile-specific scale parameter. Last, $MQ_q(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{b}_{i,q}; \psi)$ indicates the q-th M-quantile of the conditional response distribution. For $\mathbf{b}_{i,q} = \zeta_{k,q}$, this is modeled according to the following expression:

$$MQ_q(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{b}_{i,q}; \psi) = MQ_q(y_{ij} \mid \mathbf{x}_{ij}, \zeta_{k,q}; \psi) = \mathbf{x}'_{ij}\beta_q + \mathbf{w}'_{ij}\zeta_{k,q},$$

where ψ is the influence function corresponding to $\rho_q(\cdot)$, β_q is a p-dimensional vector of fixed effects, and $\mathbf{w}_{ij}$ is a subset of $\mathbf{x}_{ij}$ containing the covariates whose effect is assumed to vary with clusters.

Based on the modeling assumptions introduced so far, the joint *conditional* density of all responses coming from the i-th cluster, $\mathbf{y}_i$, is given by:

$$f_q\left(\mathbf{y}_i \mid \zeta_{k,q}; \beta_q, \sigma_q, \mathbf{X}_i\right) = \prod_{j=1}^{n_i} f_q\left(y_{ij} \mid \zeta_{k,q}; \beta_q, \sigma_q, \mathbf{x}_{ij}\right),$$

with $f_q(y_{ij} \mid \zeta_{k,q}; \beta_q, \sigma_q, \mathbf{x}_{ij})$ being the GALI distribution in equation (1) with $\mathbf{b}_{i,q} = \zeta_{k,q}$, and $\mathbf{X}_i$ being the matrix of all observed covariates for cluster i .

Denoting by Φ_q the vector of all model parameters associated to the q-th M-quantile, this can be efficiently estimated from the following observed data log-likelihood function:

$$\ell_q(\Phi_q) = \sum_{i=1}^{m} \log\left\{\sum_{k=1}^{K} f_q\left(\mathbf{y}_i \mid \zeta_{k,q}; \beta_q, \sigma_q, \mathbf{X}_i\right) \pi_{k,q}\right\}. \tag{2}$$

As it may be noticed, the expression in equation (2) resembles the likelihood of a finite mixture model defined on K distinct components; the maximization of such a function can be directly pursued by exploiting an extended version of the standard EM algorithm [6]. To conclude this section, it is worth to highlight that the assumption of GALI distributed responses is merely ancillary and it is simply introduced to cast standard estimation of M-quantile regression models into a maximum likelihood context.

4 Results

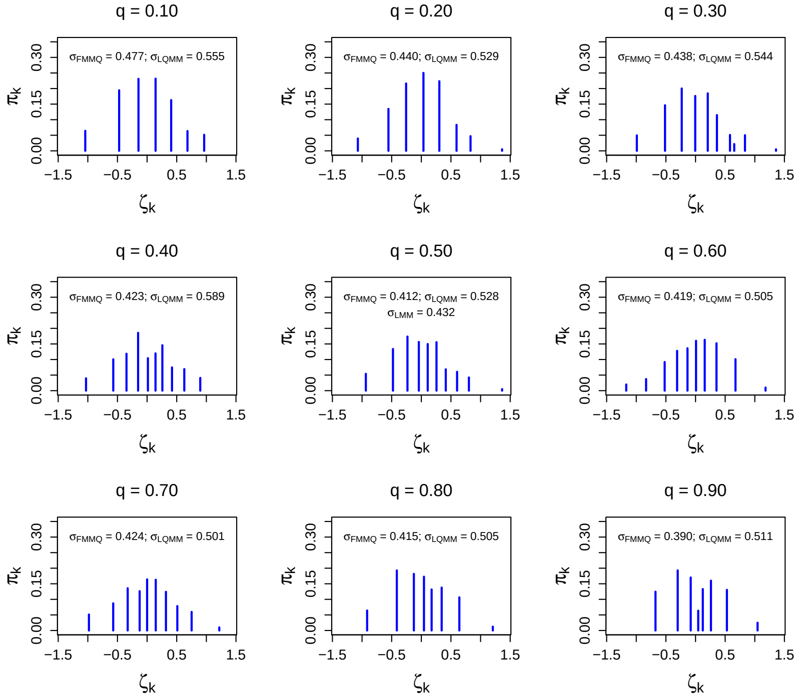
In this section, we exploit the model described so far to analyze PMetro data introduced in Section 2. Our aim is that of characterizing the evolution over time of air quality in Perugia in terms of a number of covariates summarizing traffic level and meteorology conditions.

A preliminary analysis was conducted with the aim of selecting the covariate that best represents the vehicular traffic in town. Such an analysis led us to retain as optimal proxy of vehicular traffic the cumulative sum of vehicles passed in the previous 4 hours. Other covariates included into the model are: the logarithmic transform of radon concentration (log-*Radon*), the temperature (*Temp*) and its quadratic transform (*Temp*^{2}), the wind speed (*WSP*), the total solar radiation on the logarithmic scale (log-*TSR*), and the rainfall (*Rain*). To account for sources of unobserved heterogeneity among days which are non captured by the available covariates, we also included in the model a discrete random intercept having unspecified distribution.

To analyse the possible differential effect of observed covariates on different parts of the (conditional) response variable distribution, we considered a grid of M-quantile levels ranging from $q = 0.1$ to $q = 0.9$. For each q, we run the estimation algorithm for a varying number of mixture components and retained the optimal solution according to the BIC index. In this respect, we selected $K = 10$ for $q = 0.3, 0.4, 0.5, 0.6, 0.7$, $K = 8$ for $q = 0.2, 0.8, 0.9$, and $K = 7$ for $q = 0.1$. That is, the optimal K is higher when focusing on the center of the conditional response distribution and reduces at the tails. Here, a reduced amount of information is available and, therefore, a reduced amount of (unexplained) variability is present.

We represent in Figure 1 the estimated distribution of the random parameters for varying q; at the top of this figure, we also report the standard deviation of such a distribution – σ_{FMMQ}. For comparison, we also report the standard deviation of the random parameters obtained under a standard linear quantile mixed model (LQMM) specification [7] and, for for $q = 0.5$, the estimated standard deviation obtained from a linear mixed model (LMM) [9]. By looking at these values, we observe that σ_{LQMM} is always lower than σ_{FMMQ}. This may be due to the slight (positive) asymmetry and/or bi-modality of the random parameter distribution, especially for some q values. While this feature is easily captured by the FMMQ approach, the same is not true when considering the a parametric specification for $b_{i,q}$.

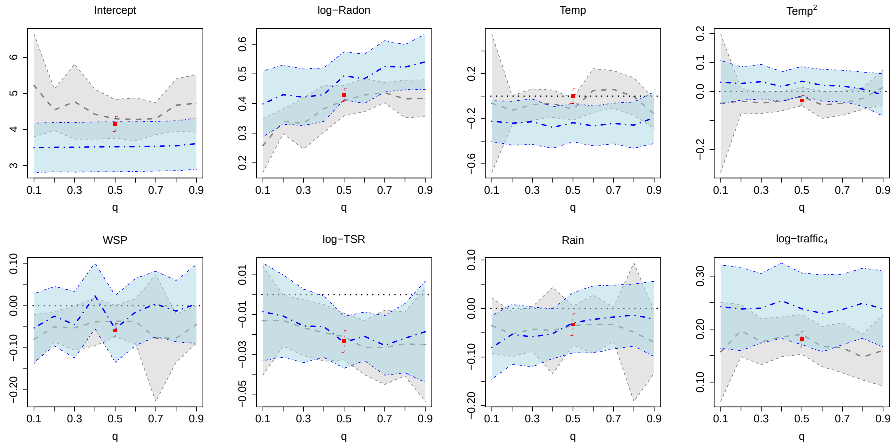
Fig. 1 Estimated random coefficient distribution for varying q. In each plot, the estimated standard deviation of the random parameter distribution is reported.



Finally, to understand whether the effect of the covariates differs across M-quantiles, we report in Figure 2 the evolution of fixed parameter estimates for different q, together with the corresponding 95% confidence intervals. For comparison, we also report for each M-quantile q the estimates obtained under the LQMM and LMM specifications. For all the covariates, confidence intervals obtained under LQMM and FMMQ always overlap suggesting that the two methodologies lead to similar results. In particular, the radon concentration estimate is always positive and significant, and, in particular, increases with q. This implies that the effect of radon on fine particle count is stronger when concentrations are higher. Total solar radiation has a similar behavior in the opposite direction: it has a negative effect on fine particle concentration, and this effect is stronger when concentrations are higher.

We would have expected an increasing pattern for vehicular traffic. However, results reported in Figure 2 highlight that this estimate remains quite constant with q. This implies that the relative impact of an increase in vehicular traffic remains the same for lower and for higher concentrations of fine particle concentrations. This pattern is similar to that from LQMM.

Fig. 2 Fixed parameter estimates and 95% confidence intervals under FMMQ (grey), LQMM (blue) and LMM (red), for varying q.



5 Conclusions

In this paper, we analyzed data on air quality to assess the effect of vehicular traffic. Fine particle concentration depends on a complex process and we included meteorological variables in our model to try to isolate the effect of traffic. We used a finite mixture of M-quantile regression models to investigate the effect of the covariates at different locations of the response distribution, to account for the clustered structure of the data without resorting to a parametric (usually Gaussian) distribution for the random parameters, and to deal with outlying observations. We used data for a whole year in order to account for seasonal trends.

Since vehicular traffic is the only variable that policy makers can control to reduce fine particle concentration, we can use the findings in this paper to provide guidelines. A common act enforced to reduce pollution in Italy is to restrict circulation in designated areas and days, alternating vehicles with an odd/even plate number. This implies roughly halving the number of vehicles. From the findings in our paper, we can state that reducing by 50% the number of vehicles implies a reduction of 10% of fine particle concentration, other things being equal. This relative reduction is essentially the same in spite of the fact that we are in day with a relatively low or high fine particle concentration. Of course, this also implies that the effect will be higher in absolute terms.

References

1. Aitkin, M.: A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and Computing **6**, 251–262 (1996)
2. Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. Biometrics **55**, 117–128 (1999)
3. Alfò, M., Salvati, N., Ranalli, M.G.: Finite mixtures of quantile and M-quantile regression models. Statistics and Computing **27**(2), 547–570 (2017)
4. Bianchi, A., Fabrizi, E., Salvati, N., Tzavidis, N.: M-quantile regression: diagnostics and parametric representation of the model. Working paper (2015). Available at `http://www.sp.unipg.it/surwey/dowload/publications/24-mq-diagn.html`
5. Castellini, S., Moroni, B., Cappelletti, D.: PMetro: Measurement of urban aerosols on a mobile platform. Measurement **49**, 99–106 (2014)
6. Dempster, A., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**, 1–38 (1977)
7. Geraci, M., Bottai, M.: Linear quantile mixed models. Statistics and Computing **24**, 461–479 (2014)
8. Kneib, T.: Beyond mean regression. Statistical Modelling **13**(4), 275–303 (2013)
9. Laird, N.M., Ware, J.H.: Random-effects models for longitudinal data. Biometrics **38**(4), 963–974 (1982)
10. Pöschl, U.: Atmospheric aerosols: composition, transformation, climate and health effects. Angewandte Chemie International Edition **44**(46), 7520–7540 (2005)
11. Seinfeld, J.H., Pandis, S.N.: Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons (2012)

Recent Advances on Extreme Value Theory

Extremes of high-order IGARCH processes

Valori estremi per processi IGARCH di ordine elevato

Fabrizio Laurini

Abstract The extremal properties of GARCH processes are of wide interest for market risk management. Only for simple GARCH(1,1) extremes have been widely discussed. Much remains to be found about the dependence structure of extreme values for higher order GARCH. Although recent research has identified the multivariate regular variation property of stationary GARCH(p,q) processes, currently there are no methods for numerically evaluating extreme components, like the average length of an extreme period. Only very simple special cases are well understood, but these are of little practical relevance, as bounded distribution of the error term is assumed. We present a unified toolkit that tackles the above critics and it is usable for Integrated GARCH(p,q) processes, assuming innovations with unbounded support or asymmetry. With our method we are able to generate the forward tail chain of the process to derive all extremal features. The convergence of our numerical algorithm is very fast due to an efficient implementation of a particle filtering simulation technique.

Abstract *Le proprietà estreme dei processi GARCH sono di interesse centrale per la gestione del rischio di mercato. Soltanto per semplici GARCH(1,1) i valori estremi sono stati caratterizzati. Molto resta da studiare riguardo la dipendenza dei valori estremi per processi GARCH di ordine più elevato. Sebbene recentemente si siano usate connessioni con le proprietà di variazione regolare multivariata dei processi GARCH(p,q), al momento non ci sono metodi in grado di quantificare tali caratteristiche estreme. Pertanto, soltanto casi speciali sono stati compresi a fondo, ma questi sono spesso irrilevanti da un punto di vista pratico, visto che viene assunta una forma sul termine d'errore con distribuzione limitata. Si presenta un insieme di tecniche unificato volto a superare tutti questi inconvenienti ed è usabile anche per processi GARCH(p,q) Integrati con innovazioni a supporto illimitato e asimmetriche. Il metodo si basa sulla cosiddetta forward tail chain per derivare tutti gli*

F. Laurini
Department of Economics and Management, University of Parma, Via J. F. Kennedy, 6, Parma, e-mail: fabrizio.laurini@unipr.it

aspetti rilevanti nel processo dei valori estremi. La convergenza dell'algoritmo è veloce grazie all'utilizzo di un'implementazione efficiente di particle filtering.

Key words: Extremes, IGARCH, Particle filtering, Regular variation

1 Introduction

The risk management in the stock markets, commonly called *market risk management*, suggests the use of statistical tools and models which aim at reducing the potential size of losses, occurring by sudden drops or growth in the stock market. Such losses are even amplified when the volatility of the stock market is substantial. Hence, modeling and forecasting the temporal evolution of the market volatility is of great concerns for financial institutions.

Consider the daily log-returns $X_t = \log P_t - \log P_{t-1}$, $(X_t \in \mathbb{R})$ where P_t, $t = 1, 2, \ldots$, is the price of a generic asset. Then a broad class of models, mostly adopted to describe the market volatility, is the generalized autoregressive conditionally heteroscedastic (GARCH) introduced by [3]. For market risk management one of the most important issue is the presence of extreme values of daily log-returns. Therefore, understanding the extreme properties for such processes is fundamental, and this can be achieved by considering the marginal and the clustering properties of GARCH processes.

GARCH(p,q) models, for integers p and q, have the form

$$X_t = \sigma_t Z_t \quad \text{with} \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \quad t = 1, 2, \ldots, \quad \text{and} \quad \alpha_0 > 0. \tag{1}$$

For fixed t Z_t and σ_t are independent. The independent and identically distributed (IID) sequence $\{Z_t\}_{t \geq 1}$ is assumed to be symmetric with $E(Z_t^2) = 1$. Conditions on the parameters $\alpha_i, i = 1, \ldots, q$ and $\beta_j, j = 1, \ldots, p$ are discussed in Section 2.

GARCH(p,q) processes $\{X_t\}_{t \geq 1}$ satisfies mixing conditions, so that the key parameter for quantifying the impact of extreme values is the extremal index $\theta_X \in (0, 1]$. The extremal index θ_X measures the level of clustering of extreme values, with the clustering of extreme increasing for θ_X decreasing.

An important interpretation of the extremal index is provided using the cluster size distribution $\pi_X(i), i = 1, 2, \ldots$, since $\sum_{i=1}^{\infty} i \pi_X(i) = (\theta_X)^{-1}$, so θ_X is the reciprocal of the limiting mean cluster size of extreme values. The special case $\theta_X = 1$ means no clustering of extremes. Extremes of GARCH(p,q) models have been studied by [1], but formulae for θ_X do not exist.

[2] were the first to propose computational algorithms for the evaluation of θ_X. Their algorithms make a very strict assumption that the innovation Z_t has bounded support, ruling out many important distributions used by practitioners, e.g., Z_t being Gaussian or t-distributed. We propose an entirely new algorithm that does not re-

quire these assumptions, and critically allows unbounded support for Z_t, as we take $Z_t \sim \mathrm{ST}(0,1,\lambda,\nu)$ scaled to have unit variance.

The relevant feature of the proposed algorithm is to simulate directly the tail chain of a GARCH(p,q) model. The tail chain of a process, introduced by [2], will be discussed later. The approach used here extends the algorithm of [4] for GARCH$(1,1)$ processes.

We derive the theory for obtaining the extremal index of a GARCH(p,q) process, and we provide a Monte Carlo algorithm for the numerical evaluation θ_X and associated cluster size distribution. Precisely, with our algorithm we first obtain the cluster size distribution $\pi_{X^2}(\cdot)$ and the extremal index θ_{X^2} for the square of the process, and then derive $\pi_X(\cdot)$ and θ_X. All results do not require the symmetry of Z.

2 Technical background, notation and assumptions

2.1 SRE representation for GARCH

Let us start by defining stationarity for GARCH(p,q) processes. We focus on the squared GARCH process, $X_t^2{}_{t\geq 1}$, and rewrite the process as a stochastic recurrence equation (SRE) as this enables the exploitation of a range of established results for such processes, e.g., the existence of results for the marginal distribution.

Let the $(p+q)$ vector $\mathbf{Y}_t$, the $(p+q)\times(p+q)$ matrix $\mathbf{A}_t$ and the $(p+q)$ vector $\mathbf{B}_t$ be

$$\mathbf{Y}_t = \begin{pmatrix} X_t^2 \\ \vdots \\ X_{t-q+1}^2 \\ \sigma_t^2 \\ \vdots \\ \sigma_{t-p+1}^2 \end{pmatrix} \qquad \mathbf{A}_t = \begin{pmatrix} \alpha^{(q-1)} Z_t^2 & \alpha_q Z_t^2 & \beta^{(p-1)} Z_t^2 & \beta_p Z_t^2 \\ I_{q-1} & 0_{q-1} & I_{q-1} & 0_{q-1} \\ \alpha^{(q-1)} & \alpha_q & \beta^{(p-1)} & \beta_p \\ 0_{q-1} & 0_{p-1} & I_{p-1} & 0_{p-1} \end{pmatrix} \qquad \mathbf{B}_t = \begin{pmatrix} \alpha_0 Z_t^2 \\ 0_{q-1} \\ \alpha_0 \\ 0_{q-1} \end{pmatrix} \tag{2}$$

where $\alpha^{(s)} = (\alpha_1,\ldots,\alpha_s) \in \mathbb{R}^{(s)}$, $\beta^s = (\beta_1,\ldots,\beta_s) \in \mathbb{R}^s$, I_s is the identity matrix of size s, 0_s is a square matrix of zeros of size s and 0_s is a column vector of zeros having length s. In each case here if $s<0$ then these terms are to be interpreted as being dimensionless. Then it follows that the squared GARCH(p,q) processes satisfies the SRE

$$\mathbf{Y}_t = \mathbf{A}_t \mathbf{Y}_{t-1} + \mathbf{B}_t, \qquad t \geq 1, \tag{3}$$

where $\{\mathbf{A}_t\}_{t\geq 1}$ and $\{\mathbf{B}_t\}_{t\geq 1}$ are each sequences of IID stochastic matrices and vectors. The formulation of the SRE via (2) is less parsimonious than that of [1], but has the benefit of covering all GARCH(p,q) processes, even when $p=q=1$.

It is necessary and sufficient that there is a negative top Lyapunov exponent of $\mathbf{A}_t$ for the existence of a unique, strictly stationary solution of SRE (3). Under the condition $E \ln^+ \|\mathbf{A}_t\| < \infty$ (here $\ln^+ x = \ln x$, if $x \geq 1$ and 0 otherwise), the top Lyapunov exponent is

$$\gamma = \lim_{t \to \infty} \frac{1}{t} \ln \|\mathbf{A}_t \mathbf{A}_{t-1} \cdots \mathbf{A}_1\| \tag{4}$$

almost surely, so that, via expression (4) a relatively simple simulation can be performed to obtain γ.

If $\sum_{j=1}^{p} \beta_j < 1$ then $\gamma < 0$. This stationary condition covers various forms of GARCH process including the IGARCH(p,q) process which has the property that

$$\sum_{i=1}^{q} \alpha_j + \sum_{j=1}^{p} \beta_j = 1. \tag{5}$$

For second-order stationarity a stronger condition is required, namely the left hand side of equation (5) is required to be strictly less than 1. This condition implies that $\gamma < 0$, and the second moment of X_t^2 are finite, and so is the fourth moment of X_t. So an IGARCH(p,q) is strictly stationary but has infinite variance and so is not second-order stationary.

2.2 Tail chain process and regular variation for squared GARCH

Taking a heavy tailed process $\{\mathbf{Y}_t\}_{t \geq 1}$ as strictly stationary, the tail chain is defined in the following way. When $u \to \infty$, if for any $t \geq 1$

$$(\mathbf{Y}_0/u, \mathbf{Y}_1/X_0, \ldots, \mathbf{Y}_t/\mathbf{Y}_0) \mid \|\mathbf{Y}_0\| > u,$$

converges weakly to $(\hat{\mathbf{Y}}_0, \hat{\mathbf{Y}}_1, \ldots \hat{\mathbf{Y}}_t)$. The tail process $\{\hat{\mathbf{Y}}_t\}_{t \geq 1}$ exists if and only if $\{\mathbf{Y}_t\}_{t \geq 1}$ is jointly regularly varying.

[1] show that there exists a unique stationary solution to the SRE (3) and this solution exhibits a multivariate regular variation property, i.e., for any $t \geq 1$, any norm $\|\cdot\|$ and all $r > 0$,

$$\frac{\Pr(\|\mathbf{Y}_t\| > rx, \mathbf{Y}_t/\|\mathbf{Y}_t\| \in \cdot)}{\Pr(\|\mathbf{Y}_t\| > x)} \xrightarrow{v} r^{-\kappa} \Pr(\mathbf{D}_t \in \cdot), \qquad \text{as } x \to \infty, \tag{6}$$

where $\xrightarrow{v}$ denotes vague convergence, $\kappa \geq 0$, and $\mathbf{D}$ is $p+q$ dimensional random vector in the unit sphere (with respect to a norm $\|\cdot\|$) defined by $\mathbb{S}^{p+q} \subset \mathbb{R}^{p+q}$. If condition (6) holds $\mathbf{Y}_t$ exhibits multivariate regularly variation with index κ and $\mathbf{D}$ is termed the spectral tail process of the vector $\mathbf{Y}_t$.

The subsequent results link γ to κ. There is structure imposed on both κ and $\mathbf{D}$ by the GARCH(p,q) process. In particular, κ is the unique positive solution of the equation

$$\lim_{t \to \infty} \frac{1}{t} \ln E \left(\|\mathbf{A}_t \mathbf{A}_{t-1} \cdots \mathbf{A}_1\|^{\kappa} \right) = 0. \tag{7}$$

For all the numerical evaluations, we will use the norm $\|\mathbf{A}\| = \sum |a_{ij}|$. In general κ can be found only by numerical solution of equation (7) via Monte Carlo. However for any IGARCH$(1,1)$ process $\kappa = 1$; see [1].

A consequence of the multivariate regular variation property (6) is that all the marginal variables of $\mathbf{Y}_t$ have regularly varying tails with index κ, so in particular for $r \geq 1$

$$\Pr(X_t^2 > rx \mid X_t^2 > x) \to r^{-\kappa}, \qquad \text{as } x \to \infty. \tag{8}$$

3 Tail chain for IGARCH process with asymmetric Z

3.1 Particle filter algorithm

In this Section an algorithm to sample from $\mathbf{D}$ is presented. The property $E(\|\mathbf{AD}_0\|^\kappa) = 1$, proved by [2], is extensively used to this aim. After obtaining κ via Monte Carlo (7), that require as inputs the coefficients and a sample from Z^2, the steps are the following

- Initialize with the estimate of the empirical distribution of $\widetilde{\mathbf{D}}_0$. The empirical distribution is estimated using m extreme values from a simulated squared GARCH(p,q), where the extremes are the m largest values of the squared sequence.
- From the empirical distribution $\widetilde{\mathbf{D}}_0$ we initialize the procedure by taking J particles with equal weight $w_0^{(j)} = 1/J$ and sample from $\widetilde{\mathbf{D}}_0$ with probabilities given by $w_0^{(j)}$. At this step of the sampling is made with replacement, i.e. it is possible to set $J > m$.
- The empirical distribution after the first run of the algorithm is computed by first using the transition
$$\mathbf{D}_1^\star = \mathbf{A}^{(j)}\widetilde{\mathbf{D}}_0, \qquad j = 1,\ldots,J, \tag{9}$$
where the $\mathbf{A}^{(j)}$ are independent copies of $\mathbf{A}$. Since we neglect the random vector $\mathbf{B}$ when computing the transition (9), then $\mathbf{D}_1^\star$ has to be normalized. A proper (yet empirical) distribution can be obtained by simply scaling the $\mathbf{D}_1^\star$, i.e.
$$\widetilde{\mathbf{D}}_1 = \frac{\mathbf{D}_1^\star}{\|\mathbf{D}_1^\star\|}.$$
- The new particles weights are subsequently updated exploiting $E(\|\mathbf{AD}_0\|^\kappa) = 1$. We take advantage of that property by first storing

$$w_j^\star = \|\mathbf{D}_1^\star\|^\kappa, \quad \text{normalized with} \quad w_j^{(1)} = \frac{w_j^\star}{\sum_{j=1}^J w_j^\star}, \qquad j = 1,\ldots,J. \tag{10}$$

- In general particle weights $w_j^{(1)}$ from iteration 1 are no longer identical, and that will be the case in all iterations.
- Iterate S times using recursions similar to (9) and (10), where at each fixed replicate s, $(s = 1,\dots,S)$ the J particles of $\widetilde{\mathbf{D}}_s$ are sampled, with replacement, with updated weights $w_j^{(s)}$, $j = 1,\dots,J$.
- In our algorithm we check that these conditions hold at convergence. We noticed that with small S the weights stabilize, and we have a good sample from $\mathbf{D}_t$, i.e. $\widetilde{\mathbf{D}}_S \to \mathbf{D}_t$.

3.2 Random thinning for the IGARCH

Mapping from the squared process to the "original" process require to rule out extremes in the X_t^2 which do not belong to the X_t process. First note that if X_t^2 is regularly varying with index $\kappa > 0$ then if

$$\Pr(Z_t > x \mid |Z_t| > x) \to \delta \qquad \text{as } x \to \infty, \tag{11}$$

where $0 < \delta < 1$ then it follows that X_t is a regularly varying random variable, with index $\kappa/2$, in both its upper and lower tails.

To translate results about the tail chain of the squared GARCH process we take the first component of $\hat{\mathbf{Y}}_t$ to be denoted as $\{\hat{Y}_t\}$. To study properties for the tail chains of the GARCH $\{\hat{X}_t\}_{t\geq 1}$ process we adopt a similar strategy to [4]. It is key to recognise that there are two tails chains for $\hat{X}_t$, a lower and an upper tail chain $\hat{X}_t^l$ and $\hat{X}_t^u$ respectively, with $\hat{X}_t^u = B_t(\hat{Y}_t)^{1/2}$ and $\hat{X}_t^l = -B_t(\hat{Y}_t)^{1/2}$ where B_t is a sequence of IID Bernoulli(δ) variables, with δ given by limit (11), where $B_t = \{-1, 1\}$ with respective probabilities $\{1-\delta, \delta\}$.

An extreme event for the tail chain $\{\hat{Y}_t\}_{t\geq 1}$ of the squared GARCH does not occur in the upper tail $\{\hat{X}_t^u\}_{t\geq 1}$ and lower tail $\{\hat{X}_t^l\}_{t\geq 1}$ with respective probabilities $P^u(\delta)$ and $P^l(1-\delta)$ where

$$P^u(\delta) = \sum_{i=1}^{\infty} \pi_{X^2}(i)\delta^i$$

and $\pi_{X^2}(i)$ is the probability that a cluster of length i is in the $\{X_t^2\}$ series.

For the upper and lower tail behaviour of $\{X_t\}$ it follows that the respective extremal indices are

$$\theta_X^u(\delta) = \delta^{-1}\theta_{X^2}\{1 - P^u(\delta)\} \tag{12}$$

and $\theta_X^l(1-\delta)$ respectively, where θ_{X^2} is the extremal index of the squared GARCH process. Similarly, the upper and lower tail limiting cluster size distributions of $\{X_t\}$ are given by

$$\pi_X^u(i,\delta) = \Big\{1 - P(\delta)\Big\}^{-1} \sum_{j=i}^{\infty} \pi_{X^2}(j) \binom{j}{i} \delta^i (1-\delta)^{j-i} \text{ for } i = 1, 2, \dots. \tag{13}$$

and $\pi^l_X(i, 1-\delta)$. Thus once we have derived $\pi_{X^2}(\cdot)$, obtaining $\pi_X(\cdot)$ is immediate for both tails of the GARCH process.

4 Some results for the GARCH(2,2)

In this example a GARCH$(2,2)$ process is considered with standard Gaussian innovation Z_t and parameters $\alpha_1 = \alpha_2 = 0.2$, $\beta_1 = 0.3$ and $\beta_2 = 0.25$.

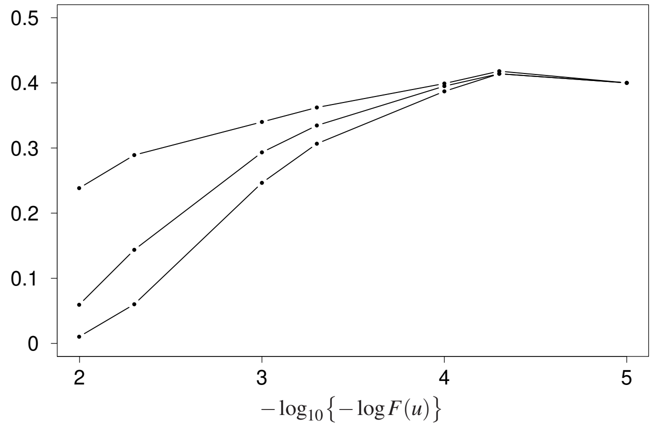


Fig. 1 Runs estimation of the extremal index for a stationary GARCH$(2,2)$ process.

To check our results, we simulate 10^7 values from such a configuration. An estimate of the extremal index with the runs method, which require run length (here $m = 100, 500, 1000$) and a sequence of high thresholds u_n, is considered. The length of the sample size reduces considerably the bias of runs method, even if only one simulation is considered. In Figure 1 we plot u_n on the $-\log_{10}\{-\log F(u_n)\}$ scale. Such choice standardises the upper tail of the distribution. From this plot we can conjecture that the extremal index might lie in the interval 0.25–0.45, but we cannot say more.

From the same GARCH$(2,2)$ structure we now turn to our algorithm, with Monte Carlo value of $\kappa = 1.83$. Figure 2 shows the numerical evaluation of θ and the (little) sensitivity of our algorithm to the number of initial seeds and threshold selection. For each fixed level of threshold, darker points in Figure 2 correspond to lower number of initial seeds. More precisely, white empty circles show results θ with a number of initial seeds higher than other circles and thus more reliable (in this example white filled circles correspond to nearly 250000 initial cluster of extremes, while black filled circles are about 10000 initial seeds).

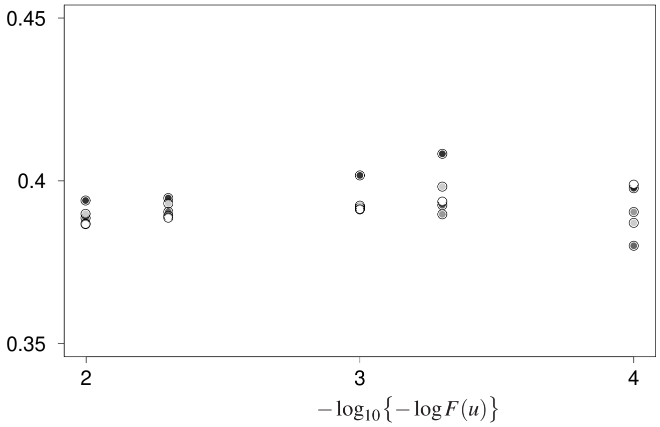


Fig. 2 Sensitivity to θ_X for the GARCH(2,2) process for a range of thresholds and a variety of initial seeds. For a fixed threshold darker circles correspond to a smaller number of initial seeds.

5 Summary and final remarks

We simulate a number of independent seeds conditioning on the event of being at extreme levels. For each initial seed we exploit the autoregressive property of GARCH processes and simulate only clusters of extreme values. The accuracy of our method relies on the number of such cluster of extremes. The usefulness of simulating from within a cluster is two-fold, as we avoid computation inefficiency derived by simulating long sequences of GARCH processes and reduce the influence on the subjective choice of a suitable high threshold.

Acknowledgements The work is part of a broader project involving Paul Fearnhead and Jonathan Tawn, Lancaster University - UK, which are kindly acknowledged. Simone Padoan read carefully an early draft and made a number of useful comments.

References

1. Basrak, B., Davis, R. A., Mikosch, T.: Regular variation of GARCH processes. Stoch Proc Applicat, **99**, 95–115 (2002)
2. Basrak, B. and Segers, J.: Regularly varying multivariate time series. Stoch Proc Applicat, **119**, 1055–1080 (2009).
3. Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, **31**, 307–327.
4. Laurini, F, Tawn, J.A.: The extremal index for GARCH(1, 1) processes. Extremes, **15**, 511–529 (2012)

Spatial Economic Data Analysis

Spatial heterogeneity in principal component analysis: a study of deprivation index on Italian provinces
Eterogeneità spaziale nell'analisi delle componenti principali: uno studio sull'indice di deprivazione sulle province italiane

Paolo Postiglione[1], M. Simona Andreano[2], Roberto Benedetti[3], Alfredo Cartone

Abstract Principal Component Analysis (PCA) is a tool often used for the construction of composite indicators even at the local level ([18]). In general, when we are dealing with spatial data, the method of PCA, in its classical version, is not appropriate for the synthesis of simple indicators. The objective of this paper is to introduce a method to take into account the spatial heterogeneity in PCA, extending the contribution introduced by [19]. The proposed method will be implemented for the definition of a deprivation index on Italian provinces.

Abstract *L'analisi delle componenti principali (ACP) è uno strumento spesso utilizzato per la costruzione di indicatori compositi anche a livello locale ([18]). In generale, quando stiamo lavorando su dati spaziali, l'ACP, nella sua versione classica, non è appropriata per la sintesi di indicatori semplici. L'obiettivo di questo lavoro è di introdurre un metodo che considera l'eterogeneità spaziale nell'ACP, estendendo l'idea di [19]. Il metodo proposto sarà implementato per la definizione di un indice di deprivazione nelle province italiane.*

Key words: Simulated annealing, GWPCA, composite indicators, spatial effects.

1. Introduction

Principal Component Analysis (PCA, [12]) is a statistical method largely adopted in empirical applications. PCA returns a set of independent variables of correlated variables by decomposing the eigen-structure of the variance-covariance matrix ([13]). Typical output of PCA are vectors of loadings corresponding to eigenvectors and new sets of coordinates corresponding to components. PCA is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that

[1] University of Chieti-Pescara, Department of Economic Studies, email: postigli@unich.it

[2] Universitas Mercatorum, email: s.andreano@unimercatorum.it

[3] University of Chieti-Pescara, Department of Economic Studies, email: benedett@unich.it

[4] University of Chieti-Pescara, Department of Economic Studies, email: alfredo.cartone@unich.it

still contains most of the information in the large set. It is also used to the aim of exploring the data.

PCA is often applied on geographically distributed data ([3]). Spatial data present particular characteristics that should be considered when applying statistical technique: spatial dependence and spatial heterogeneity that are the two inherent characteristics of spatial data. Spatial dependence can be defined as "the propensity for nearby locations to influence each other and to possess similar attributes" ([8]). On the other hand, as evidenced by [1], there are two distinct forms of spatial heterogeneity: structural instability and heteroscedasticity. Structural instability concerns the presence of varying structural parameters over space. Heteroscedasticity leads to different error variances of the spatial units.

In this paper, our main aim is to consider the problem of spatial heterogeneity when using PCA for geographically distributed data. In particular, we will address the idea that coefficient estimates can vary across space, leading to spatial structural instability. As argued by [19], the analysis and the assessment of heterogeneity for geographically distributed data is one of the main challenges for the spatial analysts. Empirical models that do not take into account for structural heterogeneities may show serious misspecification problems ([20]).

PCA can also be employed to define composite indicators ([17]). The use of composite indicators is common in practical analyses because we often meet multidimensionality in the real world ([16]). The loadings of PCA may be used as weights in the building of the composite indicators. Some Authors (see, for example, [4]) criticize this use of PCA, because the weights from PCA are defined through a statistical technique and may not reflect the relevance of the single variable for the underlying phenomenon. However, weights from PCA may be less "subjective" because these are not assigned by the researcher and are data-driven, differently from the case of "normative" weights. See [2]) for a discussion about the methods for deriving composite indicators.

Spatial heterogeneity has been considered in PCA through the approach denoted as geographically weighted principal components analysis (GWPCA) ([5]). This method allows for differences in the loadings and scores structure due to spatial instability. The output of GWPCA is represented by estimates of the covariance matrix and sets of components for each locality ([10]). In this way, distinct composite indicators are defined differently for each locality. It is clear that the interpretation of such a list of different composite indicators at local level is not entirely straightforward.

In this paper, we propose to use a modified version of simulated annealing (SA) introduced by [19] to identify zones of local stationarity in the eigenvalues and in the corresponding eigenvectors defined by PCA. The presence of the heterogeneity is a criterion to divide the sample of observations (i.e. regions) into smaller homogeneous groups. Therefore, in our case, we are able to define a composite indicator for each partition identified by SA algorithm.

The use of PCA for deriving composite indicators of deprivation has been extensively explored ([18]). Deprivation may be defined "as a state of observable and demonstrable disadvantage relative to the local community or the wider society or nation to which an individual, family or group belongs" ([21]). Its measurement considers several dimensions from both the social and economic sphere to assess the presence of disparities. In this paper, our aim is to define a composite indicator of deprivation for group of Italian provinces.

The layout of the paper is the following. Section 2 is devoted to briefly summarize the methodological contribution of the paper. In particular, we review the main characteristics of PCA and how SA can be applied to identify zone of local stationarity for eigenvalues and eigenvectors. Section 3 contains the description of our data set and shows the results of the composite indicator for Italian provinces. Finally, section 4 concludes.

2. The methodology

PCA is based on the analysis of a matrix $\mathbf{X}_{nm}$ where $i = 1, \dots, n$ denotes the statistical units and $j = 1, \dots, m$ the variables, respectively. The central idea of PCA is the representation of units in q-dimensional subspaces (with $q < m$) retaining the maximum of statistical information. The reduction of data dimensionality allows us easier interpretative analysis. A primary result in PCA is ([13]):

$$\mathbf{A\Lambda A}^t = \mathbf{\Sigma} \tag{1}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues, $\mathbf{A}$ is the corresponding matrix of loadings (i.e., the eigenvectors), and $\mathbf{\Sigma}$ is the variance/covariance matrix. The eigenvalues in $\mathbf{\Lambda}$ represent the variance of the principal component, $\mathbf{Y}_j$ defined as:

$$\mathbf{Y}_j = \mathbf{Xa}_j \tag{2}$$

where $\mathbf{a}_j$ is the j-th column of the loading matrix $\mathbf{A}$ of $\mathbf{\Sigma}$ and represents the contribution of each variable in $\mathbf{X}$ to the j-th principal component $\mathbf{Y}_j$.

In practice, the component scores related to components $q + 1$ to m represent the Euclidean distances alongside the axes of the corresponding orthogonal vectors to a q-dimensional linear subspace. The first q loadings are chosen so that this subspace contains the highest proportion of the total variance of the data points. In essence, PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables. The first q components are described by:

$$\mathbf{Y} = \mathbf{XA}_q \tag{3}$$

where $\mathbf{Y}$ is the score matrix, $\mathbf{A}_q$ is the loading matrix with the only q columns of $\mathbf{A}$. [13] demonstrates that the best (least squares) rank q approximation to $\mathbf{X}$ is $\mathbf{XA}_q\mathbf{A}_q^t$ and the residual matrix $\mathbf{S}$ can be defined as:

$$\mathbf{S} = \mathbf{X} - \mathbf{XA}_q\mathbf{A}_q^t = \mathbf{XA}_{-q}\mathbf{A}_{-q}^t \tag{4}$$

where $\mathbf{A}_{-q}$ is the loading matrix with the first q columns removed. In the case of application of PCA to spatially distributed data, the underlying implicit hypothesis is that the variance and covariance structure of the process is constant throughout the geographical area under investigation. This assumption is obviously not realistic

([10]). Therefore, it is necessary to relax this hypothesis, to consider in some way the spatial effects in the definition of the principal components.

A first appropriate technique for PCA for spatial data is represented by Geographically Weighted Principal Component Analysis, (GWPCA, [5], [9]). The equation (1) can be generalized to the case of GWPCA as ([10]):

$$\mathbf{A}(u_i, v_i)\mathbf{\Lambda}(u_i, v_i)\mathbf{A}(u_i, v_i)^t = \mathbf{\Sigma}(u_i, v_i) \tag{5}$$

where $\mathbf{\Lambda}(u_i, v_i)$ is the diagonal matrix of local eigenvalues, $\mathbf{A}(u_i, v_i)$ is the corresponding matrix of local eigenvectors, $\mathbf{\Sigma}(u_i, v_i)$ is the local variance-covariance matrix, and (u_i, v_i) are the coordinates of spatial unit i.

The output of GWPCA consists in different loadings and component scores defined for each spatial unit. In practice, GWPCA defines completely different index for each spatial unit as function of distinct loadings. This produces remarkable difficulties in the interpretation of the results.

To simplify the reading of the phenomena, in this paper we propose to apply simulated annealing (SA) algorithm to PCA to identify groups of spatial units that are supposed to share the same eigenvectors (i.e., the same composite indicators). This approach was introduced by [19] and improved by [20] for the analysis of economic growth. The main idea of this framework is that the appropriate treatment of spatial heterogeneity is substantially equivalent to partition an area in groups of geographical zones not necessarily conterminous that have similar component scores. Following this methodology, the output is not represented by different loadings for each spatial unit as in the case of GWPCA, but distinct loadings for every groups of regions identified by SA.

SA is a stochastic relaxation algorithm that was originally introduced in statistical mechanics by [15] and [14]. [7] observes that a spatial combinatorial optimization problem might be described through a Markov Random Field (MRF). The probability measure of a MRF using Gibbs distribution is defined through the energy function $U(\mathbf{X}, \mathbf{k})$, that represent in our algorithm the objective function to be minimized, and a control parameter, T (see [6]; [19]). $U(\mathbf{X}, \mathbf{k})$ depends on observed data $\mathbf{X}$, and the label vector $\mathbf{k} = (k_1, k_2, \dots, k_i, \dots, k_n)$, which categorizes the heterogeneous zones, identifying clusters of regions. $U(\mathbf{X}, \mathbf{k})$ is defined by considering two different effects: a measure of the goodness of fit of the model, and a proximity constraint that describes the extent of aggregation of the spatial units. In particular, at the l-th iteration of the procedure, the energy function is defined as:

$$U(\mathbf{X}, \mathbf{k}) = \beta \sum_{i=1}^{n} I_i \; - (1-\beta) \sum_{r=1}^{n} \sum_{s=1}^{n} \mathbf{c}_{rs} \mathbf{1}_{(k(j)_r = k(j)_s)} \tag{6}$$

where the first part in the right-hand-side is the interaction term, with $I_i = \sum_{k=q+1}^{p} s_i^2$, with s_i the entry of the matrix of the residual matrix $\mathbf{S}$ defined by equation (4); while the second one is the penalty term defined through a Potts model (see [19]). Specifically, $\mathbf{c}_{rs}$ is the element (r, s) of a binary contiguity matrix, $\mathbf{1}_{(k(j)_r = k(j)_s)}$ is the indicator function of the event and $k(j)_r = k(j)_s$, and $(1-\beta)$ is a parameter that discourages configurations with not conterminous units. The parameter $(1-\beta)$ is chosen by the researcher and models the importance of the proximity of the spatial units. Note that the two parts of the energy function (6) are balanced with complementary weight. At the initial value of control parameter T_0, each unit i, is randomly classified as $k_{i,0}$, where $k_{i,0} \in \{1,2,\dots,K\}$ with K is the number of clusters. This step defines the initial configuration S_0. At the $(l+1)$-th iteration, given a current configuration S_l, a different configuration $S_l \neq S_{l+1}$ is randomly chosen, defining a new energy function $U(S_{l+1})$ the is compared with the previous one $U(S_l)$. The old configuration S_l is substituted by the new S_{l+1} in accordance to the probability:

$$Pr_{l,l+1} = max\left\{1, exp\left(-\frac{U(S_{l+1})-U(S_l)}{T_l}\right)\right\} \tag{7}$$

It is worth noting that probability (7) allows to avoid entrapments in local minimum, by defining positive probability for the change of configuration also when the objective function $U(S)$ increases. In essence, more likely patterns (i.e. configurations with lower states of energy) are always accepted, but it is also possible to accept also poorer configurations.

3. Empirical evidence

The proposed methodology is applied to define a deprivation index for Italian provinces. Data set derive from the 15th Population and Housing Census (2011) by Italian National Statistical Institute.

In this paper, a set of ten variables is adopted to build an area-based indicator of material deprivation. The choice of variables has been carried on according to the definition of deprivation index by [17] that suggest choosing a small set of variables able to capture socio-economic deprivation and assist policy makers in a wide set of decisions, for example, public health and tracking inequalities. The variables cover both economic and social domains. Income, educational attainment (proportion of people without high school diploma, School), and employment (Empl) are considered together with social conditions, as the proportion of people living alone (Unip), the percentage of separated, widowed, or divorced people (SVD), and the proportion of single parent families living in each area (Sin_Par). Furthermore, to have a better definition of the deprivation, we include other indicators, and use some of the variables proposed by [11] to assess the level of material deprivation: lack of car possession among resident families (Car), percentage of families living in house of property (Hou), and available surface in residence houses per person (Sqm) are added to assess the level of material deprivation. Moreover, the percentage of foreigners living in the Province (Frg) is considered as an additional variable, particularly to evaluate situation of social exclusion.

PCA is performed on Italian provinces and four components are selected which capture 80% of the variance in the data set. In Table 1, the loadings of the first four components are reported.

The eigenvector corresponding to the first component is characterized by a dominance of the economic variables: employment, income, and house dimension. These are negatively correlated to deprivation. Lack of car possession is positively linked to deprivation, displaying that owning a car decreases the level of material deprivation. On the other side, social variables tend to be lower in magnitude. In the second component, the picture is substantially different with social variables higher in magnitude, and negatively correlated to the deprivation level. Interestingly, the percentage of foreign people reverses its sign, being negative in the first component and positive in the second.

Table 1: Loadings for first 4 components of global PCA.

	PC 1	PC 2	PC 3	PC 4
Empl	-0.500	0.087	-0.101	0.141
Income	-0.307	0.245	-0.316	-0.062
School	0.110	-0.187	0.083	0.879
Sin_Par	-0.125	-0.602	0.027	-0.230
SVD	-0.161	-0.595	-0.100	-0.193
Unip	-0.294	-0.285	-0.428	0.273
Car	0.386	-0.156	-0.375	0.123
Frg	-0.440	0.238	-0.119	0.105
Hou	-0.153	-0.106	0.686	0.114
Sqm	-0.387	-0.085	0.253	0.024
Eigenvalues	1.84	1.50	1.20	1.01
Proportion of variance	0.34	0.22	0.14	0.10
Cumulative variance	0.34	0.56	0.70	0.80

In standard PCA, homogeneity across space is assumed, and the same set of loadings may be used to derive an indicator of deprivation for the whole Country. Nevertheless, spatial heterogeneity could characterize the structure of the variance-covariance matrix. Therefore, the hypothesis of spatial homogeneity could be relaxed, allowing loadings to change according to different spatial configurations.

To avoid potential drawbacks of spatial heterogeneity in PCA, SA is adopted for identifying clusters of spatial units.

In this paper we identify three different clusters, and this spatial configuration is considered for further analysis. The selected combination produces an improvement in the proportion of explained variance when compared to the standard PCA. Finally, a level of $(1-\beta)$=0.3 is chosen, and 4 components are retained for all configurations.

In Figure 1, the groups are mapped, where white denotes the first cluster, light grey the second group, and dark grey the third regime.

The first group (i.e., white) is mainly composed by provinces in the North-East of the Peninsula and some part of Tuscany. Provinces closer to Alps and the Centre - especially in the Apennine mountains - compose the second groups (i.e., light gray), while the third regime (i.e., dark grey) characterizes mostly the Southern part of the Country and few Provinces of the North (e.g. Turin).

As expected, the three clusters show substantial differences in terms of their indicators structures. The eigenvectors of the first component of the three groups are shown in Table 2. Particularly interesting is the impact of the social variables. While in the first group social variables contribute positively to the level of deprivation, in the other spatial clusters the effects (i.e., Sin_Par, SVD, Unip) on the deprivation is negative. Other significant differences in the loadings structure can be found in the heterogeneous effect on deprivation of school attainment and the different magnitude of employment in the diverse regimes.

Figure 1: Clusters of spatial units obtained by Simulated Annealing.

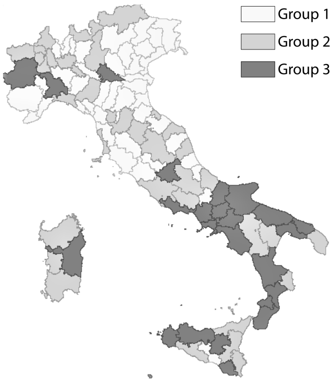


Table 2: Loadings of first component for each group obtained from Simulated Annealing.

	Group 1	Group 2	Group 3
Empl	-0.318	-0.462	-0.360
Income	-0.293	-0.378	-0.007
School	0.283	0.192	0.037
Sin_Par	0.429	-0.245	-0.365
SVD	0.432	-0.300	-0.379
Unip	0.243	-0.349	-0.415
Car	0.380	0.330	0.236
Frg	-0.366	-0.351	-0.280
Hou	-0.081	0.240	-0.320
Sqm	-0.135	-0.211	-0.429
Proportion of Variance	0.40	0.35	0.39

4. Conclusion

In this paper we propose a method for considering spatial heterogeneity in PCA. In fact, when dealing with geographically distributed data, the application of the classical framework of PCA could be misleading and lead to incorrect results. To overcome this drawback, the proposed method extends to PCA, the SA algorithm introduced by [19] for analyzing regional economic growth.

Applying SA to PCA let to highlight different structures of the multivariate phenomenon taking into account the presence of spatial heterogeneity. Results show the differences in the computing the composite indicator in each cluster. Especially the effect of social variables varies from first to second and third groups and a substantial difference of the North provinces with the rest of the Country is highly evident. However, results of SA help policy maker in the interpretation of the global phenomenon, improving interpretability of the indicator levels while considering different spatial regimes.

References

1. Anselin, L.: Spatial econometrics: Methods and models. Kluwer Academic Publishers, Dordrecht (1988).
2. Decanq, K., Lugo, M.A.: Weights in multidimensional indices of wellbeing: An overview. Econom. Rev. 32: 7-34 (2013).
3. Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., McLoone, S.: Principal component analysis on spatial data: An overview. Ann. Assoc. Am. Geogr. 103: 106-128 (2013).
4. De Muro, P., Mazziotta, M., Pareto, A.: Composite indices of development and poverty: An application to MDGs. Soc. Indic. Res. 104: 1-18 (2011).
5. Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically weighted regression - The analysis of spatially varying relationships. Chichester, UK: Wiley (2002).
6. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721-741 (1984).
7. Geman, D., Geman, S., Graffigne, C., Dong, P.: Boundary detection by constrained optimization. IEEE Trans. Pattern Anal. Mach. Intell. 12, 609–628 (1990).
8. Goodchild, M.F.: Geographical data modeling. Comput. Geosci. 18: 401–408 (1992).
9. Harris P., Brunsdon C., Charlton. M. (2011). Geographically weighted principal components analysis. Int. J. Geogr. Inf. Sci. 25: 1717–36.
10. Harris, P., Clarke, A., Juggins, S., Brunsdon, C., Charlton M.: Enhancements to a geographically weighted principal component analysis in the context of an application to an environmental data set. Geogr. Anal. 47: 146-172 (2015).
11. Havard, S., Deguen, S., Bodin, J., Louis, K., Laurent, O., Bard, D.: A small-area index of socioeconomic deprivation to capture health inequalities in France. Soc. Sci. Med. 67: 2007-2016 (2008).
12. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of Educ. Psychol. 24: 417-441 (1933).
13. Jolliffe, I.T.: Principal component analysis. Springer (2002).
14. Kirkpatrik, S., Gelatt, Jr C.D., Vecchi, M.P.: Optimization by simulated annealing. Sci. 220: 671-680 (1983).
15. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092 (1953).
16. OECD: OECD Core set of indicators for environmental performance reviews. Environ. Monogr. 83 (1993).
17. Pampalon, R., Raymond, G.: A deprivation index for health and welfare planning in Quebec. Chronic Dis. Can. 21:104-13 (2000).
18. Pampalon, R., Hamel, D., Gamache, P.: Health inequalities, deprivation, immigration and aboriginality in Canada: A geographic perspective. Can. J. Public Health 101:470-4 (2010).
19. Postiglione, P., Andreano, M.S., Benedetti, R.: Using constrained optimization for the identification of convergence clubs. Comput. Econ. 42: 151-174. (2013).
20. Postiglione, P., Andreano, M.S., Benedetti, R.: Spatial clusters in EU productivity growth. Growth Chang. 48: 40-60 (2017).
21. Townsend, P.: Deprivation. J. So. Policy 16: 125-146 (1987).

Spatial Functional Data Analysis

Object oriented spatial statistics for georeferenced tensor data

Statistica spaziale orientata agli oggetti per dati tensoriali georeferenziati

Alessandra Menafoglio and Davide Pigoli and Piercesare Secchi

Abstract We address the problem of analysing a spatial dataset of manifold-valued observations. We propose to model the data by using a local approximation of the Riemannian manifold through a Hilbert space, where linear geostatistical methods can be developed. We discuss estimation methods for the proposed model, and consistently develop a Kriging technique for tensor data. The methodological developments are illustrated through the analysis of a real dataset dealing with covariance between temperatures and precipitation in the Quebec region of Canada.

Abstract Si considera il problema dell'analisi di osservazioni georeferenziate a valori in una varietà Riemanniana. Si propone di modellare i dati usando approssimazioni locali della variet stessa attraverso opportuni spazi di Hilbert, dove metodi geostatistici lineari possono essere sviluppati. Sono discussi metodi di stima per il modello proposto, ed consistentemente sviluppato un metodo di Kriging per dati tensoriali. Gli sviluppi metodologici sono illustrati attraverso l'analisi di un dataset reale riguardante l'analisi di matrici di covarianza tra temperature e precipitazioni nella regione del Quebec, in Canada.

Key words: Object oriented data analysis, spatial statistics, covariance matrices, tangent space approximation

Alessandra Menafoglio
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, Italy e-mail: alessandra.menafoglio@polimi.it

Davide Pigoli
Department of Mathematics, King's College London, The Strand, London, United Kingdome-mail: davide.pigoli@kcl.ac.uk

Piercesare Secchi
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, Italy e-mail: piercesare.secchi@polimi.it

1 Introduction

The statistical analysis of spatial complex data has recently received much attention in the literature, motivated by the increasing availability of heterogenous datasets in environmental field studies. In this framework, Object Oriented Spatial Statistics (O2S2) (Menafoglio and Secchi, 2017) is a recent system of ideas and methods that allows the analysis of complex data when their spatial dependence is an important issue. The foundational idea of O2S2 is to interpret data as *objects*: the *atom* of the geostatistical analysis is the entire object, which is seen as an indivisible unit rather than a collection of features. In this view, the observations are interpreted as random points within a space of objects – called *feature space* – whose dimensionality and geometry should properly represent the data features and their possible constraints.

In this communication, we focus on the problem of analyzing a set of spatial tensor data. These are georeferenced data whose feature space is a Riemannian manifold. Informally, Riemannian manifolds are *mildly* non-Euclidean spaces, in the sense that they are non-Euclidean, but can be locally approximated through a Hilbert space. In this setting, the linear geostatistics paradigm (Cressie, 1993) cannot be directly applied, as the feature is not close with respect to the Euclidean geometry (e.g., a linear combination of elements in the manifold does not necessarily belong to the manifold). However, following Pigoli *et al.* (2016), we shall discuss the use of a tangent space approximation to locally describe the manifold through a linear space, where the linear methods of Menafoglio et al. (2013) can be applied.

Although the presented approach is completely general, for illustrative purposes we will give emphasis to the case of positive definite matrices. The latter find application in the analysis and prediction of measures of association, such as the covariance between temperature and precipitation measured in the Quebec region of Canada, which are displayed as green ellipses in Figure 1 (data source: Environment Canada on the website http://climate.weatheroffice.gc.ca).

2 A tangent space approximation to kriging for tensor data

To set the notation, call $\mathscr{M}$ a Riemannian manifold and, given a point P in $\mathscr{M}$, let $\mathscr{H}$ be the tangent space at the point P, $\mathscr{H} = T_P\mathscr{M}$. The latter is a Hilbert space when equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ in $\mathscr{H}$. Given two points, the shortest paths between these points on the manifold is called geodesics. Under technical assumptions on $\mathscr{M}$, for every pair $(P;T) \in \mathscr{M} \times T_P\mathscr{M}$, there is a unique geodesic curve $g(t)$ such that $g(0) = P$ and $g(0) = T$. The exponential map is defined as the smooth map from $T_P\mathscr{M}$ to $\mathscr{M}$, which maps a tangent vector $T \in T_P\mathscr{M}$ to the point at $t = 1$ of the geodesic starting in P in direction T. We denote by $\exp_P$ the exponential map in P, and by $\log_P$ its inverse. More details on these definitions and on the properties of Riemannian manifolds can be found, e.g., in (Lee, 2012) and a detailed example for the case of the manifold of positive definite symmetric matrices is discussed in Section 3.

Given a spatial domain $D \subseteq \mathbb{R}^d$ and n locations $s_1,...,s_n$ in D, we indicate by $S_{s_1},...,S_{s_n}$ the manifold-valued observations at those locations (e.g., the covariance matrix between temperature and precipitation of Figure 1). As in classical geostatistics, we assume the data to a partial observation of a random field $\{S_s, s \in D\}$, valued in $\mathscr{M}$. For a location s in the spatial domain D, we model the random element S_s, taking value in $\mathscr{M}$, as

$$S_s(\mathbf{a},P) = \exp_P(A(\mathbf{f}(s);\mathbf{a}) + \delta_s), \tag{1}$$

where, $A(\mathbf{f}(s);\mathbf{a})$ is a drift term defined in the tangent space $\mathscr{H}$, and δ_s is a zero-mean stochastic residual. In this work, we focus on drift terms expressed in a linear form

$$A(\mathbf{f}(s);\mathbf{a}) = \sum_{l=0}^{L} f_l(s) \cdot a_l,$$

where $a_0,...,a_L$ are coefficients belonging to $\mathscr{H}$ and $f_l(s)$ are scalar regressors. We further assume that the random field $\{\delta_s, s \in D\}$, is a zero-mean globally second-order stationary and isotropic random field in the Hilbert space $\mathscr{H}$, with covariogram C (Menafoglio et al., 2013), i.e., for s_i, s_j in D,

$$C(\|s_i - s_j\|_d) = \mathbb{E}[\langle \delta(s_i), \delta(s_j) \rangle^2_{\mathscr{H}}],$$

$\|s_i - s_j\|_d$ denoting the distance between s_i, s_j in D. We denote by $\Sigma \in \mathbb{R}^{n\times n}$ the covariance matrix of the array $\delta = (\delta_{s_1},...,\delta_{s_n})^T$ in $\mathscr{H}^n$, that is $\Sigma_{ij} = C(\|s_i - s_j\|_d^2)$, and call $R \in \mathscr{H}^n$ the array of residuals $R_i = A(\mathbf{f}(s);\mathbf{a}) - \log_P(S_i)$. Given the array R and a matrix $A \in \mathbb{R}^{p\times n}$, we define the matrix operation AR as $(AR)_i = \sum_{j=1}^n A_{ij} R_j$, $i = 1,...,p$.

Given the observations $S_{s_1},...,S_{s_n}$, we now aim to estimate the model (1), and make prediction at unsampled locations. To estimate $(P,\mathbf{a})$ accounting for the spatial dependence, a generalized least square (GLS) criterion, based on minimizing the functional

$$(\widehat{P},\widehat{a}) = \underset{P\in\mathscr{M},\mathbf{a}\in\mathscr{H}^{L+1}}{\operatorname{argmin}} ||\Sigma^{-1/2}R||^2_{\mathscr{H}^n}, \tag{2}$$

can be used. In (2), $\mathscr{H}^n$ denotes the cartesian space $\mathscr{H} \times \cdots \times \mathscr{H}$, which is a Hilbert space when equipped with the inner product $\langle \mathbf{x},\mathbf{y} \rangle_{\mathscr{H}^n} = \sum_{i=1}^n \langle x_i, y_i \rangle_{\mathscr{H}}$. Given Σ, problem (2) can be solved iteratively, by alternatively minimizing the GLS functional in (2) with respect to P given $\mathbf{a}$ and viceversa. Since in practice both the parameters and the spatial dependence are unknown, one needs to resort to a nested iterative algorithm. The complexity of such minimization is problem dependent, and may require the development of specific optimization techniques.

Given the estimated $(\widehat{P},\widehat{a},\widehat{\Sigma})$, the spatial prediction can be performed by using the tangent space model as follows. In the Hilbert space $\mathscr{H}$, the simple kriging predictor for δ_{s_0} is well-defined and it is obtained as $\sum_{i=1}^n \lambda_i^0 \widehat{\delta}_{s_i}$, where $\widehat{\delta}_{s_i}$ indicates the estimated residual at s_i, $\widehat{\delta}_{s_i} = A(\mathbf{f}(s_i);\widehat{a}) - \log_{\widehat{P}}(S_i)$, and the vector of kriging weights $\lambda_0 = (\lambda_1^0,...,\lambda_n^0)$ is found as $\lambda_0 = \widehat{\Sigma}^{-1} c$, with $c = (\widehat{C}(||s_1 -$

$s_0||_d), \ldots, \widehat{C}(||s_n - s_0||_d))^T$. The spatial prediction of S at the target location s_0 is then

$$\widehat{S}_0 = \exp_{\widehat{P}}(\widehat{a}_0^{GLS}(\widehat{P}) + \sum_{l=1}^{L} \widehat{a}_l^{GLS}(\widehat{P}) f_l(s_0) + \sum_{i=1}^{n} \lambda_i^0 \widehat{\delta}_{s_i}),$$

where $\mathbf{f}(s_0)$ is the vector of covariates given at the location s_0. Uncertainty quantification of such estimate can be performed by resampling methods, e.g., via bootstrap (Pigoli *et al.*, 2016).

3 Analysis of covariance matrices in the Quebec region

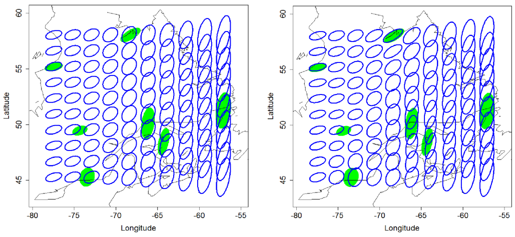
We here discuss the application of the method recalled in Section 2 to the covariance matrices displayed in Figure 1. Those data where estimated from temperature-precipitation data recorded in the month of January along the years 1983-1992.

Recall that the covariance matrix of a p-variate random variable belongs to the Riemannian manifold $PD(p)$ of positive definite matrices of dimension p, which is a convex subset of $\mathbb{R}^{p(p+1)/2}$ but it is not a linear space. The tangent space $T_P PD(p)$ to $PD(p)$ in the point $P \in PD(p)$ can be identified with the space of symmetric matrices of dimension p, $Sym(p)$. A Riemannian metric in $PD(p)$ is then induced by the inner product in $Sym(p)$. Following (Pigoli *et al.*, 2016), we consider the scaled Frobenius inner product in $Sym(p)$, which induces the exponential map $\exp_P(A) = P^{\frac{1}{2}} \exp(P^{-\frac{1}{2}} A P^{-\frac{1}{2}}) P^{\frac{1}{2}}$, and the logarithmic map $\log_P(D) = P^{\frac{1}{2}} \log(P^{-\frac{1}{2}} D P^{-\frac{1}{2}}) P^{\frac{1}{2}}$, where $\exp(A)$ stands for the exponential matrix of $A \in Sym(p)$, and $\log(C)$ for the logarithmic matrix of $C \in PD(p)$.

The linear model for the drift in the tangent space was set to $A(\phi_i, \lambda_i) = a_0 + a_1 \phi_i$, (ϕ, λ) denoting longitude and latitude. Such model was chosen by Pigoli *et al.* (2016) as to guarantee the stationarity of the residuals of (2). The drift coefficients and the structure of spatial dependence were estimated by numerically optimizing functional (2). The estimated drift and the predicted field are displayed in Figure 1a-b. A possible meteorological interpretation is associated with the exposition of the region toward the sea. Indeed, the drift model accounts for the distance between the location of interest and the Atlantic Ocean, which is likely to influence temperatures, precipitations and their covariability.

4 Conclusion and discussion

Object Oriented Spatial Statistics allows dealing with general types of data, by using key ideas of spatial statistics, revised according to a geometrical approach. In this communication we focused on the spatial analysis of tensor data, through the use of a tangent space approximation. Such approximation is appropriate to threat observations whose variability on the manifold is not too high. Simulation studies



(a) Estimated drift (b) Kriging prediction

Fig. 1 Kriging of the (temperature, precipitation) covariance matrix field during January, with a drift term depending on longitude. A covariance matrix S at location s is represented as an ellipse centered in s and with axis $\sqrt{\sigma_j}e_j$, where $S\mathbf{e}_j = \sigma_j e_j$ for $j = 1,2$. Horizontal and vertical axes of the ellipses represent temperature and precipitation respectively. In subfigure (a) and (b) green ellipses indicate the data, blue ellipses the estimated drift and the kriging interpolation, respectively. Modified from (Pigoli *et al.*, 2016).

(Pigoli *et al.*, 2016) showed that the method is robust to a moderate increase of the variability on the manifold. However, in cases characterized by a very high variability, more complex models should be used. A recent extension of the model, which is currently investigated by the authors, regards the use of local tangent space models to describe the field variability. This approach is based on the idea of embedding the model here illustrated in a novel computation framework developed in (Menafoglio *et al.*, 2018) and based on the idea to repeatedly partition the domain through Random Domain Decompositions. Such an extension will potentially allow to improve the characterization of the field variability, and the associated predictions.

References

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Lee, J. (2012) *Introduction to Smooth Manifolds*, 218, Springer Science & Business Media.

Menafoglio, A., P. Secchi, and M. Dalla Rosa (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics 7*, 2209–2240.

Menafoglio, A. and P. Secchi (2017). Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European Journal of Operational Research 258*(2), 401–410.

Menafoglio, A., Gaetani, G., Secchi, P. (2018) Random domain decompositions for object-oriented kriging over complex domains, *MOX-report 10/2018*, Politecnico di Milano.

Pigoli, D., Menafoglio, A., Secchi, P. (2016) Kriging prediction for manifold-valued random fields. *Journal of Multivariate Analysis*, **145**, 117–131.

A Spatio-Temporal Mixture Model for Urban Crimes

Un Modello Mistura per Dati Spazio-Temporali Relativi alla Criminalità

Ferretti Angela, Ippoliti Luigi and Valentini Pasquale

Abstract This paper considers the determinants of severe crimes at the census-tract level in Pittsburgh, Pennsylvania. We develop a mixture panel data model to describe the number of severe crimes that allows for temporal as well as spatial correlation, together with significant heterogeneity across census tracts. We use traditional Bayesian mixtures admitting uncertainty about the number of groups. We focus on pooling regression coefficients across clusters, implying that census-tracts belonging to the same cluster are similar. The clustering is done in a data-based fashion.

Abstract *In questo articolo ci proponiamo di studiare le determinanti dei reati gravi verificatisi nei distretti della città di Pittsburgh (Pennsylvania). A tal fine, si propone una mistura di modelli di regressione per dati panel che consente di cogliere la correlazione temporale e spaziale, nonchè l'eterogeneità tra i distretti. Assumendo come incognito il numero delle componenti della mistura, il modello consente di pervenire ad una classificazione in cui i gruppi si distinguono per diversi profili di covariate.*

Key words: Crime places, Distribution dynamics, Spatio-temporal models, Dirichlet process

Ferretti Angela
DeC, University ”G.d'Annunzio” of Chieti-Pescara, e-mail: angela.ferretti29@gmail.com

Ippoliti Luigi
DeC, University ”G.d'Annunzio” of Chieti-Pescara e-mail: luigi.ippoliti@unich.it

Valentini Pasquale
DeC, University ”G.d'Annunzio” of Chieti-Pescara e-mail: pasquale.valentini@unich.it

1 Introduction

Any individual behaviour is a product of interaction between the person and the setting [1]. In recent years, the spatio-temporal urban distribution of crimes is receiving growing attention not only from researchers (criminologists, sociologists, economists, geographers, etc) but also from law enforcement agencies. In particular, in [11] it is highlighted a need to “integrate geographic and temporal representation and analyses” and in [10] it is stated that “the most under-researched area of spatial criminology is that of spatio-temporal crime patterns”. In this paper, we aim at addressing such needs by proposing a mixture panel data model for high-dimensional urban crime count data. The model allows to include temporal and spatial effects, socio-economic census tract characteristics and random effect components to take care of the heterogeneity existing across census tracts. Additionally, our model extends a “traditional” panel data model in many ways. For example, model coefficients are pooled across clusters, implying that census tracts belonging to the same cluster are similar. However the number of clusters is not known in the data but it is determined by the Dirichlet process.

The paper is organized as follows. In section 2 we present our model and the econometric methodology. In section 3 we describe our empirical dataset and discuss the empirical results.

2 Model Specification

Let y_{it} be the number of Part I offenses in census tract i $(i = 1,...,N)$ at time t $(t = 1,...,T)$. Let us assume that, conditionally on the mean, the y_{it}'s are mutually independent with Poisson distribution, $y_{it} \sim Po(\exp(\eta_{it}))$. The logarithm of the conditional mean is given as follows:

$$\eta_{it} = \nu + \phi \eta_{it-1} + \rho \sum_{l=1}^{N} W_{il} \eta_{lt} + x_{it}' \beta + \delta_{it} \qquad (1)$$

where x_{it} is a vector including strictly exogenous variables, ν is the intercept, β is a vector of regression parameters, η_{it-1} is the temporally lagged value of η_{it}, W_{il} is a generic element of a matrix W reflecting contiguity relations between the N census tracts, ρ is a scalar parameter reflecting the strength of spatial dependence, and δ_{it} can be related to a set of common determinants as $\delta_{it} = \xi_t' \gamma_i + \varepsilon_{it}$ where $\varepsilon_{it} \sim N(0, \sigma^2)$. The linear combination $\xi_t' \gamma_i$ transfers the contemporaneous correlation from the errors to the conditional expectation part of the model. Since in our empirical analysis we choose to include the first latent factor (principal component) as common regressor ξ_t, γ_i can be interpreted as a parameter vector of factor loadings specific for each census tract.

We consider a generalization of model (1) as used in [9], where the number of Part I offenses is modeled as a mixture of C^* unobserved clusters, whose coefficients

are pooled only across census tracts having similar characteristics. Particularly, the cluster-specific coefficients imply that census tracts belonging to the same cluster are defined by common effects, while census tracts belonging in different clusters have structural differences in their severe crimes' determinants.
Let $r_i = j$ index that census tract i belongs to cluster j, for $j = 1,...,C^*$ clusters in total, and with a respective probabilities $\pi_{i1} = P[r_i = 1], \pi_{i2} = P[r_i = 2], ..., \pi_{iC^*} = P[r_i = C^*]$, where $0 \leq \pi_{i1}, \pi_{i2}, ..., \pi_{iC^*} \leq 1$ and $\sum_{j=1}^{C^*} \pi_{ij} = 1$. The model we use can be written as

$$\eta_{it} = \nu_j + \phi_j \eta_{it-1} + \rho \sum_{l=1}^{N} W_{il} \eta_{lt} + x'_{it} \beta_j + \delta_{it} \tag{2}$$

$$if \quad r_i = j \qquad\qquad j = 1,...,C^*;$$

where ϕ_j and $\beta_j, j = 1,2,...,C^*$ are cluster specific coefficients.

2.1 Dirichlet Process Mixture Model

One of features of model (2) is the number of clusters C^*. In order to address to this question, in this paper we adopt a truncated Dirichlet process model to define the prior over the mixing probabilities based on some (large) upper bound C (see [4]).

Denote the respose variable $\eta_i = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT})'$ and a set of covariates $z_{it} = \left(1, \eta_{it-1}, \sum_{l=1}^{N} W_{il} \eta_{lt}, x'_{it}\right)'$ observed at time t for the ith individual. With no loss of generality here, we rewrite model (2) as $\eta_{it} = z'_{it} \tau_j + \varepsilon_{it}$ where $\tau_j = (\nu_j, \rho, \phi_j, \beta'_j)'$ and $\delta_{it} = \varepsilon_{it}$. Then the density of the mixture is

$$f(\eta_i | z, \theta) = \sum_{j=1}^{C} \pi_j \left(\prod_{t=1}^{T} f_j\left(\eta_{it} | z_{it}, \psi_j\right) \right)$$

where $\theta = \{\alpha, \pi_{1:C}, \psi_{1:C}\}$, $\psi_j = \{\tau_j, \sigma_j^2\}$, $\sum_{j=1}^{C} \pi_j = 1$ with $0 \leq \pi_j \leq 1$, f_j are the C component densities and α is a precision parameter of the Dirichlet process. The mixture model can be realized through the configuration indicators r_i for each observation η_i with prior $P(r_i = j | \pi) = \pi_j$, so that we obtain the standard hierarchical model:

$$(\eta_i | r_i = j, z_i, \psi_j) \sim f_j(\eta_i | \psi_j). \qquad (\psi_j | G) \sim G \qquad (G | \alpha, G_0) \sim DP(\alpha, G_0). \tag{3}$$

where $G(\cdot)$ is an uncertain distribution function, $G_0(\cdot)$ is the prior mean of $G(\cdot)$ and $\alpha > 0$ the total mass, or precision of the DP. From the *Pólya Urn Scheme*,

$$\psi_j | \psi_1, \psi_2, ..., \psi_{j-1} \sim \frac{\alpha}{j-1+\alpha} G_0(\cdot) + \frac{1}{j-1+\alpha} \sum_{k=1}^{j-1} \delta_{\psi_k}(\cdot) \tag{4}$$

where $\delta_{\psi_k}(\cdot)$ is the point mass distribution at ψ_k. The truncated Dirichlet process prior is such that

$$\pi_1 = V_1, \qquad \pi_j = V_j \times \prod_{i=1}^{j-1}(1 - V_i), \tag{5}$$

$j > 1$, where V_i has a Beta distribution $Be(1,\alpha)$, $i < C$ indipendently over i and $V_C = 1$. Prior specification for each component j ($j = 1,...,C$) is completed with the following distribution,

$$G_0(\tau_j, h_j^{-1}) = N(\tau_j|\tau_0, h_j^{-1})Ga(h_j|a,b) \tag{6}$$

where $h_j^{-1} = \sigma_j^2$; and with a Gamma prior $\alpha \sim Ga(\varsigma_1, \varsigma_2)$. Placing a prior on α ([4]) allows us to draw inferences about the number of mixture components through the role of α of the *Pólya Urn Scheme* as the prior number of observations in each component.

3 Application

In this section we apply model (2) to study the determinants of census tract severe crimes in Pittsburgh, Pennsylvania. We first describe the full dataset and next we give details on empirical results.

3.1 Data

The crime dataset that we used includes monthly (January 2008 to December 2013) counts of Part I and Part II offenses for each of the 138 2000 census tracts in Pittsburgh, Pennsylvania. Part I offenses, also known as index crimes, regroup serious felonies in the following eight categories: criminal homicide, forcible rape, robbery, aggravated assault, burglary, larceny-theft (except motor vehicle theft), motor vehicle theft and arson. Part I offenses consist of the number of offenses in these categories that are known to law enforcement. Part II offenses include 21 categories of non-serious felonies and misdemeanors for which only arrest data were collected. A more detailed description of these variables are provided in [7].

The dependent variable in our study, y_{it}, is the number of Part I offenses in census tract i for $i = 1,...,138$ in month t for $t = 1,...,72$. Potential covariates include the log of number of Part I offenses in census tract at time $t-1$, the log of Part II offenses lagged by 1 month as leading indicator and the spatially lagged state variable. In addition, in order to account for heterogeneity across census tracts, we collected data on the following 15 time-invariant socio-economic variables from the Census 2000 (US Census Bureau and Social Explorer Tables): log of median income (*Lmi*), civilian unemployment rate (*Cur*), poverty rate (*Pvr*), percentage of population with less than a high school degree (*Hdl*), percentage of population with a bachelor degree or higher (*Bdh*), rental housing units as percentage of occupied housing units (*Rhu*), percentage of households having been in the same house for more than 1 year (*Sh1*), percentage of female-headed households (*Fhh*), housing units vacancy rate (*Hvr*), percentage of total population that is African-American (*Paa*), log of total population (*Ltp*), log of population density per square mile (Lpd), dropout rate age 16–19 (*Dra*), percentage of total population under 18 ($U18$) and group quarter proportion (*Gqp*). Finally, missing values for our socio-economic covariates in 14 census tracts that do not have a regular resident population were replaced by dummies.

3.1.1 Results

By means of a Markov chain Monte Carlo approach, posterior inference was based on the last $50,000$ draws (after a burn-in of $5,000$) using every 5th member of the chain to avoid autocorrelation within the sampled values. From the computational viewpoint, we first sample η_{it} from its marginal distribution using the adaptive rejection sampling [2] and then we draw all the other parameters. Conditional on η_{it}, the full conditional posterior distributions take convenient functional forms and can be easily sampled from. Convergence of the chains of the model was monitored visually through trace plots as well as using the R-statistic of [3] on two chains simultaneously started from different initial points.

Results indicate the existence of two clusters of census tracts in Pittsburgh, with 42 census tracts (30%) belonging to cluster 1 and 96 census tracts (70%) belonging to cluster 2. Also, we note that the temporal correlation parameters referring to the Part II crime data, only impact in cluster 1. This suggests that crime hot spots may arise first as a concentration of soft crimes that later hardens into more serious crimes. Consistently, if a large number of Part I offenses happen at time $t-1$, a huge number of the same crimes will occur at time t. The spatial dependence in Part I offenses is relevant in both clusters, so we do have a spatial diffusion of certain types of crime in Pittsburgh. This result could provide a useful tool for efficient allocation of law enforcement resources.

Of the 15 socio-economic determinants of severe crimes, only the *Cur* seems to be relevant in both clusters. The positive influence of civilian unemployment rate on the number of Part I offenses confirms the social organization theory according to which bad socio-economic conditions, such as job unavailability, give rise to criminal motivation. For the rest of potential determinants we find that they can have an impact in one cluster but not the other. While *Ldp* and *Fhh* do not provide any impact in group 1, they become important in group 2. Here, census tracts with a small population size or lack of residential instability should enjoy lower number of Part I offenses. Furthermore, the variables *Lmi*, *Pvr* and *U18* appear with a negative sign in cluster 2. In contrast, *Hvr* and *Hdl* represent important determinants for the Part I offenses with positive and negative signs, respectively, only in cluster 1.

Overall, this study shows that criminal dynamics have different features across the two clusters, with differences which cannot be captured by traditional regression analyses.

References

1. Felson, M., Clarke, R.V.: Opportunity Makes the Thief: Practical Theory for Crime Prevention. In: Police Research Studies. London (1998)
2. Gelman, A.: Inference and Monitoring Convergence. In Gilks, W.R and Richardson, S. and Speigelhalter, D. (eds). Chapman & Hall:Boca Raton; 131-143 (1996)
3. Gilks, W.R. and Wild, P. : Adaptive Rejection sampling for Gibbs Sampling. Journal of the Royal Statistical Society, Series C, **41**, 337–348 (1992)
4. Ishwaran, H., James, L.: Approximate Dirichlet process computing in finite normal mixture: Smoothing and prior information. J. Comput. Graph. Stat. **11**, 508–532 (2002)
5. Kikuchi, G.: Neighborhood Structures and Crime: A Spatial Analysis. El Paso, Texas (2010)
6. Korobilis, D., Gilmartin, M.: On regional unemployment: An empirical examination of the determinants of geographical differentials in the UK. Scott. J. Polit. Econ. (2012) doi: 10.1111/j.1467-9485.2011.00575.x
7. Liesenfeld, R., Richard, J.F., Vogler, J.: Likelihood-based inference and prediction in spatio-temporal panel count models for urban crime. J. Appl. Econ. (2016) doi: 10.1002/jae.2534
8. Manolopoulou, I., Chan, C., West, M.: Selection sampling from large data ses for targeted inference in mixture modeling. Bayesian Anal. **3**, 429–450 (2010)
9. Paap, R., Franses, P.H., van Dijk, D.: Does Africa grow slower than Asia, Latin America and the Middle East? Evidence from a new data-based classification method. J. Dev. Econ. **77**, 553–570 (2005)
10. Ratcliffe, J.H.: Crime mapping: spatial and temporal challenges. Handbook of Quantitative Criminology, Piquero, A.R., Weisburd, D. (eds). Springer: New York, 524 (2013)
11. Roth, R.E., Ross, K.S., Finch, B.G., Luo, W., MacEachren, A.M.: Spatiotemporal crime analysis in U.S. law enforcement agencies: current practices and unmet needs. J. Gov. Inf. **30**, 226–240 (2013)

Statistical Methods for Service Quality

Cumulative chi-squared statistics for the service quality improvement: new properties and tools for the evaluation

Il chi quadrato cumulato per la qualità dei servizi: nuove proprietà e strumenti per la valutazione

Antonello D'Ambra, Antonio Lucadamo, Pietro Amenta, Luigi D'Ambra

Abstract In service quality evaluation, data are often categorical variables with ordered categories and collected in two way contingency table. The Taguchi's statistic is a measure of the association between these variables as a simple alternative to Pearson's test. An extension of this statistic for three way contingency tables handled in two way mode is introduced. We highlight its several properties, the approximated distribution, a decomposition according to orthogonal quantities reflecting the main effects and the interaction terms, and an extension of cumulative correspondence analysis based on it.

Abstract *Nella valutazione della qualità dei servizi erogati, i dati rappresentano spesso variabili qualitative ordinali raccolte in tabelle di contingenza a due vie. L'indice di Taguchi è una misura dell'associazione esistente tra queste variabili e nasce come un'alternativa al test di Pearson in presenza di variabili ordinali. In questo lavoro viene presentata una estensione di questo indice per tabelle di contingenza a tre vie. Se ne evidenziano diverse proprietà, la distribuzione approssimata, una decomposizione rispetto a quantità che riflettono gli effetti principali e l'interazione, nonchè un'estensione dell'analisi delle corrispondenze.*

Key words: Two and three way contingency tables, Chi-squared statistic, Taguchi's statistic, Interaction term, Main effects, Service quality evaluation.

Antonello D'Ambra
Department of Economics, Second University of Naples, Italy, Corso Gran Priorato di Malta, Capua, e-mail: antonello.dambra@unina2.it

Antonio Lucadamo, Pietro Amenta
Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Italy, Piazza Arechi II, Benevento, e-mail: alucadam@unisannio.it, amenta@unisannio.it

Luigi D'Ambra
Department of Economics, Management and Institutions, University of Naples, Italy, Via Cinthia Monte Sant'Angelo, Napoli, e-mail: dambra@unina.it

1 Introduction

Service companies have given increasing importance to customer satisfaction (hereafter CS) over the years worldwide. Measuring the quality of a service is indeed a fundamental and strategic function for every firms because it allows checking the level of efficiency and effectiveness perceived by users. In service quality evaluation, data are often categorical variables with ordered categories and usually collected in two way contingency table. To determine the nature of the association, tests involving the Pearson chi-squared statistic are generally considered. However, the statistic does not take into account the structure of ordered categorical variables [1]. To overcome this problem, Taguchi [6, 7] developed a simple statistic that does take into consideration the structure of an ordered categorical variable. It does so by considering the cumulative frequency of the cells of the contingency table across the ordered variable. An extension of this statistic for three way contingency tables handled in two way mode is introduced in section 4, highlighting some properties and its approximated distribution. Moreover, an extension of correspondence analysis based on the suggested new statistic is proposed to study the association from a graphical point of view. It highlights the impacts of the main effects and the interaction terms on the association. This is obtained in section 5 by means of a decomposition of the new statistic according to orthogonal quantities reflecting several effects.

An application on real data about service quality evaluation using all the theoretical results will be shown in the extended version of this paper.

2 Notations

Let A, B, and Y be categorical variables with $i = 1, \dots, I$, $k = 1, \dots, K$ and $j = 1 \dots, J$ categories, respectively, and suppose $(A_1, B_1, Y_1), \dots, (A_n, B_n, Y_n)$ is a random sample of the random vector (A, B, Y). The basic data structures in this paper are two and three-way contingency tables $\mathbf{N}$ and $\check{\mathbf{N}}$ of orders (I, J) and (I, K, J) with frequencies $\{n_{ij}\}$ and $\{n_{ikj}\}$ counting the numbers of observations that fall into the cross-categories $i \times j$ and $i \times k \times j$, respectively. $\mathbf{N}$ cross classifies n statistical units according to two categorical variables A and Y while $\check{\mathbf{N}}$ according to three categorical variables A, B, and Y. Table $\check{\mathbf{N}}$ is handled in this paper in two way mode by row unfolding it according to variables A and B: the resulting two way contingency table $\tilde{\mathbf{N}}$ is then of size $[(I \times K) \times J]$ with general term n_{ikj}. We consider row variables A and B as predictors and the column variable Y as response, reflecting a unidirectional association between the categorical variables (rows versus column). Moreover, suppose that Y has an ordinal nature with increasing scores.

We denote by p_{ij} the probability of having an observation fall in the i-th row and j-th column of the table, with $\mathbf{P} = \{p_{ij} = n_{ij}/n\}$. $p_{i.} = \sum_{j=1}^{J} p_{ij}$ and $p_{.j} = \sum_{i=1}^{I} p_{ij}$ denote the probabilities that A and Y are in categories i and j, respectively.

Considering the two way table $\mathbf{N}$, let $z_{is} = \sum_{j=1}^{s} n_{ij}$ and $z_{.s} = \sum_{j=1}^{s} n_{.j}$ be the cumulative count and the cumulative column total up to the j-th column category, respectively, with $s = 1,\dots,J-1$. $d_s = z_{.s}/n$ denotes the cumulative column proportion. Let $\tilde{\mathbf{P}} = \{p_{ikj} = n_{ikj}/n\}$ be the joint relative frequency distribution. Moreover, let $\mathbf{D}_I = \{p_{i.} = \sum_{j=1}^{J} p_{ij}\}$ and $\mathbf{D}_J = \{p_{.j} = \sum_{i=1}^{I} p_{ij}\}$ be diagonal matrices containing the row and column sum of $\tilde{\mathbf{P}}$, respectively. Let $\mathbf{D}_{IK} = diag(p_{ik.})$ (marginal row) and $\mathbf{D}_J = diag(p_{..J})$ (marginal column) be also the diagonal matrices with generic elements $p_{ik.} = \sum_{j=1}^{J} p_{ikj}$ and $p_{..j} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ikj}$, respectively.

Lastly, denote $C_{iks} = \sum_{j=1}^{s} n_{ikj}$ with $s = 1,\dots,J-1$ the cumulative frequencies of the $\{ik\}$-th row category up to the s-th column categories. Their consideration provides a way of ensuring that the ordinal structure of the column categories is preserved. Similarly, denote, $\tilde{d_s} = \sum_{j=1}^{s} p_{..j}$ the cumulative relative frequency up to the s-th column category.

3 The Taguchi's statistics in a nutshell

Taguchi [6, 7] proposed a measure of the association between categorical variables where one of them possesses ordered categories by considering the cumulative sum of cell frequencies across this variable. He introduced this measure as a simple alternative to Pearson's test in order to consider the impact of differences between adjacent ordered categories on the association between row and column categories. In order to assess the unidirectional association between the row and (ordered) column variables, Taguchi [6, 7] proposed the following statistic

$$T = \sum_{s=1}^{J-1} \frac{1}{d_s(1-d_s)} \sum_{i=1}^{I} n_{i.} \left(\frac{z_{is}}{n_{i.}} - d_s\right)^2 \qquad (1)$$

with $0 \leq T \leq [n(J-1)]$. This statistic performs better than Pearson's chi-squared statistic when there is an order in the categories on the columns of the contingency table and it is more suitable for studies (such as clinical trials) where the number of categories within a variable is equal to (or larger than) 5 [8].

Takeuchi and Hirotsu [8] and Nair [3] showed also that the T statistic is linked to the Pearson chi-squared statistic $T = \sum_{s=1}^{J-1} \chi_s^2$ where χ_s^2 is Pearson's chi-squared for the $I \times 2$ contingency tables obtained by aggregating the first s column categories and the remaining categories $(s+1)$ to J, respectively. For this reason, the Taguchi's statistic T is called the *cumulative chi-squared statistic* (hereafter CCS). Nair [3] considers then the class of CCS-type statistics

$$T_{CCS} = \sum_{s=1}^{J-1} w_s \left[\sum_{i=1}^{I} n_{i.} \left(\frac{z_{is}}{n_{i.}} - d_s\right)^2\right] \qquad (2)$$

corresponding to a given set of weights $w_s > 0$. The choice of different weighting schemes defines the members of this class. Examples of possible choices for w_j are to assign constant weights to each term (i.e. $w_s = 1/J$) or assume it proportional to the inverse of the conditional expectation of the s-th term under the null hypothesis of independence (i.e. $w_s = [d_s(1-d_s)]^{-1}$). It is evident that T_{CCS} subsumes T in the latter case. Moreover, Nair shows that T_{CCS} with $w_s = 1/J$ (that is $T_N = \sum_{s=1}^{J-1}(1/J)\sum_{i=1}^{I} N_{i.}(z_{is}/n_{i.} - d_s)^2$) has good power against ordered alternatives.

Nair [3, 4] highlighted the main properties of the CCS-type tests by means of a matrix decomposition of this statistic into orthogonal components. Lastly, Taguchi’s statistics can be also viewed as an approximate sum of likelihood ratios within the regression model for binary dependent variables following a scaled binomial distribution, providing in this way a different interpretation of this statistic [2]. Refer to [2] for a wider and deeper study with other new interpretations and characteristics of this statistic.

4 Cumulative Correspondence Analysis and Taguchi’s Statistics for three way contigency tables handled in two way mode

In this paper we introduce a new extension of the Taguchi’s statistic on a three-way contingency table, where one of the variables consists of ordered responses, handled in two way mode. We name ”Multiple Taguchi’s statistic” the following measure of the unidirectional association between the rows and (ordered) column variables

$$T^M = \sum_{s=1}^{J-1} \frac{1}{\tilde{d_s}(1-\tilde{d_s})} \left[\sum_{i=1}^{I} \sum_{k=1}^{K} n_{ik.} \left(\frac{C_{iks}}{n_{ik.}} - \tilde{d_s} \right)^2 \right] \qquad 0 \le T_M \le n(J-1) \tag{3}$$

Likewise formulas (1) and (2) it is also possible to consider a class of CCS-type statistics T^M_{CSS} corresponding to a given set of weights $w_s > 0$. The choice of different weighting schemes defines the members of this class.

It is possible to show that there is a link between Multiple Taguchi’s statistic, Pearson Chi-Squared statistic and C-Statistics $T^M = \sum_{s=1}^{J-1} \chi^2(s) = \frac{n}{n-1} \sum_{s=1}^{J-1} C(s)$. Here $\chi^2(s)$ and $C(s) = (n-1)[\sum_{i=1}^{I} \sum_{j=1}^{s} p_{ik.} \left(p_{ikj}/p_{ik.} - p_{..j}\right)^2]/(1-\sum_{j=1}^{J} p^2_{..j})$ are the Pearson chi-squared and the C-statistics, respectively, for a $[(I \times K) \times 2]$ contingency table obtained by aggregating the first j column categories and the remaining categories $(j+1)$ to J.

It is possible to highlight other properties by means of a matrix decomposition of the CCS-type statistic T^M_{CSS} into orthogonal components. For instance, this allows to introduce the Multiple Taguchi’s statistic at heart of a new cumulative extension of correspondence analysis (hereafter MTA). Main goal of MTA is to show how similar cumulative categories are with respect to joined nominal ones from a graphical point of view. We represent the variations of column categories rather than the categories on the space generated by cumulative frequencies. Let define the matrix

$$\mathbf{R} = \mathbf{D}_{IK}^{-1} \tilde{\mathbf{P}} \mathbf{A}^T \mathbf{W}^{\frac{1}{2}} \tag{4}$$

where $\mathbf{W}$ is a diagonal square matrix of dimension $[(J-1)\times(J-1)]$ with general term w_s and $\mathbf{A}$ the following $[(J-1)\times J]$ matrix

$$\mathbf{A} = \begin{bmatrix} 1-\tilde{d}_1 & -\tilde{d}_1 & \dots & -\tilde{d}_1 \\ 1-\tilde{d}_2 & 1-\tilde{d}_2 & \dots & -\tilde{d}_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1-\tilde{d}_{J-1} & 1-\tilde{d}_{J-1} & \dots & -\tilde{d}_{J-1} \end{bmatrix}$$

$\mathbf{A}$ can be also written as $\mathbf{A} = \mathbf{M} - [\mathbf{D}(\mathbf{1}_{J-1}\mathbf{1}_J^T)]$ where $\mathbf{M}$ is unitriangular lower matrix of dimension $(J-1)\times J$, $\mathbf{D} = diag(\tilde{d_s})$, with $\mathbf{1}_{J-1}$ and $\mathbf{1}_J$ column vectors of one of dimension $(J-1)$ and J, respectively. The CSS-type Multiple Taguchi's statistic T^M_{CSS} is then given by

$$T^M_{CSS} = n\times||\mathbf{R}||^2_{\mathbf{D}_{IK}} = n\times||\mathbf{D}^{-1}_{IK}\tilde{\mathbf{P}}\mathbf{A}^T\mathbf{W}^{\frac{1}{2}}||^2_{\mathbf{D}_{IK}} = n\times trace\left(\mathbf{D}^{-\frac{1}{2}}_{IK}\tilde{\mathbf{P}}\mathbf{A}^T\mathbf{W}\mathbf{A}\tilde{\mathbf{P}}^T\mathbf{D}^{-\frac{1}{2}}_{IK}\right)$$

Let $GSVD(\mathbf{R})_{\mathbf{D}_{IK},\mathbf{I}}$ denotes the generalized singular value decomposition of matrix $\mathbf{R} = \{r_{ikj}\}$ of rank M such that $\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$, where $\mathbf{U}$ is an $[I\times M]$ matrix of left singular vectors such that $\mathbf{U}^T\mathbf{D}_I\mathbf{U} = \mathbf{I}_M$, $\mathbf{V}$ is an $[(J-1)\times M]$ matrix of right singular vectors such that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_M$ and $\boldsymbol{\Lambda}$ is a positive definite diagonal matrix of order M of singular values of $\mathbf{R}$ of general term λ_m $(m = 1,\dots,M)$. Total inertia is given by

$$||\mathbf{R}||^2_{D_{IK}} = trace(\mathbf{R}^T\mathbf{D}_{IK}\mathbf{R}) = \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J} p_{ik.}r^2_{ikj} = \sum_{m=1}^{M}\lambda^2_m = \frac{T^M_{CSS}}{n}$$

Finally, row and column standard coordinates for the graphical representation of the association between predictors and response categorical variables are then given by $\mathbf{F} = \mathbf{U}\boldsymbol{\Lambda}$ and $\mathbf{G} = \mathbf{V}\boldsymbol{\Lambda}$, respectively.

According to the Nair's approach [3, 4] we show how the distribution of T^M_{CSS} is approximated using Satterthwaite's method [5]. Let $\boldsymbol{\Gamma}$ be the $(J-1)\times(J-1)$ diagonal matrix of the nonzero singular-values of $\mathbf{A}^T\mathbf{W}\mathbf{A}$ and consider the singular value decomposition $\mathbf{A}^T\mathbf{W}^{\frac{1}{2}} = \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Z}^T$ with $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$ such that $\mathbf{A}^T\mathbf{W}\mathbf{A} = \mathbf{Q}\boldsymbol{\Gamma}^2\mathbf{Q}^T$. The CSS-type Multiple Taguchi's statistic T^M_{CSS} is given by

$$\begin{aligned} T^M_{CSS} &= n\times trace\left(\mathbf{D}^{-\frac{1}{2}}_{IK}\tilde{\mathbf{P}}\mathbf{A}^T\mathbf{W}\mathbf{A}\tilde{\mathbf{P}}^T\mathbf{D}^{-\frac{1}{2}}_{IK}\right) = n\times trace\left(\mathbf{D}^{-\frac{1}{2}}_{IK}\tilde{\mathbf{P}}\mathbf{Q}\boldsymbol{\Gamma}^2\mathbf{Q}^T\tilde{\mathbf{P}}^T\mathbf{D}^{-\frac{1}{2}}_{IK}\right) \\ &= n\times trace\left(\mathbf{S}\boldsymbol{\Gamma}^2\mathbf{S}^T\right) \end{aligned}$$

where $\mathbf{S} = \mathbf{D}_I^{-\frac{1}{2}}\tilde{\mathbf{P}}\mathbf{Q}$. It is possible to show that the i-th elements S_{is} of column vector $\mathbf{S}_s$ are asymptotically iid with a $N(0,1)$ distribution as $n\rightarrow\infty$, with $s = 1,\dots,J-1$ and $i = 1,\dots,(I\times K)-1$. Then, under the hypothesis of homogeneity and given the row and column probabilities, the components $\mathbf{S}_s^T\mathbf{S}_s = \sum_{i=1}^{I\times K}S^2_{is}$ are asymptotically iid with a $\chi^2_{[(I\times K)-1]}$ distribution. Consequently, under the null hypothesis, the limiting distribution of the CSS-type Multiple Taguchi's statistic T^M_{CSS} is a linear combination of chi-squared distributions

$$T^M_{CSS} = n \times trace\left(\mathbf{S}\boldsymbol{\Gamma}^2\mathbf{S}^T\right) \xrightarrow[H_0]{d} \sum_{s=1}^{J-1} \gamma_s \times \chi^2_{[(I\times K)-1]}(s)$$

where $\chi^2_{[(I\times K)-1]}(s)$ is the chi-squared distribution for the s-th component $\mathbf{S}^T_s\mathbf{S}_s$ $(s = 1,\ldots,J-1)$ and γ_s are elements of matrix $\boldsymbol{\Gamma}$. By using Satterthwaite’s two-moment approximation [5], the asymptotic distribution of T^M_{CSS} can be then approximated [3, 4] by $d^M_{CSS} \times \chi^2_{\nu^M_{CSS}}$ with $\nu^M_{CSS} = (d^M_{CSS})^{-1}\sum_{s=1}^{J-1}\gamma_s$ degrees of freedom and $d^M_{CSS} = [(I\times K)-1]^{-1}(\sum_{s=1}^{J-1}\gamma_s^2/\sum_{k=1}^{J-1}\gamma_s)$.

5 Orthogonal decomposition of Multiple Taguchi’s statistic

The Multiple Taguchi’s statistic is a measure of association that contains both main effects and interaction term. The main effects represent the change in the response variables due to the change in the level/categories of the predictor variables, considering the effects of their addition. The interaction effect represents the combined effect of predictor categorical variables on the ordinal response variable.

The interpretation of MTA graphical results can be improved if we highlight the impact of these effects on the association. The Multiple Taguchi’s statistic can be then decomposed in different orthogonal quantities:

$$T^M_{CSS} = T^{A\cup B} + T^{A\times B} = T^A + T^{B|A} + T^{A\times B} = T^{A|B} + T^B + T^{A\times B}$$

where $T^{A\cup B}$ reflects the main effects and represents the change in the response variables due to the change on the levels/categories of the predictor variables considering their joining effects, T^A (or T^B) represents the Taguchi’s statistic calculated between Y and A (or B) after a row aggregation of variable B (or A), while $T^{B|A}$ (or $T^{A|B}$) is the Taguchi’s statistic between Y and B (or A) where the effects of variable A has been partialled out (or B). Finally, $T^{A\times B}$ is the interaction effect and represents the combined effect of predictor variables on the response variable. In particular, there is an interaction between two predictor variables when the effect of one predictor variable varies as the levels/categories of the other predictor vary. If the interaction is not significant, it is possible to examine the main effects. Instead, if the interaction is statistically significant and of strong entity, then, it is not useful to consider the main effects.

In order to separate the main effects and the interaction term, the approach starts from a constraints matrix. Let $\mathbf{T}_{A\cup B} = [\mathbf{T}_A|\mathbf{T}_B]$ be the matrix of dummy variables with $\mathbf{T}_A = (\mathbf{1}_K \otimes \mathbf{I}_I)$ (factor A), $\mathbf{T}_B = (\mathbf{I}_K \otimes \mathbf{1}_I)$ (factor B) and such that formula (4) can be written as $\mathbf{R} = \mathbf{D}^{-1}_{IK}\mathbf{H}^T_{1/D_{IK}}\tilde{\mathbf{P}}\mathbf{M}^T\mathbf{W}^{\frac{1}{2}}$ where $\mathbf{H}_{1/D_{IK}} = \mathbf{I}_{IK} - [\mathbf{1}_{IK}(\mathbf{1}^T_{IK}\mathbf{D}_{IK}\mathbf{1}_{IK})^{-1}\mathbf{1}^T_{IK}\mathbf{D}_{IK}]$ is the orthogonal projector onto the null space of $\mathbf{1}_{IK}$ in metric $\mathbf{D}_{IK}$ with $\mathbf{1}_{IK}$ unitary column vectors of dimension IK. $\mathbf{H}_{1/D_{IK}}$ eliminates the row marginal effect from the relationship between rows and columns. The main effects are given by

$$\mathbf{R}_{A\cup B} = \mathbf{H}_{1/D_{IK}}\mathbf{T}_{A\cup B}(\mathbf{T}_{A\cup B}^{T}\mathbf{H}_{1/D_{IK}}^{T}\mathbf{D}_{IK}\mathbf{T}_{A\cup B})^{-1}\mathbf{T}_{A\cup B}^{T}\mathbf{H}_{1/D_{IK}}^{T}\tilde{\mathbf{P}}\mathbf{M}^{T}\mathbf{W}^{\frac{1}{2}}$$

Since $\mathbf{R}_{A\times B} = \mathbf{R} - \mathbf{R}_{A\cup B}$ then we obtain the following norm decomposition

$$||\mathbf{R}||_{D_{IK}}^{2} = ||\mathbf{R}_{A\cup B}||_{D_{IK}}^{2} + ||\mathbf{R}_{A\times B}||_{D_{IK}}^{2} \tag{5}$$

Similarly, a double decomposition of the main effects in orthogonal quantities is

$$||\mathbf{R}_{A\cup B}||_{D_{IK}}^{2} = ||\mathbf{R}_{A}||_{D_{IK}}^{2} + ||\mathbf{R}_{B|A}||_{D_{IK}}^{2} = ||\mathbf{R}_{B}||_{D_{IK}}^{2} + ||\mathbf{R}_{A|B}||_{D_{IK}}^{2} \tag{6}$$

where $\mathbf{R}_{A} = \mathbf{H}_{1/D_{IK}}\mathbf{T}_{A}(\mathbf{T}_{A}^{T}\mathbf{H}_{1/D_{IK}}^{T}\mathbf{D}_{IK}\mathbf{T}_{A})^{-1}\mathbf{T}_{A}^{T}\mathbf{H}_{1/D_{IK}}^{T}\mathbf{P}\mathbf{M}^{T}\mathbf{W}^{\frac{1}{2}}$ and $\mathbf{R}_{B} = \mathbf{H}_{1/D_{IK}}\mathbf{T}_{B}$ $(\mathbf{T}_{B}^{T}\mathbf{H}_{1/D_{IK}}^{T}\mathbf{D}_{IK}\mathbf{T}_{B})^{-1}\mathbf{T}_{B}^{T}\mathbf{H}_{1/D_{IK}}^{T}\mathbf{P}\mathbf{M}^{T}\mathbf{W}^{\frac{1}{2}}$. Decomposition (6) shows that $A\cup B \neq (A+B)$ because $A\cup B = (A+B|A) = (B+A|B)$ since A and B are not orthogonal factors [9]. If we consider a balanced design then we have $R_{A|B} = R_{A}$ and $R_{B|A} = R_{B}$ so that we can write $||\mathbf{R}_{A\cup B}||_{D_{IK}}^{2} = ||\mathbf{R}_{A}||_{D_{IK}}^{2} + ||\mathbf{R}_{B}||_{D_{IK}}^{2}$ and decomposition (5) is now $||\mathbf{R}||_{D_{IK}}^{2} = ||\mathbf{R}_{A}||_{D_{IK}}^{2} + ||\mathbf{R}_{B}||_{D_{IK}}^{2} + ||\mathbf{R}_{A\times B}||_{D_{IK}}$.

Table 1 Multiple Taguchi's statistic decomposition

Decomposition	Index	$\tilde{d}$	Statistic	degrees of freedom
Main effects	$T^{A\cup B}$	$\tilde{d}^{A\cup B} = \left(\frac{1}{\tilde{d}^{A}} + \frac{1}{\tilde{d}^{B\|A}}\right)^{-1}$	$T^{A\cup B}/\tilde{d}^{A\cup B}$	$v^{A\cup B} = \frac{1}{\tilde{d}^{A\cup B}}\sum_{s=1}^{J-1}\gamma_s$
Interaction	$T^{A\times B}$	$\tilde{d}^{A\times B} = \left(\frac{1}{\tilde{d}^{M}} - \frac{1}{\tilde{d}^{A}} - \frac{1}{\tilde{d}^{B\|A}}\right)^{-1}$	$T^{A\times B}/\tilde{d}^{A\times B}$	$v^{A\times B} = \frac{1}{\tilde{d}^{A\times B}}\sum_{s=1}^{J-1}\gamma_s$
Total	T_{CSS}^{M}	$\tilde{d}_{CSS}^{M} = \frac{1}{[(I\times K)-1]}\frac{\sum_{s=1}^{J-1}\gamma_s^2}{\sum_{s=1}^{J-1}\gamma_s}$	$T_{CSS}^{M}/\tilde{d}_{CSS}^{M}$	$v_{CSS}^{M} = \frac{1}{\tilde{d}_{CSS}^{M}}\sum_{s=1}^{J-1}\gamma_s$

Table 2 Alternative $T^{A\cup B}$ decompositions

Decomposition	Index	$\tilde{d}$	Statistic	degrees of freedom
Factor A	T^{A}	$\tilde{d}^{A} = \frac{1}{(I-1)}\frac{\sum_{s=1}^{J-1}\gamma_s^2}{\sum_{s=1}^{J-1}\gamma_s}$	$T^{A}/\tilde{d}^{A}$	$v^{A} = \frac{1}{\tilde{d}^{A}}\sum_{s=1}^{J-1}\gamma_s$
Factor A|B	$T^{A\|B}$	$\tilde{d}^{A\|B} = \frac{1}{(I-1)}\frac{\sum_{s=1}^{J-1}\gamma_s^2}{\sum_{s=1}^{J-1}\gamma_s}$	$T^{A\|B}/\tilde{d}^{A\|B}$	$v^{A\|B} = \frac{1}{\tilde{d}^{A\|B}}\sum_{s=1}^{J-1}\gamma_s$
Factor B	T^{B}	$\tilde{d}^{B} = \frac{1}{(K-1)}\frac{\sum_{s=1}^{J-1}\gamma_s^2}{\sum_{s=1}^{J-1}\gamma_s}$	$T^{B}/\tilde{d}^{B}$	$v^{B} = \frac{1}{\tilde{d}^{B}}\sum_{s=1}^{J-1}\gamma_s$
Factor B|A	$T^{B\|A}$	$\tilde{d}^{B\|A} = \frac{1}{(K-1)}\frac{\sum_{s=1}^{J-1}\gamma_s^2}{\sum_{s=1}^{J-1}\gamma_s}$	$T^{B\|A}/\tilde{d}^{B\|A}$	$v^{B\|A} = \frac{1}{\tilde{d}^{B\|A}}\sum_{s=1}^{J-1}\gamma_s$

Let $\mathbf{R}_{(A\cup B)} = \mathbf{U}_{A\cup B}\boldsymbol{\Lambda}_{A\cup B}\mathbf{V}_{A\cup B}^{T}$ denote $GSVD(\mathbf{R}_{A\cup B})_{D_{IK},I}$ where $\mathbf{U}_{A\cup B}$ is an $(I\times M)$ matrix of right singular vectors such that $\mathbf{U}_{A\cup B}^{T}\mathbf{D}_{IK}\mathbf{U}_{A\cup B} = \mathbf{I}_{M}$ $[M = rank(\mathbf{R}_{A\cup B})]$,

$\mathbf{V}_{A\cup B}$ is $[(J-1)\times M]$ matrix of right singular vectors such that $\mathbf{V}_{A\cup B}^{T}\mathbf{V}_{A\cup B}=\mathbf{I}_M$ and $\boldsymbol{\Lambda}$ is a positive definite diagonal matrix of order M of singular value of $\mathbf{R}_{A\cup B}$ with general term $\lambda_m^{A\cup B}$ with $m=1,\ldots,M$. Row and column standard coordinates of the main effects are given by $\mathbf{F}_{A\cup B}=\mathbf{U}_{A\cup B}\boldsymbol{\Lambda}_{A\cup B}$ and $\mathbf{G}_{A\cup B}=\mathbf{V}_{A\cup B}\boldsymbol{\Lambda}_{A\cup B}$, respectively. It’s also possible to plot the interaction term and the single effects in the same way.

6 Conclusion

Taguchi introduced his statistic as simple alternative to Pearson’s chi-squared test for two way contingency tables. Actually, χ^2 does not perform well when we have a contingency table cross-classifying at least one ordinal categorical variable. In this paper an extension of this statistic for three way contingency tables handled in two way mode has been introduced highlighting some properties. The approximated distribution of the CCS-type Multiple Taguchi’s statistic T_{CSS}^{M}, by using Satterthwaite’s method, has been also suggested. In this paper, an extension of Correspondence Analysis based on the decomposition of the CCS-type Multiple Taguchi’s statistic has been moreover proposed. The interpretation of the graphical results has been improved highlighting the impact of the main effects and the interaction terms on the association. This is obtained with a decomposition of statistic T_{CSS}^{M} according to orthogonal quantities reflecting several effects.
Finally, an extended version of this paper will include an application on real data about service quality evaluation. All the theoretical results will be used showing also the graphical outputs. We will be also able to evaluate the impact of the main effects and interaction term on association among the categorical variables.

References

1. Agresti A.: Categorical Data Analysis, third edition. John Wiley & Sons, (2013)
2. D’Ambra, L., Amenta, P., D’Ambra, A.: Decomposition of cumulative chi-squared statistics, with some new tools for their interpretation, Statistical Methods and Applications, doi 10.1007/s10260-017-0401-3, (2017)
3. Nair, V.N.: Testing in industrial experiments with ordered categorical data. Technometrics **28**(4), 283–291 (1986)
4. Nair, V.N.: Chi-squared type tests for ordered alternatives in contingency tables. Journal of American Statistical Association **82**, 283–291 (1987)
5. Satterthwaite, F.: An approximate distribution of estimates of variance components. Biometrical Bullettin (2), 110–114 (1946)
6. Taguchi, G.: Statistical Analysis. Tokyo: Maruzen (1966)
7. Taguchi, G.: A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. Saishin Igaku **29**, 806–813 (1974)
8. Takeuchi, K., Hirotsu, C.: The cumulative chi square method against ordered alternative in two-way contingency tables. Tech. Rep. 29, Reports of Statistical Application Research. Japanese Union of Scientists and Engineers (1982)
9. Takeuchi, K., Yanai, H., Mukherjee, B.N.: The Foundations of Multivariate Analysis. John Wiley & Sons (Asia) Pte Ltd., (1982)

A robust multinomial logit model for evaluating judges' performances

Un modello multinomiale robusto per valutare la performance dei giudici.

Ida Camminatiello and Antonio Lucadamo

Abstract Principal component multinomial regression is a method for modelling the relationship between a set of high-dimensional regressors and a categorical response variable with more than two categories. This method uses as covariates of the multinomial model a reduced number of principal components of the regressors. Because the principal components are based on the eigenvectors of the empirical covariance matrix, they are very sensitive to anomalous observations. Several methods for robust principal component analysis have been proposed in literature. In this study we consider ROBPCA method. The new robust approach will be applied for assessing judges' performances.

Abstract *La regressione multinomiale sulle componenti principali è un metodo per modellare la relazione tra un set di regressori ad alta dimensionalità e una variabile di risposta nominale con più di due modalità. Questo metodo usa come covariate del modello multinomiale un numero ridotto di componenti principali estratte dai regressori. Poiché le componenti principali si basano sugli autovettori della matrice di covarianza empirica, sono molto sensibili alle osservazioni anomale. Diversi metodi robusti per l'analisi in componenti principali sono stati proposti in letteratura. In questo studio consideriamo il metodo ROBPCA. Il nuovo approccio robusto sarà applicato per valutare la performance dei giudici.*

Key words: principal component analysis, multinomial logit model, outliers, judges' performances.

[1] Ida Camminatiello, University of Campania; email: ida.camminatiello@unicampania.it
Antonio Lucadamo, University of Sannio; email: antonio.lucadamo@unisannio.it

1. Introduction

The court computerization of the last decades allows us to create available databases with complete information about the judicial flows. Here, we aim to focus on the causes of different judges' performances in the court of Naples.

The dataset shows strongly correlated regressors, so the most proper statistic methodology to analyse this kind of data could be Principal component multinomial regression (Camminatiello, Lucadamo, 2010; Lucadamo, Leone, 2015).

A previous research on Florence Court (Camminatiello, Lombardo, Durand, 2017) highlighted the presence of outliers among judges.

The aim of the paper is to study the dependence relationship between the judges' performances and some indicators of the judges' workload taking into account multicollinearity and outlier problems which make the estimation of the multinomial model parameters inaccurate because of the need to invert nearsingular and ill-conditioned information matrices.

A robust method for logistic regression (Rousseeuw, Christmann, 2003) and robust logistic ridge regression (Ariffin, Midi, 2014) have been proposed in literature, we propose a robust approach for the principal component multinomial regression (PCMR).

We proceed in the following way. In the second section we describe the PCMR. In the third section we list the most important robust methods for estimating the variance/covariance matrix and propose a robust approach to PCMR. In the fourth section we apply our robust approach for evaluating the judges' efficiency and calculate the correct classification rate for comparing three different models.

2. From multinomial logit regression to robust methods for principal component multinomial regression.

Multinomial logit model (MNL) is the simplest model in discrete choice analysis when more than two alternatives are in a choice set. The model becomes unstable when there is multicollinearity among predictors (Ryan, 1997). To improve the estimation of the MNL parameters, Camminatiello and Lucadamo (2010) proposed the PCMR.

PCMR uses as covariates of the multinomial model a reduced number of principal components (PCs) of the regressors. Because these components are based on the eigenvectors of the empirical covariance matrix, they are very sensitive to anomalous observations. Several methods for robustifying principal component analysis (PCA) have been proposed in literature.

If the number of observations is sufficiently large with respect to the number of variables, the classical covariance matrix can be replaced by minimum covariance determinant (MCD) estimator, minimum volume ellipsoid (MVE) estimator (Rousseeuw and Leroy 1987), S-estimators (Davies 1987), reweighted MCD

(Rousseeuw and van Zomeren, 1990), FAST-MCD (Rousseeuw and Van Driessen, 1999).

For high-dimensional data, a ROBust method for PCA, called ROBPCA, has recently been developed (Hubert, Rousseeuw and Vanden Branden, 2012). ROBPCA starts by reducing the data space to the affine subspace spanned by *n* observations. A convenient way to perform it is by a singular value decomposition of the mean-centred data matrix. In the second stage $h < n$ "least outlying" data points are found by Stahel-Donoho affine invariant outlyingness. In the third stage, the algorithm robustly estimates the location and scatter matrix of the data points projected into a subspace of small to moderate dimension by using the FAST-MCD estimator. ROBPCA ends by yielding the robust principal components. Like classical PCA, the ROBPCA method is location and orthogonal equivariant.

2.1. Robust Principal Component Multinomial Regression - RobPCMR

Several authors applied ROBPCA to formulate other robust techniques (Hubert, Vanden Branden, 2003; Hubert, Verboven, 2003; Rousseeuw, Christmann, 2003). We investigate using ROBPCA before PCMR to deal with multicollinearity and outlier problems in MNL. We proceed in the following way.

Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_\mathrm{i}, \dots, \mathbf{x}_\mathrm{p}]$ be a set of *p* quantitative regressors and $\mathbf{y}$ a categorical response variable with more than two categories observed on *n* statistical units.

At first step, robust principal component multinomial regression (RobPCMR) creates the robust PCs of the regressors $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_\mathrm{i}, \dots, \mathbf{z}_\mathrm{p}]$ which are linear combinations of the original variables $\mathbf{Z} = \mathbf{XV}$. At second step the multinomial model is carried out on the set of robust PCs. At third step, the number of robust PCs, $\mathrm{a} < p$, to be retained in the model, is selected according to different tools (Camminatiello and Lucadamo, 2010).

At fourth step, the multinomial model is carried out on the subset of robust PCs chosen. The probability, for the individual *i*, to choose the alternative *c* can be expressed in terms of *a* robust PCs as:

$$\pi_i^{(a)}(c) = \frac{\exp\left\{\sum_{j=1}^{p} \sum_{k=1}^{a} z_{ik} v_{kj} \beta_{jc}^{(a)}\right\}}{\left\{\sum_{b=1}^{s} \exp \sum_{j=1}^{p} \sum_{k=1}^{a} z_{ik} v_{kj} \beta_{jb}^{(a)}\right\}} = \frac{\exp\left\{\sum_{k=1}^{a} z_{ik} \gamma_{kc}^{(a)}\right\}}{\left\{\sum_{b=1}^{s} \exp \sum_{k=1}^{a} z_{ik} \gamma_{kb}^{(a)}\right\}} \qquad (1)$$

where $\gamma_{kb}^{(a)} = \sum_{j=1}^{p} v_{kj} \beta_{jb}^{(a)}$ are the robust coefficients to be estimated on the subset of *a* robust PCs and $\beta_{jb}^{(a)}$ are the robust PCMR parameters obtained after the extraction of the *a* components.

Finally, the robust MNL parameters can be expressed in function of original variables (**X** matrix)

$$\mathbf{Z}^{(a)}\boldsymbol{\gamma}^{(a)} = \mathbf{X}\mathbf{V}^{(a)}\boldsymbol{\gamma}^{(a)} = \mathbf{X}\boldsymbol{\beta}^{(a)} \tag{2}$$

where $\boldsymbol{\beta}^{(a)} = \mathbf{V}^{(a)}\boldsymbol{\gamma}^{(a)}$ is the matrix of robust parameters expressed in function of original variables; $\boldsymbol{Z}^{(a)}$ is the matrix of robust PCs; $\boldsymbol{\gamma}^{(a)}$ is the matrix of robust parameters on a robust PCs for the s alternatives; $\mathbf{V}^{(a)}$ is the matrix of robust eigenvectors.

To measure the performance of a method, several criteria can be utilised (Camminatiello, Lombardo, Durand, 2017; Camminatiello, Lucadamo, 2010). Here, we focus on rate of well classified which we expect higher compared to PCMR and MNL.

3. A robust model to predict judges' performances.

Our study concerns the causes of different judges' performances in the court of Naples. The performance evaluation is based, among others, on the time that each judge employs to solve the disputes. According to many available publications about similar problems (Camminatiello, Lombardo, Durand, 2017), we aim to study if number of: pendings, hearings, dossiers, incoming and defined proceedings can influence the judges' performances. The response variable is on categorical scale, with four modalities from 1 (Low) to 4 (High), measured for 136 judges.

To evaluate how judges' performances can be influenced by explicative variables, we divide our sample in two sub-samples. The first one, composed by the 70% of the observations, is the sample used to estimate the model parameters (estimation sample). The second one (validation sample) is considered to test the goodness of the obtained estimates. In both the cases we calculate the rate of well classified judges and we compare the results obtained by applying the MNL, the PCMR and the RobPCMR.

In table 1 the results obtained by the three methods on the estimation sample are shown. From the second to the fifth column we have the percentage of well classified, calculated for four different models: the MNL estimated on all the regressors and on the two significant ones (via stepwise regression); the PCMR run on the first PC (which accounts for 93.4% of the variance, has the eigenvalue higher than one and furthermore is the only significant regressor); the RobPCMR carried out on the first robust PC (which accounts for 89.8 % of the variance, has the eigenvalue higher than one and is the only significant regressor).

It is easy to observe that for the estimation sample the classical MNL performs better than other models: it could be due to an over-fitting problem.

Table 1: Percentage of correct classified calculated for four different models on the estimation sample

	MNL (all)	*MNL (sign)*	*PCMR (1)*	*RobPCMR (1)*
% correct classified	70.1%	63.9%	59.8%	58.8%

To verify the goodness of the techniques it is then necessary to consider the results on the validation sample (table 2).

In this case it is evident that for all the methods, but for the RobPCMR, the percentages are lower than before.

Table 2: Percentage of correct classified calculated for four different models on the validation sample

	MNL (all)	*MNL (sign)*	*PCMR (1)*	*RobPCMR (1)*
% correct classified	53.8%	56.4%	53.8%	59.0%

In fact, looking at the table 2, we can notice that the MNL, considering all the variables, leads to the same classification rate obtained by the PCMR with only one component, while for the MNL, taking into account only the two significant variables, the percentage of well classified increases. It is surprising that RobPCMR result shows a rate of correct classification higher than before (59.0% against 58.8%). This may indicate the ability of the method in parameter estimation.

It is also interesting to notice what happens when we consider more components in the analysis, both for PCMR and for RobPCMR.

For this reason we show in table 3 and 4 the results obtained when we consider a different number of components as explicative variables.

Table 3: Percentage of correct classified, at varying the number of components, for the estimation sample

Number of components	*PCMR*	*RobPCMR*
1	59.8%	58.8%
2	60.8%	60.8%
3	60.8%	63.9%
4	61.9%	63.9%
5	70.1%	62.9%

For the PCMR, the results on the estimation sample show that, when the number of components increases, the classification improves. Considering all the components the result is equal to that obtained using all the variables in the classical MNL (Camminatiello and Lucadamo, 2010).

For RobPCMR instead, there is an improving in the correct classification rate at beginning, but when we consider all the components, the result is lower than one obtained with 3 and 4 components.

If we consider the validation sample the results confirm both for the PCMR and for the RobPCMR that the selection of the significant components (in this case explaining the most part of the variability too) is an useful solution to obtain a good rate of correct classification.

Table 4: Percentage of correct classified, at varying the number of components, for the validation sample

Number of components	*PCMR*	*RobPCMR*
1	53.8%	59.0%
2	53.8%	59.0%
3	51.3%	56.4%
4	48.7%	58.9%
5	51.3%	56.4%

Obviously for RobPCMR, as already done in previous studies for PCMR, a complete simulation study is necessary to generalize the results.

4. Conclusion and perspective

In this paper we carried out a robust model for evaluating the judges performances in presence of outliers and strongly correlated covariates. .To solve these problems, we proposed to use as covariates of the multinomial model a reduced number of robust PCs of the predictor variables.

The application showed that the proposed approach is a valid alternative on real data. However, an extensive simulation study is needed in order to verify that it is resistant towards many types of contamination, to compare the results with other robust approaches for PCA proposed in literature and to select optimal dimension of the model. The procedure should lead to lower variance estimates of model parameters comparing to PCMR. The variance can be estimated by bootstrap resampling.

Finally, an extension to MNL of the approaches proposed in literature, for dealing with multicollinear and outlier problems in the logit model (Ariffin, Midi, 2014) could be interesting as well as an extension to ordinal logit regression of the approach here proposed.

5. References

1 Ariffin, S.B., Midi H.: Robust Logistic Ridge Regression Estimator in the Presence of High Leverage Multicollinear Observations. In: 16th Int. Conf. Math. Comput. Methods Sci. Eng. pp 179-184 (2014)

2 Camminatiello, I., Lucadamo, A.: Estimating multinomial logit model with multicollinear data. Asian Journal of Mathematics and Statistics 3 (2), 93-101 (2010)

3 Camminatiello, I., Lombardo, R., Durand, J.F.: Robust partial least squares regression for the evaluation of justice court delay. Qual Quant 51 (2), 813-27 (2017). https://doi.org/10.1007/s11135-016-0441-z

4 Davies, L.: Asymptotic Behavior of S-Estimators of Multivariate Location and Dispersion Matrices. The Annals of Statistics 15, 1269-1292 (1987).

5 Hubert, M., Rousseeuw, P.J., Vanden Branden, K.: ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics 47 (1), 64-79 (2012) doi: 10.1198/004017004000000563

6 Hubert, M., Vanden Branden, K.: Robust Methods for Partial Least Squares Regression. *J Chemometr* 17, 537-549 (2003)

7 Hubert, M., Verboven, S.: A Robust PCR Method for High-Dimensional Regressors. *J Chemometr* 17, 438-452 (2003)

8 Lucadamo, A, Leone, A.: Principal component multinomial regression and spectrometry to predict soil texture. *J Chemometr.,* **29** (9), 514-520 (2015).

9 Rousseeuw, P. J., Christmann, A.: Robustness Against Separation and Outliers in Logistic Regression. Computational Statistics and Data Analysis 43, 315-332.

10 Rousseeuw, P.J. Leroy, A.M.: Robust regression and Outlier Detection. Wiley, New York (1987).

11 Rousseeuw, P.J., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212-223 (1999)

12 Rousseeuw, P. J., Van Zomeren, B. C.: Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association 85, 633-651 (1990).

13 Ryan, T.P.: Modern Regression Methods. Wiley, New York (1997)

Complex Contingency Tables and Partitioning of Three-way Association Indices for Assessing Justice Court Workload

Tabelle di Contingenza Complesse e Partizioni di Indici di Associazione per la Valutazione del Carico di Lavoro nei Tribunali Giudiziari

Rosaria Lombardo, Yoshio Takane and Eric J Beh

Abstract A comprehensive study is conducted on the partition of two common indices for three-way contingency tables under several representative hypotheses about the expected frequencies (hypothesized probabilities). Specifically, the partition of the classical (symmetrical) three-way Pearson index and of the asymmetrical three-way Marcotorchino index are considered under a general *Scenario 0* from which known different scenarios are derived: 1) where the hypothesized probabilities are homogeneous among the categories [12], and 2) when the hypothesized probabilities are estimated from the data [7, 6].

Abstract *In questo lavoro si presenta uno studio completo sulla partizione di due indici di associazione per le tabelle di contingenza a tre vie, in base a diverse ipotesi sulle frequenze attese (probabilitá ipotizzate). Nello specifico, si propone la partizione dell'indice di Pearson (simmetrico) e dell'indice di Marcotorchino a tre vie (asimmetrico) in uno Scenario 0 da cui derivano note partizioni: 1) dove le probabilitá ipotizzate sono uniformi [12] e 2) quando le probabilitá ipotizzate sono stimate dai dati [7, 6].*

Key words: Three-way association indices, Hypothesized probabilities, Observed frequencies, Chi-Squared distribution

R. Lombardo
Economics Department, University of Campania, via Gran Priorarato di Malta, Capua (CE), Italy, Tel.: +390810601382, Fax: +390823622984, e-mail: rosaria.lombardo@unina2.it

Y. Takane
University of Victoria, 5173 Del Monte Ave, Victoria BC, V8Y 1X3, Canada e-mail: yoshio.takane@mcgill.ca

E.J. Beh
School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, 2308, NSW, Australia e-mail: eric.beh@newcastle.edu.au

1 Introduction

Measuring and assessing the association among categorical variables can be undertaken by partitioning a multi-way association index into bivariate, trivariate and higher order terms. The partitions presented in this paper are based on the work by Lancaster (1951) which considered an ANOVA-like partitions of Pearson's chi-squared statistic implemented for the analysis of a $2 \times 2 \times 2$ table. Here, we focus on Pearson's statistic [10] and on Marcotorchino's index [15] for studying the symmetrical and asymmetrical association among three categorical variables, respectively. In this paper we show that, under complete independence of the three categorical variables, a general scenario - called *Scenario 0* - is defined from which different known partitions of Pearson's statistic and Marcotorchino's index can be derived as special cases. Examples of these known partitions include those of [7, 6], [14], [3, 4] and [12]. In *Scenario 0*, the probabilities are prescribed by the analyst instead of being estimated by using the margins of the empirical distribution that underly the data. The reason is that *a priori* knowledge of phenomena can suggest differently; see [1, 12].

The paper is organised in the following manner. After introducing the notation, a general partition of a three-way array in $\Re^{I\times J\times K}$ is discussed in Section 2. The key scenarios of model dependence are described in Section 3. Section 4 presents the general partition for Pearson's statistic. A practical demonstration of this partition under *Scenario 0* is given in Section 5. Some concluding remarks are made in Section 6.

2 Partitioning three-way association indices

Consider a general three-way contingency table $\underline{\mathbf{X}} = \left(x_{ijk}\right) \in \Re^{I\times J\times K}$ (for $i = 1,...,I$, $j = 1,...,J$, and $k = 1,...,K$) that summarises the cross-classification of n individuals/units according to I row, J column and K tube categories. These sets of categories form the row, column and tube variables x_I,x_J, and x_K, respectively. Let $\underline{\hat{\mathbf{P}}} = (\hat{p}_{ijk})$ be the array of the observed joint relative frequencies (of dimension $I \times J \times K$), so that $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}/n = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{p}_{ijk} = 1$.

Let $\underline{\mathbf{P}} = (p_{ijk})$ be the array of the joint probability values of the cells in the three-way tables. Define $\mathbf{p}_I\{= p_{i\bullet\bullet}\}$, $\mathbf{p}_J\{= p_{\bullet j\bullet}\}$, $\mathbf{p}_K\{= p_{\bullet\bullet k}\}$ to be the vectors of the marginal probabilities associated with the three variables x_I,x_J, and x_K, respectively.

In case complete three-way independence is hypothesized, it holds that $p_{ijk} = p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}$, where p_{ijk} is the $(i,\ j,\ k)$th joint probability and $p_{i\bullet\bullet}$, $p_{\bullet j\bullet}$ and $p_{\bullet\bullet k}$ are the marginal probabilities of the ith row category, the jth column category, and the kth tube category, respectively.

When studying the association among variables that are symmetrically associated, Pearson's chi-squared statistic is always appropriate. Instead, when variables are asymmetrically associated (for example, x_I may be considered as the response variable while x_J and x_K are treated as the predictor variables) then Marcotorchino's index [15, 14, 4] may be considered more suitable.

Let $\underline{\boldsymbol{\Pi}}$ be a three-way array whose general element is $\pi_{ijk} = \left(\frac{\hat{p}_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} - 1\right)$ and define $\underline{\boldsymbol{\Pi}}_M = \left(\frac{\hat{p}_{ijk}}{p_{\bullet j\bullet}p_{\bullet\bullet k}} - p_{i\bullet\bullet}\right)$. Let $\underline{\mathbf{M}} = \mathbf{D_I} \otimes \mathbf{D_J} \otimes \mathbf{D_K}$ and $\underline{\mathbf{M}}_{tau} = \mathbf{I} \otimes \mathbf{D_J} \otimes \mathbf{D_K}$, be the metric related to the arrays $\underline{\boldsymbol{\Pi}}$ and $\underline{\boldsymbol{\Pi}}_M$, respetively. Given the traditional definition of inner products and quadratic norm in the space $\Re^{I\times J\times K}$ [16, p. 7],[6], observe that the quadratic norm of $\underline{\boldsymbol{\Pi}}$ with metric $\underline{\mathbf{M}}$ represents Pearson's mean square statistic and can be expressed as

$$\frac{\chi^2}{n} = \Phi^2 = \| \underline{\boldsymbol{\Pi}} \|^2_M = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\pi^2_{ijk}, \tag{1}$$

which measures the overall discrepancy between a set of observed frequencies and the expected frequencies.

Similarly, the quadratic norm of $\underline{\boldsymbol{\Pi}}_M$ with metric $\underline{\mathbf{M}}_{tau}$ represents Marcotorchino's index numerator and can be expressed as

$$\tau_M^{num} = \| \underline{\boldsymbol{\Pi}} \|^2_{M_{tau}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}(\pi^M_{ijk})^2, \tag{2}$$

which measures the discrepancy between a set of conditional frequencies and the expected frequencies. The denominator of this index is considered a constant term, since it does not depend on the predictor variables, and is usually neglected. Marcotorchino's index, called τ_M, is a generalization of the τ index presented by [8] for studying the predictability issue in two-way contingency tables. For more information on this index, refer to [15, 14] and [4, p.461].

For partitioning purpose, examine the orthogonal projections of a general three-way array $\underline{\mathbf{X}}$ beloging to the space $\Re^{I\times J\times K}$ onto the subspaces $\Re^0$, $\Re^I$, $\Re^J$, $\Re^K$, $\Re^{JK}$, $\Re^{IJ}$, $\Re^{IK}$ and $\Re^{IJK}$. Then, according to the ANOVA-like decomposition of the elements of an array we get

$$x_{ijk} = a + b_i + c_j + d_k + e_{ij} + f_{ik} + g_{jk} + h_{ijk}, \tag{3}$$

so that there exists a fixed main term ($a = x_{\bullet\bullet\bullet}$), three univariate main terms ($b_i = x_{i\bullet\bullet} - x_{\bullet\bullet\bullet}$, $c_j = x_{\bullet j\bullet} - x_{\bullet\bullet\bullet}$ and $d_k = x_{\bullet\bullet k} - x_{\bullet\bullet\bullet}$), three bivariate terms ($e_{ij} = x_{ij\bullet} - x_{\bullet\bullet\bullet}$, $f_{ik} = x_{i\bullet k} - x_{\bullet\bullet\bullet}$ and $g_{jk} = x_{\bullet jk} - x_{\bullet\bullet\bullet}$) and a trivariate effect ($h_{ijk} = x_{ijk} - x_{ij\bullet} - x_{i\bullet k} - x_{\bullet jk} + x_{i\bullet\bullet} + x_{\bullet j\bullet} + x_{\bullet\bullet k} - x_{\bullet\bullet\bullet}$), defined onto the sub-spaces $\Re^0$, $\Re^I$, $\Re^J$, $\Re^K$, $\Re^{IJ}$, $\Re^{IK}$, $\Re^{JK}$ and $\Re^{IJK}$, respectively. The determination of the partition terms depend upon the definition of the array $\underline{\mathbf{X}}$ and on the metric related to each subspace, as it will be illustrated in Section 2.1

2.1 Index's partition term

In general, the uniqueness of each term of the partition in Equation 3, for example say b_i, is verified if the following two conditions are satisfied

1. each term belongs to the related space: $b_i \in \Re^I$
2. each term is orthogonal to the space vectors:

$$(\mathbf{x} - \mathbf{b}_M) \perp \Re^I \leftrightarrow < \mathbf{x} - \mathbf{b}, \beta >_M = 0 \qquad \forall \beta \in \Re^I$$

as consequence

$$< \mathbf{x}, \beta >_M = < \mathbf{b}, \beta >_M$$

By setting $\boldsymbol{x} = \boldsymbol{\pi}$, using the metric $\underline{\mathbf{M}}$ and expanding the inner product, we get

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\pi_{ijk}\beta_i &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}b_i\beta_i \\ \sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\pi_{ijk} &= b_i \\ \pi_{i\bullet\bullet} &= b_i\,.\end{aligned}$$

So the projection of $\underline{\boldsymbol{\Pi}}$ onto the subspace $\Re^I$ is defined by

$$\begin{aligned} b_i &= \pi_{i\bullet\bullet} - \pi_{\bullet\bullet\bullet} \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K} p_{\bullet j\bullet}p_{\bullet\bullet k}\left(\frac{\hat{p}_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} - 1\right) \\ &= \left(\frac{\hat{p}_{i\bullet\bullet}}{p_{i\bullet\bullet}} - 1\right)\end{aligned}$$

The weighted quadratic norm of this term is equal to the first term on the right-hand side of Equation 4.
The orthogonal projections of $\underline{\boldsymbol{\Pi}}$ onto the remaining subspaces can be similarly defined.
Furthermore, changing the metric in $\underline{\mathbf{M}}_{tau} = \boldsymbol{I} \otimes D_J \otimes D_K$ and taking into account the array $\underline{\boldsymbol{\Pi}}_M$, we can get the orthogonal projections of Marcotorchino's index in a similar way.

3 The key scenarios of model dependence

Under the three-way model of complete independence, we present a new general scenario - referred to as *Scenario 0* - which can be applied to each variable of a three- or multi-way contingency table rather than to contingency tables as a whole (see Section 4). In *Scenario 0*, the marginal probabilities, $\mathbf{p}_I$, $\mathbf{p}_J$, and $\mathbf{p}_K$, can take on any prescribed values that satisfy the classical probability laws (e.g., $p_{i\bullet\bullet} \geq 0$, and $\sum_{i=1}^{I} p_{i\bullet\bullet} = 1$). We do not assume that the probabilities of $\mathbf{p}_I$, $\mathbf{p}_J$, and $\mathbf{p}_K$ are user-defined but are dictated by situational demand.

- *Scenario 1*. A special case of *Scenario 0* where it is hyphotesised the marginal homogeneity under independence such that $\mathbf{p}_I = \mathbf{1}_I/I, \mathbf{p}_J = \mathbf{1}_J/J$, and $\mathbf{p}_K = \mathbf{1}_K/K$, where $\mathbf{1}_I, \mathbf{1}_J$ and $\mathbf{1}_K$ are unitary vector of length I, J and K, respectively. In this case the degree of freedom, *df*, for the three-way chi-squared statistic is $IJK - 1$. This scenario is at the core of the chi-squared partition proposed by Loisel and Takane by using orthogonal transformations of variables [12].
- *Scenario 2*. A special case of *Scenario 0* in which the probabilities in $\mathbf{p}_I$, $\mathbf{p}_J$, and $\mathbf{p}_K$ are estimated from the data, so that $\mathbf{p}_I = \hat{\mathbf{p}}_I$, $\mathbf{p}_J = \hat{\mathbf{p}}_J$ and $\mathbf{p}_K = \hat{\mathbf{p}}_K$, are prescribed to be equal to the observed marginal proportions of the three categorical variables. In this scenario, the *df* for the three-way chi-squared statistic is $(IJK-1)-(I-1)-(J-1)-(K-1)$
- *Scenario 3*. A special case of *Scenario 0* in which the specification of the probabilities in $\mathbf{p}_I$, $\mathbf{p}_J$, and $\mathbf{p}_K$ can be a mix of both *Scenario 1* and *Scenario 2*.

For all scenarios, the partition terms of Pearson's mean square coefficient and Marcotorchino's index in the space $\Re^{I\times J\times K}$ are likely to be different. However, irrespective of which special case of *Scenario 0* is considered, these statistics and each term of their partition, can be tested (asymptotically) for statistical significance using a χ^2 distribution.

4 Partitioning Pearson's statistic

For a sake of brevity, here we illustrate only the partition of Φ^2 under *Scenario 0*. The row, column and tube marginal probabilities, $\mathbf{p}_I$, $\mathbf{p}_J$, and $\mathbf{p}_K$, respectively, can be *a priori* known or estimated from the data. This general partition is

$$\begin{aligned}
\Phi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{1}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} \left(\hat{p}_{ijk} - p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\right)^2 \\
&= \sum_{i=1}^{I} \frac{1}{p_{i\bullet\bullet}} \left(\hat{p}_{i\bullet\bullet} - p_{i\bullet\bullet}\right)^2 + \sum_{j=1}^{J} \frac{1}{p_{\bullet j\bullet}} \left(\hat{p}_{\bullet j\bullet} - p_{\bullet j\bullet}\right)^2 + \sum_{k=1}^{K} \frac{1}{p_{\bullet\bullet k}} \left(\hat{p}_{\bullet\bullet k} - p_{\bullet\bullet k}\right)^2 \\
&\quad + \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{1}{p_{i\bullet\bullet}p_{\bullet j\bullet}} \left(\hat{p}_{ij\bullet} - \hat{p}_{i\bullet\bullet}p_{\bullet j\bullet} - \hat{p}_{\bullet j\bullet}p_{i\bullet\bullet} + p_{i\bullet\bullet}p_{\bullet j\bullet}\right)^2 \\
&\quad + \sum_{i=1}^{I}\sum_{k=1}^{K} \frac{1}{p_{i\bullet\bullet}p_{\bullet\bullet k}} \left(\hat{p}_{i\bullet k} - \hat{p}_{i\bullet\bullet}p_{\bullet\bullet k} - \hat{p}_{\bullet\bullet k}p_{i\bullet\bullet} + p_{i\bullet\bullet}p_{\bullet\bullet k}\right)^2 \\
&\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{1}{p_{\bullet j\bullet}p_{\bullet\bullet k}} \left(\hat{p}_{\bullet jk} - \hat{p}_{\bullet j\bullet}p_{\bullet\bullet k} - \hat{p}_{\bullet\bullet k}p_{\bullet j\bullet} + p_{\bullet j\bullet}p_{\bullet\bullet k}\right)^2 \\
&\quad + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{1}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} \left(\hat{p}_{ijk} - \hat{p}_{ij\bullet}p_{\bullet\bullet k} - \hat{p}_{i\bullet k}p_{\bullet j\bullet} - \hat{p}_{\bullet jk}p_{i\bullet\bullet}\right. \\
&\qquad \left. + \hat{p}_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k} + \hat{p}_{\bullet j\bullet}p_{i\bullet\bullet}p_{\bullet\bullet k} + \hat{p}_{\bullet\bullet k}p_{i\bullet\bullet}p_{\bullet j\bullet} - p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\right)^2 \\
&= \Phi_I^2 + \Phi_J^2 + \Phi_K^2 + \Phi_{IJ}^2 + \Phi_{IK}^2 + \Phi_{JK}^2 + \Phi_{IJK}^2 .
\end{aligned} \tag{4}$$

Here we can see that there are seven terms in the partition. The first term is the row main effect which, when multiplied by n and under the null hypothesis $p_{ijk} = p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}$, asymptotically follows a χ^2_{I-1} distribution, whose $1-\alpha$ percentile is $\chi^2_{\alpha,I-1}$ with $I-1$ degrees of freedom. Similarly, the column main effect, $n\Phi^2_J$ asymptotically follows a χ^2_{J-1} distribution while the tube main effect, $n\Phi^2_K$, follows a χ^2_{K-1} distribution. The bivariate terms of the partition - $n\Phi^2_{IJ}$, $n\Phi^2_{IK}$ and $n\Phi^2_{JK}$ - are measures of the row-column, row-tube and column-tube association, respectively, and are asymptotically chi-squared random variables with $(I-1)(J-1)$, $(I-1)(K-1)$ and $(J-1)(K-1)$ degrees of freedom, respectively. The last term of the partition, $n\Phi^2_{IJK}$ is the measure of three-way, or *trivariate*, association between all three variables and is asymptotically a chi-squared random variable with $(I-1)(J-1)(K-1)$ degrees of freedom. This general framework allows one to consider a mixture of tests, in particular goodness-of-fit tests, and association tests of the various relationships among the variables.

5 Example

To illustrate our partition briefly, we consider Pearson's index under *Scenario 0*; see Equation (4). The data concerns a study about the justice court delay in Italy [5]. We investigate the association among *Trial length*, *Subjects* of trials and *Number of Hearings*. The *Trial length* has four categories *low* duration (from 88 to 596 days), middle-low (from 597 to 1130 days), *mlow*, middle-long (from 1131 to 1950 days), *mlong*, and *long* (from 1951 to 6930 days) duration. The column variable, *Sub-*

ject of trials, has three categories *Obbligation, Controversy*, and *Real Rights* and also the tube variable has three categories, low number of hearings (from 0 to 2), *Hlow*, middle number of hearing (from 3 to 5), *Hmedium*, and high number of hearings (from 6 to 25), *Hhigh*. For a priori knowledge of the problem, we set the row probabilities as estimated from the data, the column probabilities as uniform, i.e. $p_{\bullet j \bullet} = (1/3, 1/3, 1/3)$, and the tube probabilities as equal to $p_{\bullet\bullet k} = (0.4, 0.4, 0.2)$. The overall association is $n\Phi^2 = \chi^2 = 4122$ with df=32 there exists strong evidence to say that it is statistically significant (p-value < 0.0001). The size and percentage contribution to χ^2 and the eight terms are reported in Table 2.

Table 1 Justice data: Crosstabulation of *Trial length* by *Subject* and by *Number Hearing*.

Hlow- *Low number of hearings*			
	Subject		
Trial Length	**Obbligation**	**Controversy**	**Real Right**
low	123	26	8
mlow	63	23	19
mlong	161	27	33
long	69	69	22
Hmedium- *Middle number of hearings*			
	Subject		
Trial Length	**Obbligation**	**Controversy**	**Real Right**
low	110	9	9
mlow	82	7	4
mlong	74	6	15
long	32	11	10
Hhigh- *High number of hearings*			
	Subject		
Trial Length	**Obbligation**	**Controversy**	**Real Right**
low	181	0	8
mlow	241	4	28
mlong	126	1	27
long	183	6	70

Table 2 Partition of the three-way chi-squared association measure under Scenario 0.

	X^2_J	X^2_K	X^2_{IJ}	X^2_{IK}	X^2_{JK}	X^2_{IJK}	X^2
Index	1591.15	869.84	92.94	152.93	1202.43	212.92	4122.22
% of Inertia	38.59	21.10	2.26	3.71	29.17	5.17	100.00
df	2.00	2.00	6.00	6.00	4.00	12.00	32.00
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00

This output shows that the most dominant contributor (38.6%) to the association among the univariate terms of Equation (4) is due to the inclusion of the *Subject* variable. The unique not significant term is due to the row variable *Trial Length* (whose hypothesized probabilities are set equal to the estimated frequencies), all the others are statistically significant and the most important concerns the *Subject* of trials. Of the bivariate association terms from the partition, the most important is due to the column and tube variables, *Subject* and *Number of Hearing*, it is about 29% of the total association, while the least important bivariate association concerns the row and tube variables, *Trial Length* and *Subject*, that is 2.3% of the total association among the variables. Finally, the contribute of the trivariate association term is low (around 5%).

6 Discussion

This paper has proposed a general expression of the partition for three-way association statistics, in particular for the traditional three-way Pearson statistic (see Equation 4) when hypothesising complete independence model. Further investigation of the partitions when considering complete and partial independence will lead to other quite distinct situations in partitioning symmetric and asymmetric three-way association indices [14, 13]. Comparisons with other goodness-of-fit statistics for large-sample, like the likelihood-ratio statistic or the Wald statistic [11] will be pursued.

References

1. Andersen E. B.: The Statistical Analysis of Categorical Data. Springer-Verlag, Germany (1980)
2. Agresti, A.: Categorical Data Analysis. John Wiley & Sons, New York, USA (1990)
3. Beh, E. J., Davy, J. P.: Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table. The Australian and New Zealand Journal of Statistics, **40**, 465–477 (1998)
4. Beh, E.J., Lombardo, R.: Correspondence Analysis, Theory, Practice and New Strategies. Wiley, Chichester (2014)
5. Camminatiello, I., Lombardo, R., Durand, J. F.: Robust Partial Least Squares Regression for the Evaluation of Justice Court Delay. Quality and Quantity Journal, 813–827 (2017)
6. Carlier, A., Kroonenberg, P.M.: Decompositions and Biplots in Three-way Correspondence Analysis. Psychometrika, **61**, 355–373 (1996)
7. Choulakian, V.: Exploratory analysis of contingency tables by loglinear formulations and generalizations of correspondence analysis. Psychometrika, **53**, 235–250 (1988)
8. Goodman, A. L., Kruskal, W. H.: Measures of Association for Cross Classifications. Journal of the American Statistical Association, **49**, 732–764 (1954)
9. Gray, L. N., Williams, J. S.: Goodman and Kruskals tau b: Multiple and Partial Analogs. American Statistical Association Proceedings of the Social Statistics Section, 444–448 (1975)
10. Lancaster, O. H.: Complex contingency tables treated by the partition of the chi-square. Journal of Royal Statistical Society, Series B, **13**, 242–249 (1951)
11. Lang, J. B.: On the Partitioning of Goodness-of-Fit Statistics for Multivariate Categorical ResponseModels. Journal of the American Statistical Association, **91**, 1117–1023 (1996)
12. Loisel, S., Takane, Y.: Partitions of Pearson's chi-square statistic for frequency tables: A comprehensive account. Computational Statistics, **31**, 1429–1452 (2016)
13. Lombardo, R.: Three-way association measure decompositions: the Delta index. Journal of Statistical Planning and Inference, **141**, 1789–1799 (2011)
14. Lombardo, R., Carlier, A., D'Ambra, L.: Nonsymmetric Correspondence Analysis for Three-Way Contingency Tables. Methodologica, **4**, 59–80 (1996)
15. Marcotorchino, M.: Utilisation des Comparaisons par Paires en Statistique des Contingencies: Partie I, Etude du Centre Scientifique, IBM, France, No F 069, (1984)
16. Saxe, K.: Beginning Functional Analysis. Springer, (2002)

Finding the best paths in university curricula of graduates to improve academic guidance services

Individuare i migliori percorsi di carriera universitaria per migliorare l'efficacia dei servizi d'orientamento d'Ateneo

Silvia Bacci and Bruno Bertaccini

Abstract Within the more general quality assurance pathways undertaken by the universities, inbound and on going guidance activities are assuming an increasingly strategic role with the aim to reduce the dropout rate and the time to qualification. In this contribution, the usefulness of some typical data mining solutions is evaluated in the context of students' careers analysis. More in detail, an analysis of graduates' careers paths is proposed, mainly based on the application and comparison of clustering procedures. With our proposal we aim at identifying career paths that are particularly virtuous in terms of average scores and time to qualification. Such type of information can be used by the university management for planning the career paths of freshmen.

Abstract *All'interno dei più generali percorsi di assicurazione della qualità intrapresi dagli Atenei, le attività di orientamento in ingresso e itinere stanno assumendo un ruolo sempre più strategico in ottica di riduzione del tasso d'abbandono e contenimento dei tempi di conseguimento del titolo. Questo lavoro intende valutare l'applicabilità di alcune soluzioni di data mining nel campo dell'analisi dei dati di carriera dei laureati; in particolare i percorsi scelti da coloro che hanno completato gli studi saranno analizzati attraverso alcune tecniche di clustering per l'identificazione di carriere virtuose (in termini di votazione media e tempo richiesto per il completamento del percorso) che possano essere proposte dagli organi di governo dei corsi di studio quale modello di riferimento per i nuovi iscritti.*

Key words: academic guidance services, cluster analysis, data mining, education, hidden Markov models, mixture models

Silvia Bacci
Dipartimento di Economia, Università di Perugia, Via A. Pascoli 20, 06123 Perugia (IT), e-mail: silvia.bacci@unipg.it

Bruno Bertaccini
Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti", Università degli Studi di Firenze,Viale Morgagni 59, Firenze (IT), e-mail: bruno.bertaccini@unifi.it

1 Introduction

The Italian university system is characterized by some peculiarities, *in primis* that students can freely decide when taking an exam and the specific sequence of enrolled exams. The main relevant consequence of this organizative approach are the long times to qualification. In such a context identifying career paths that are particularly virtuous in terms of average grades and time to qualification as well as the exams representing "bottlenecks" in the career flows is especially relevant for the university management in order to plan the career paths of freshmen and to improve the academic guidance services.

In this contribution we aim at studying the sequences of exams taken by a cohort of graduated students, through the comparison of some clustering approaches.

2 Data

The analysis here proposed is based on a cohort of 189 students in Business Economics at the University of Florence that enrolled the degree course in year 2012 and completed it within year 2017. In Figures 1 and 2 the sequences of first-year exams are shown, by average grade (low-high; Figure 1) and time to qualification (low-high; Figure 2). In both cases, attention is captured by differences in the observed sequences: in particular, the first-year exams that are more problematic in terms of tendency to postpone are Private Law and Mathematics.

3 Clustering approaches

Behavioral homogeneity of the cohort of students at issue is investigated through some clustering approaches:

Hierarchical cluster analysis. This a well-known approach that does not require any model specification, even if some choices are necessary (e.g., number of clusters, grouping method, type of distance). The analysis is performed on the exams' cumulative average score and on the cumulative time to qualification. This approach suffers for the presence of sparse data due to a large number of exams that are taken only by few students: as a consequence a preliminary cleanup of data is necessary.

Latent class model-based cluster analysis [4]. This is the simplest model-based clustering approach. The analysis is performed on the same variables as the previous approach. This approach has the same drawbacks as the hierarchical clustering.

Mixture Luce-Plackett model-based analysis [5]. Differently from the two above approaches, this approach works on the students' partial ranking of exams, that is for each student the corresponding sequence of exams is formulated in terms of a rank, where missing values correspond to exams that are not in the student's degree curriculum. In such a way the sequence of exams is explicitly taken into account. As main result, this approach provides a measure of liking toward each exam, separately for each cluster. The main practical drawback of this approach is that the clustering process does not account for the exam grades and, only partially, for the time gaps between exams, so that clusters are likely to overlap a lot in terms of exam grades and times to qualification. Moreover, the liking measures tend to assume high values for the very first enrolled exams and values around zero for the following exams, such that differences between clusters in terms of sequences may not be well defined.

Mixture Hidden Markov (MHM) model-based cluster analysis [2]. Similarly to the mixture Luce-Plackett model-based approach, this approach directly models the sequences of exams. More in details, a multiple multichannel MHM model is specially suitable for data at issue, as it will be cleared in the next section. Terms "multiple" and "multichannel" refer to the presence of more than one individual (i.e., cohort of students) and of more than one sequence of data for each individ-

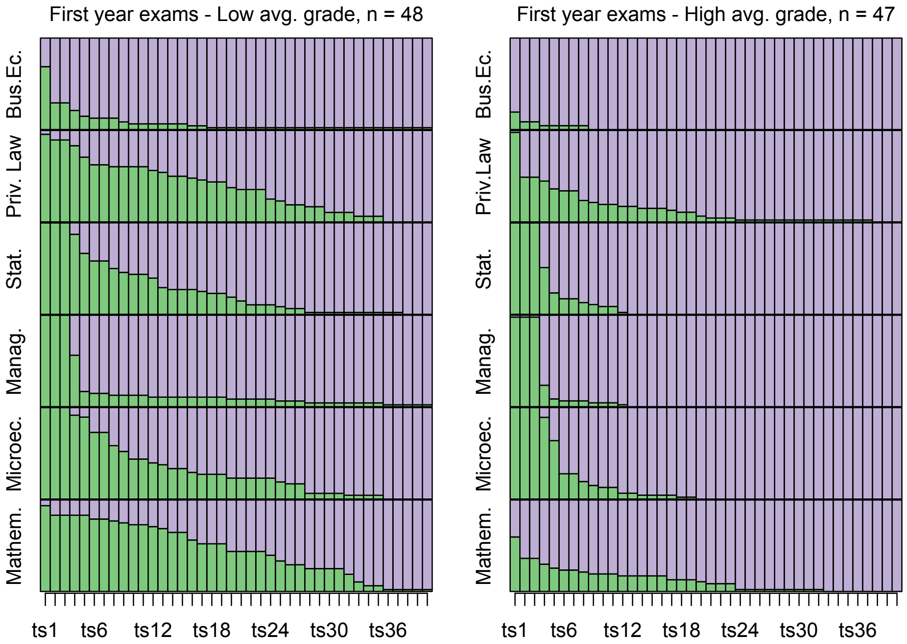


Fig. 1 Sequences of first-year exams by average grade: average grade less than first quartile (left panel) and average grade greater than third quartile (right panel). Legend: green = exam not taken; purple = exam taken.

ual (i.e., one sequence for each exam), respectively. In practice, the MHM model presents some specific advantages with respect to the other approaches: (*i*) the presence of sparse data is not a problem (exams that are not present in the degree curriculum of a student correspond to a sequence of 0s), (*ii*) the observed sequences of exams are explicitly modelled and individuals are clustered on the basis of these sequences, (*iii*) in addition to the other finite mixture approaches, it allows us an in-depth analysis of every exam within each cluster, such that weaknesses of the different typologies of students are highlighted, (*iv*) differently from the other approaches, it allows us to enclose in the analysis students that dropped out or that did not yet finished the exams of the degree course.

3.1 Mixture Hidden Markov model

HM models [3, 1] represent a nice frame to analyse sequence data. In such a context, a sequence of 0s and 1s represents the observed states, which are interpreted as

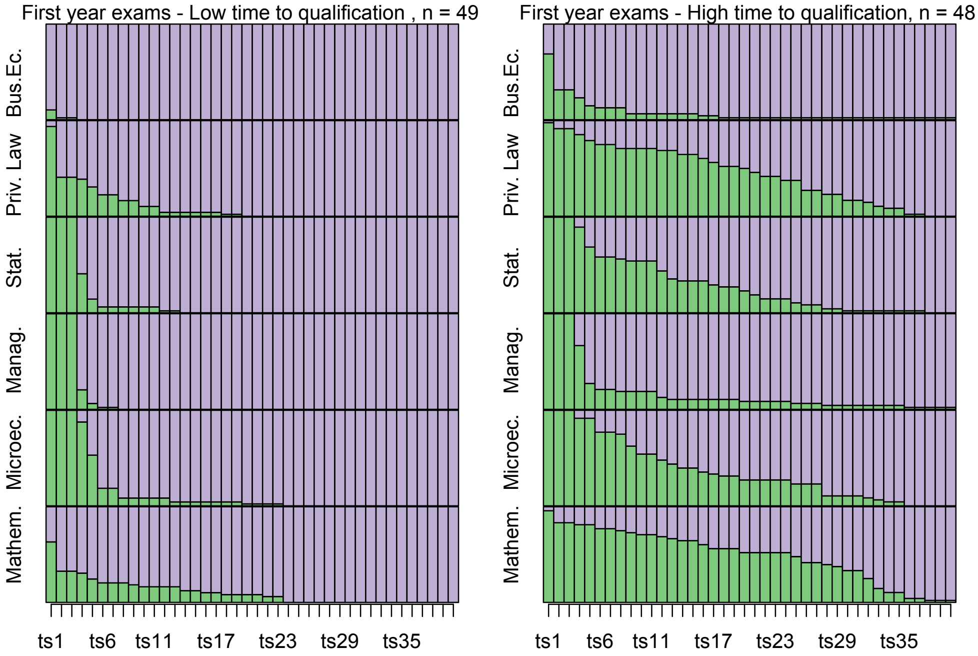


Fig. 2 Sequences of first-year exams by time to qualification: time less than first quartile (left panel) and time greater than third quartile (right panel). Legend: green = exam not taken; purple = exam taken.

probabilistic manifestations of a certain number of unobservable (i.e., hidden or latent) states. In our contribution, we assume that student i in any time point may belong to one of two hidden states: state $u_{it} = 1$ denotes a low propensity to take exams at time t and state $u_{it} = 2$ denotes a high propensity to take exams at time t. As usual in the HM models, students can move from a state to another one. In addition, we introduce the observed state $y_{itj} = y$, with $y = 1$ if student i takes exam j at time t or before, and $y = 0$ if student i did not yet taken exam j at time t; $j = 1,\dots,J$, with J denoting the total number of exams (it does not matter how many students choose an exam for their own degree curriculum), and $t = 1,\dots,T$ with T denoting the length of any sequence and corresponds to the number of exam sessions scheduled in the years 2012-2017.

Hence, we have two types of vectors for each student: vector $\mathbf{u}_i = (u_{i1},\dots,u_{it},\dots,u_{iT})$ of hidden state sequence and vector $\mathbf{y}_{ij} = (y_{i1j},\dots,y_{itj},\dots,y_{iTj})$ of observed state sequence; note that we have one vector $\mathbf{y}_{ij}$ for each exam.

The probability of observed sequence of data is formulated according to a time-homogeneous multivariate HM model:

$$p(\mathbf{Y}_{ij} = \mathbf{y}_{ij}) = \sum_{u=1}^{2} p(\mathbf{y}_{ij}|\mathbf{u}_i)p(\mathbf{u}_i) = \sum_{u=1}^{2} \left[p(y_{i1j}|u_{i1})p(u_{i1}) \prod_{t=2}^{T} p(y_{itj}|u_{it})p(u_{it}|u_{i,t-1}) \right], \tag{1}$$

with: $p(y_{itj}|u_{it})$ conditional probability of observed state given the hidden state (emission probability), $p(u_{i1})$ initial probability of starting from hidden state u_{i1}, and $p(u_{it}|u_{i,t-1})$ transition probability of moving from hidden state $u_{i,t-1}$ to hidden state u_{it}. This model is usually estimated through the maximization of the log-likelihood function, using the forward-backward algorithm.

A generalization of the HM model is represented by the MHM model, which is based on the assumption that the population is composed by homogenous groups of individuals and each of these groups follow a specific HM model. In such a context, model in equation 1 modifies as

$$p(\mathbf{Y}_{ij} = \mathbf{y}_{ij}) = \sum_{k=1}^{K} \pi_k \left\{ \sum_{u=1}^{2} p(\mathbf{y}_{ij}|\mathbf{u}_i)p(\mathbf{u}_i) \right\}, \tag{2}$$

where π_k denotes the prior probability that the sequence of observed states of an individual belongs to cluster k.

To synthetize, the mixture part of the model allows us to cluster students in groups that are homogenous in terms of observed sequences of exams and propensity to take exams. In this way a classification of students is obtained, which is comparable with those obtained in the previously described approaches. In addition, the HM part of the model allows us for an in-depth analysis of the performance of students on single exams.

4 Main results

To define homogenous sub-groups of students, we selected a number of clusters equal to three and we applied the clustering procedures above described. In Table 1 the results in terms of average grade, average time to qualification and cluster size are illustrated.

Table 1 Average grade and average time to qualification by cluster and clustering approach. In bold the best performances, in italic the worst performances.

Approach	Variables	Cluster 1	Cluster 2	Cluster 3
Hierarchical clustering				
	avg. grade (out of 30)	**26.14**	*23.69*	24.99
	avg. time (days)	**1037.04**	*1598.90*	1262.98
	# of students	54	48	87
Latent class clustering				
	avg. grade (out of 30)	25.35	*24.05*	**26.74**
	avg. time (days)	1171.83	*1436.12*	**1079.16**
	# of students	59	93	37
Mix. Luce-Plackett clust.				
	avg. grade (out of 30)	**25.59**	24.97	*24.13*
	avg. time (days)	**1305.47**	1382.06	*1447.91*
	# of students	30	148	11
MHM clustering				
	avg. grade (out of 30)	25.25	*23.99*	**25.81**
	avg. time (days)	1324.83	*1577.25*	**1219.21**
	# of students	46	68	75

It is worth to be noted that all approaches provide a cluster of best performers (in bold) and a cluster of worst performers (in italic) with respect to both criteria of average grade and time to qualification. The clustering approach based on partially ranked data is the least satisfactorily as concerns the level of separability among clusters.

Additional details about the cluster characteristics are provided by the estimated parameters of the MHM model. In Table 2 the emission probabilities of the first-year exams are shown, which describe the probability of taking an exam - in any time point - given the hidden state, that is, $p(Y_{itj} = 1|u_{it})$; these probabilities are cluster-specific. For the sake of clarity we remind that state 1 denotes a low propensity to take exams and state 2 denotes a high propensity to take exams; then, the lower $p(Y_{itj} = 1|u_{it})$, the higher the tendency to postpone exam j.

As shown in Table 1, the worst performers are allocated in cluster 2. In more detail 2, students of cluster 2 belonging to state 1 have a high tendency to postpone Mathematics and Private Law (emission probabilities equal to 26.6% and 33.3%, respectively), followed by Microeconomics (46.0%) and Statistics (57.8%). Mathematics and Private Law represent bottlenecks also for the colleagues in state 2 (emission probabilities equal to 79.1% and 83.8%, respectively).

Table 2 MHM model: Estimated emission probabilities by cluster (only first-year exams)

Exam	Hidden state	Cluster 1	Cluster 2	Cluster 3
Business Economics				
	State 1	0.893	0.925	0.983
	State 2	0.984	1.000	1.000
Private Law				
	State 1	0.636	0.333	0.655
	State 2	0.962	0.838	0.987
Statistics				
	State 1	0.609	0.578	0.740
	State 2	0.991	0.963	0.998
Management				
	State 1	0.704	0.763	0.804
	State 2	0.977	1.000	1.000
Microeconomics				
	State 1	0.493	0.460	0.677
	State 2	0.956	0.929	1.000
Mathematics				
	State 1	0.438	0.266	0.783
	State 2	0.885	0.791	0.989

On the opposite, the best performers are allocated in cluster 3. In such a cluster the main problems are observed for Private Law and Microeconomics: for both the exams the emission probabilities for state 1 are strongly smaller than 1 (65.5% for Private Law and 67.7% for Microeconomics).

Finally, cluster 1 lies in an intermediate position with respect to clusters 2 and 3 for all (the first-year) exams, with the exception of Business Economics, for which a certain tendency to postpone is observed for both the hidden states (emission probabilities are higher than those of cluster 2).

A synthetic representation of the three clusters is provided by Figure 3, where the estimated sequence of hidden states is displayed for each cluster. It is worth to be noted that around the centre of the time line cluster 2 presents a tail for state 1 that is heavier with respect to cluster 1 and, mainly, cluster 3.

5 Conclusions

The illustrated approaches of cluster analysis, with a special attention for the mixture Hidden Markov model, represent useful instruments for the academic management to detect critical exams for specific sub-groups of students and, more in general, to plan the career paths of freshmen and to improve the academic guidance services.

For the future development of this contribution we intend to extend the analysis to the entire cohort of students, including students that dropped out from the university

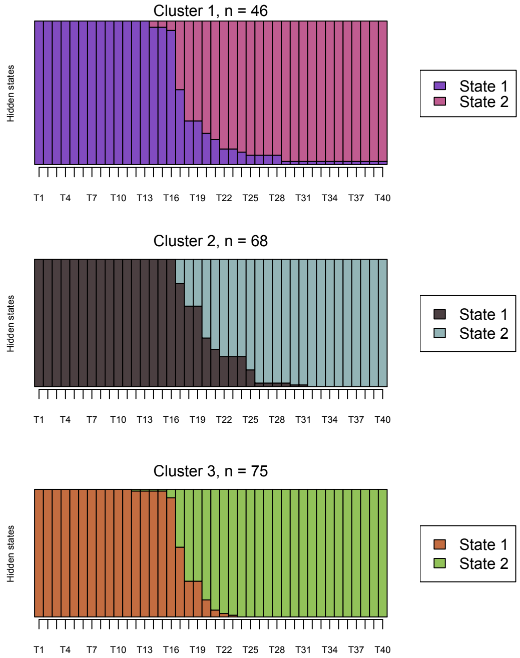


Fig. 3 MHM model: Sequences of hidden states by cluster.

and students that have to finish the exams of the degree course. We also intend to extend the analysis to account for individual characteristics (e.g., type of high school, high school graduation mark).

References

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov models for longitudinal data. Chapman & Hall/CRC, Boca Raton, FL (2012)
2. Helske, S., Helske, J.: Mixture Hidden Markov models for sequence data: the seqHMM package in R. Available via https://arxiv.org/pdf/1704.00543.pdf. Cited 25 Apr 2018
3. Zucchini, W., MacDonald, I. L.: Hidden Markov Models for Time Series: an Introduction using R. Springer, New York (2009)
4. McLachlan, G., Peel, D.: Finite mixture models. Wiley, New York (2000)
5. Mollica, C., Tardella, L.: PLMIX: An R package for modeling and clustering partially ranked data. Available via https://arxiv.org/pdf/1612.08141.pdf. Cited 25 Apr 2018

Statistical Modelling for Business Intelligence Problems

A nonlinear state-space model for the forecasting of field failures

Un modello state-space non lineare per la previsione di guasti post vendita

Antonio Pievatolo

Abstract We consider time series of field failure data (warranty claims) of domestic appliances, manufactured by different plants, with the aim of forecasting failures within the warranty period. The monthly failure profiles over two-year periods display variation across monitoring epochs and also batch-to-batch variation. A nonlinear state space model is developed to jointly represent the variation of the underlying failure rate parameters and the observed occurrence of failures, obtaining a dynamic Poisson-Lognormal model with a meaningful covariance structure of failure rates between monitoring epochs. An adaptation of the auxiliary particle filter is used for parameter learning and forecasting. A series of examples with data from two different production plants show that it is possible to obtain a small forecasting error for claims having very different patterns.

Abstract *Analizziamo serie storiche di riparazioni in garanzia di elettrodomestici, prodotti da impianti differenti, per prevedere la frequenza di guasti entro la fine del periodo di garanzia. I profili di guasto mensili in periodi di due anni mostrano variazioni sia tra periodi di monitoraggio sia tra lotti di produzione. Sviluppiamo un modello state space per rappresentare allo stesso tempo la variazione del parametro del tasso di guasto sottostante e l'apparizione dei guasti, ottenendo un modello dinamico Poisson-Lognormale con una appropriata struttura di covarianza tra periodi di monitoraggio. Utilizziamo un adattamento dell'auxiliary particle filter per la stima dei parametri statici e per la previsione. Alcuni esempi con dati da due impianti di produzione mostrano che è possibile ottenere un piccolo errore di previsione per serie di guasti post-vendita con profili molto diversi.*

Key words: Warranty claims, State-space model, Poisson-Dirichlet model, Poisson-Lognormal model, Particle learning, Covariance of failure rates

Antonio Pievatolo
CNR IMATI, Via A. Corti, 12, Milano, e-mail: antonio.pievatolo@cnr.it

1 Introduction

Warranty data have long been recognized as a source of useful information for several purposes, some of which are prediction of future claims, comparison of groups of products and identification of faulty design or material defects or undetected production line problems. [5] provided one of the first review articles on this subject.

Claims are often aggregated by epoch index (e.g. by month) for every production batch, so that available data may be represented by n_t, for the number of items produced at epoch t, and y_{tj}, for the number of claims at epoch $t+j-1$ for batch t, $j=1,\ldots,d$, where d is the duration of the warranty period.

[5] proposed to model the claim arrival process of a given production batch as a sequence of independent Poisson random variables: $Y_{tj} \sim Poisson(n_t \lambda_j)$, as $j=1,\ldots,d$, where λ_j is the expected number of failures per produced item in batch t at epoch $t+j-1$. [9] used the same model for the early detection of reliability problems. Other works are focussed on the modelling of failure times instead ([3]; [4]).

The model based on independent Poisson counts has the main drawback that λ_j does not depend on t so it cannot adequately describe batch-to-batch variation, which could be attributed to: material defects due for example to a change of supplier; changes in the production line that may affect reliability; other unmeasured batch to batch heterogeneity.

This deficiency could be addressed following [6], who introduced a rate function for car warranty data, depending on a unit-specific random effect Z_t and a usage function modelled as $U_{tj} = jZ_t$. This approach requires to select a parametric form for the rate function and does not include possible dependence among unit failure rates at close t values. Furthermore it was developed for situations in which times of occurrence of failures of each unit are available, which is not our present situation.

In this work we have proposed a state space modelling framework in which the observation equation for (conditionally independent) describes a Poisson distribution, whereas the batch-to-batch heterogeneity and the possible dependence among failures of units from batches close in time is represented by a state transition equation on failure rates. In this way the failure rates are regarded as generated by a stochastic process and no assumption on their form is needed. On the other hand, we will focus on the specification of a meaningful dependence structure among failure rates and on a particle learning algorthm, which will provide forecasts on future failure rates before the end of the warranty period and also learn about unknown model parameters.

Our methodology has been applied to claims regarding home appliences manufactured by a multinational company with markets and production plants in several countries.

2 A state space model

Let $y_t = (y_{t1}, \dots, y_{td})^T$ and $\lambda_t = (\lambda_{t1}, \dots, \lambda_{td})^T$, where, unkile the introductory section, λ_t now denotes a vector of failure rates. We consider a state space model for claims

$$\begin{aligned} y_{tj} &\sim Poisson(n_t \lambda_{tj}), \quad j = 1, \dots, d \\ \lambda_t | \lambda_{t-1} &\sim f(\lambda_t | \lambda_{t-1}; \theta) \end{aligned} \tag{1}$$

where θ is an unknown parameter. This model is flexible enough to be able to describe batch-to-batch variation, dependence between batches, dependence between claim numbers in different epochs for the same batch, variation of claim reporting patterns. The task is now to select the form of the state transition equation, which we will do by examining observations from two production plants, where epochs are months, in Figure 2. The left-hand panels show a substantial batch-to-batch variation of observed failure rates for entire batches (number of failures in two years divided by the batch size). The right-hand panels highlight the within-batch variation, that is, how the overall number of failures in two years is distributed over the epochs (months) for all observed batches, with variablity in the shape of the curves.

A model which separates these two types of variation is the following Poisson-Dirichlet model

$$\begin{aligned} y_{tj} &\sim Poisson(n_t \mu_t p_{tj}), \quad j = 1, \dots, d \\ \log(\mu_t) &= \log(\mu_{t-1}) + \sigma w_t \\ p_t &= \gamma p_{t-1} + (1-\gamma) q_t \end{aligned} \tag{2}$$

where μ_t is a scalar, $q_t \sim Dir(M\eta)$, a d-dimensional Dirichlet distribution with mass parameter M and shape parameter $\eta = (\eta_1, \dots, \eta_d)$ such that $\sum_j \eta_j = 1$, and $\gamma \in (0,1)$, $w_t \sim N(0,1)$. The unknown parameter θ includes σ, η and M. The second equation describes the dynamics of the overall batch failure rate, whereas the third one describes the within-batch variation, also defining a covariance structure among epochs in the same batch via the Dirichlet error q_t.

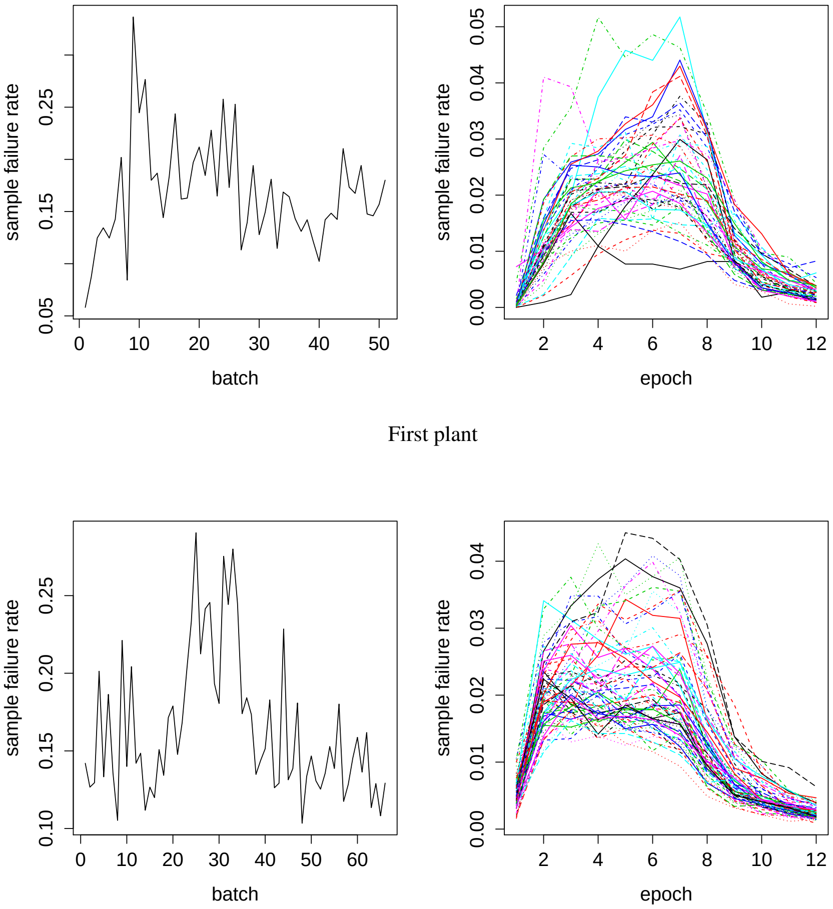
A more tractable version of this model, with a view to sequential Bayesian parameter update, can be obtained by collapsing the two state equations into one. Letting $\lambda_{tj} = \mu_t p_{tj}$, we derive the following:

$$\log \lambda_{tj} = \log \lambda_{t-1,j} + \varepsilon_{tj} \tag{3}$$

where $\varepsilon_{tj} = \log \lambda_{t-1,j} + \log(\gamma + (1-\gamma) q_{tj} / p_{t-1,j}) + \sigma w_t$ and the vector ε_t has a non-diagonal covariance matrix, which we can approximate via error-propagation formulas from the covariance structure of q_t. By doing so, we find

$$Var(\varepsilon_{tj}) \simeq \frac{(1-\gamma)^2 \eta_j (1-\eta_j)}{M+1} \frac{1}{(\gamma p_{t-1,j} + (1-\gamma)\eta_j)^2} + \sigma^2$$

which is further approximated by



First plant

Second plant

Fig. 2: First column: plot of pairs $(t, \sum_j y_{tj}/n_t)$; second column: plots of pairs $(j, y_{tj}/n_t)$, for all available values of t

$$\frac{(1-\gamma)^2}{M+1}\frac{(1-\eta_j)}{\eta_j}+\sigma^2=\tau^2\frac{(1-\eta_j)}{\eta_j}+\sigma^2$$

using η_j for $p_{t-1,j}$. Continuing with the covariances, by the error propagation formulas and using η_j for $p_{t-1,j}$, $Cov(\varepsilon_{tj},\varepsilon_{tr}) \simeq -\tau^2/2+\sigma^2$ and finally $E(\varepsilon_{tj}) \simeq 0$.

Then, the new Poisson-Lognormal state-space model is defined as

$$\begin{aligned} y_{tj} &\sim Poisson(n_t e^{\alpha_{tj}}), \quad j=1,\dots,d \\ \alpha_t &= \alpha_{t-1}+\varepsilon_t \end{aligned} \tag{4}$$

where $\alpha_t = \log\lambda_t$, $\varepsilon_t \sim N_d(0,\Sigma)$ and

$$\begin{aligned} \Sigma_{jj} &= \tau^2\frac{(1-\eta_j)}{\eta_j}+\sigma^2, j=1,\dots,d \\ \Sigma_{jr} &= -\frac{\tau^2}{2}+\sigma^2, j,r=1,\dots,d,\ j\neq r\ . \end{aligned} \tag{5}$$

Model (4)-(5) is not equivalent to the original model (2) and the normality of the error term has been assumed ex post, however this approximation procedure has provided a justification for applying a certain covariance structure to the logarithms of the failure rates. Furthermore, conditional on α_{t-1}, the model for (y_t,α_t) is the multivariate Poisson-lognormal model of [1], for which $Var(Y_{tj}) > E(Y_{tj})$, allowing for overdispersion relative to the Poisson distribution.

3 Particle filtering for claim forecasting

Let $\alpha^t = (\alpha_1,\dots,\alpha_t)$ and $y^t = (y_1,\dots,y_t)$. The filtering distribution at time t is

$$f(\alpha_t|y^t)$$

and it encodes the state of information on the claim rate vector given the observed claims. However, the needed warranty claim rate information for batch t is already provided by $\sum_j y_{tj}/n_t$, and the filter becomes useful for an early assessment of the overall claim when only a part of y_t is observed. So we seek to obtain

$$f(\alpha_t|y^{t-1},y_{t1},\dots,y_{tr_t}), \quad r_t<d,$$

where r_t is the latest observed epoch for batch t, taking advantage of the covariance structure of our state-space model.

By combining the methodology of the auxiliary particle filter (APL) of [8] and of the particle learning method of [2], we have obtained a new modified APL which also includes parameter learning on Σ and converges to the correct filtering and posterior distributions. The filter has not been applied to model (4)-(5), but to a relaxed version to allow for parameter updating using conjugacy. In particular, an inverse-Wishart initial distribution has been assigned to Σ_0, that is, $\Sigma \sim iW(\Sigma_0,\nu_0)$. Then, given that Σ is independent from y^t given α^t, its posterior distribution, conditional on the past history of the observations and of the states, is

$$\Sigma|s_t \sim iW(\Sigma_0+\sum_{r=1}^{t}(\alpha_r-\alpha_{r-1})(\alpha_r-\alpha_{r-1})^T, \nu_0+t) \tag{6}$$

where $s_t = \sum_{r=1}^{t} (\alpha_r - \alpha_{r-1})(\alpha_r - \alpha_{r-1})^T$. At time $t+1$ this distribution can then be updated by knowing only the sufficient statistics s_t and $(\alpha_{t+1} - \alpha_t)$, which, together, give s_{t+1}. The covariance structure (5) is not discarded, but is used to provide the Σ_0 parameter of the initial distribution of Σ, which, as experiments showed, must been assigned carefully.

The derivation of our modified APL is not shown here, for reasons of space. Instead we describe now the filtering algorithm, which provides, at each time t, a discrete approximation of the filtering density as a set of support points with associated weights. Additional yet undefined notation used for the algorithm, is L, the likelihood function determined by the observation equation in model (4) and the function f, which denotes a conditional density.

Given a step-t sample $(\alpha_t, s_t, \Sigma)^{(i)}$ and weights $w_t^{(i)}$, the modified APL goes through the following steps:

1. *Resampling step*: sample N index values k_i, $i = 1, \ldots, N$, using weights proportional to $L(\mu_{t+1}^{(k)}; y_{t+1}) w_t^{(k)}$, putting $\mu_{t+1}^{(k)} = \alpha_t^{(k)}$
2. *Propagation step*: sample N new particles $\alpha_{t+1}^{(i)}$ from $f(\alpha_{t+1} | \alpha_{t,ADJ}^{(k_i)}, \Sigma^{(k_i)})$ where $\alpha_{tj,ADJ}^{(k_i)} = \log(y_{t+1,j}/n_{t+1})$ if $j \leq r_{t+1}$, whereas $\alpha_{tj,ADJ}^{(k_i)} = \alpha_{tj}^{(k_i)}$ if $j > r_{t+1}$
3. Compute weights

$$w_{t+1}^{(i)} \propto \frac{L(\alpha_{t+1}^{(i)} | y_{t+1}) f(\alpha_{t+1}^{(i)} | \alpha_t^{k_i}, \theta^{(k_i)})}{L(\mu_{t+1}^{(i)} | y_{t+1}) f(\alpha_{t+1}^{(i)} | \alpha_{t,ADJ}^{k_i}, \theta^{(k_i)})}$$

   update sufficient statistics $s_{t+1}^{(i)} = s_t^{(k_i)} + (\alpha_{t+1}^{(i)} - \alpha_t^{(k_i)})(\alpha_{t+1}^{(i)} - \alpha_t^{(k_i)})^T$ and sample $\Sigma^{(i)} \sim iW(\Sigma_0 + s_{t+1}^{(i)}, \nu_0 + t + 1)$

4 Data examples

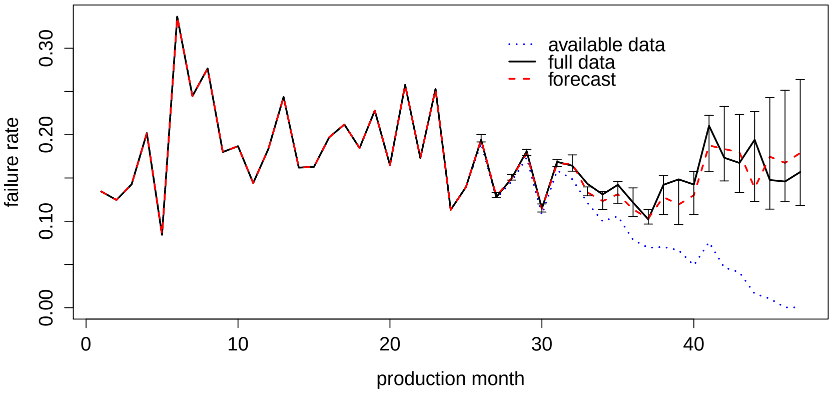
Because of the abundance of training data, the prior can be pretty informative. Therefore the initial distribution for Σ is $iW((R+1)\Sigma_0, R+d+2)$, with large R. With this parameterization, $E(\Sigma) = \Sigma_0$ a priori. The initial Σ_0 is computed using $\eta_j = \sum_t y_{tj} / \sum_t n_t$ from a training sample (such as another plant), then the filtered claim rate for batches with r_t observed claim epochs, $r_t < d$, is estimated as

$$\frac{1}{n_t} \left(\sum_{j=1}^{r_t} y_{tj} + \sum_{j=r_t+1}^{d} \sum_{i=1}^{N} w_t^{(i)} n_t \lambda_{tj}^{(i)} \right) .$$

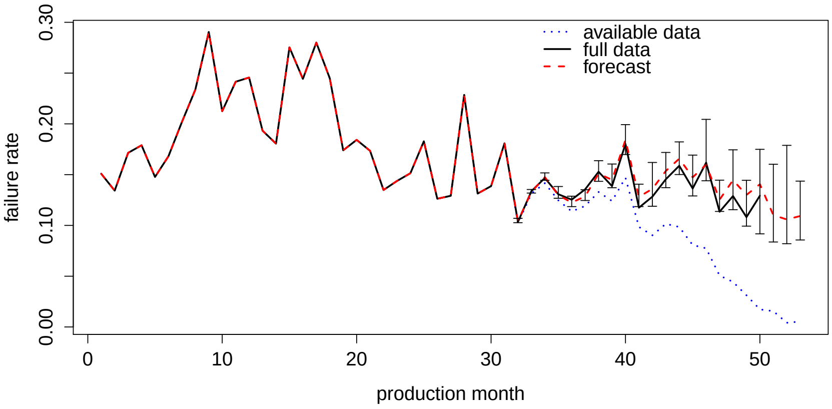
For predictive interval forecasts we resample particles and draw

$$\sum_{j=r_{t+1}}^{d} y_{tj}^{(i)} \sim Poisson(n_t \sum_{j=r_{t+1}}^{d} \lambda_{tj}^{(i)})$$

for every resampled particle. Then we take sample quantiles. The result of the forecasting procedure, including prediction intervals, is displayed in Figure 4, showing a good performance even with very little information on the current production batch.



First plant



Second plant

Fig. 4: Forecasts of claims for the two example plants versus production batch. Solid line: complete data; dashed line: forecast; dotted line: ratio between available number of claims and batch size at the time of the forecast, representing the available information (the lower the less)

5 Conclusions

The state-space Poisson-Lognormal model developed in this work has shown a good potential for making early prediction of future claims from customers during the warranty period of a domestic appliance, thanks to the design of an appropriate covariance structure of within-batch failure rates and to parameter learning. This approach is ideal for the sequential monitoring and prediction of claims when they occur as counts in predefined monitoring intervals, without the need of any detailed modelling of known disturbances such as claim reporting delays and of the usage pattern of appliances. Experiments indicate that a good elicitation of the initial value of the covariance matrix is requested, which is possible in the present situation because of the abundance of data, therefore future work can be directed to the exploration of results using a less concentrated initial distribution, as well as to other parameter learning strategies that improve the convergence of the particle filter.

Acknowledgements This work is a follow-up to [7]

References

1. Aitchison, J., Ho, C.H.: The multivariate Poisson-log normal distribution. Biometrika, **76**, 643-653 (1989)
2. Carvalho, C.M., Johannes, M.S., Lopes, H.F., Polson, N.G.: Particle learning and smoothing. Statistical Science, **25**, 88-106 (2010)
3. Hong, Y., Meeker, W.Q.: Field-failure and warranty prediction based on auxiliary use-rate information. Technometrics, **52**, 148-159 (2010)
4. King, C., Hong, Y., Meeker, W.Q.: Product Component Genealogy Modeling and Fieldfailure Prediction. Qual. Reliab. Eng. Int., **33**, 135-148 (2017)
5. Lawless, J.F.: Statistical analysis of product warranty data. Int. Stat. Rev., **66**, 41-60 (1998)
6. Lawless, J.F., Crowder, M.J., Lee, K.A.: Analysis of reliability and warranty claims in products with age and usage scales. Technometrics, **51**, 14-24 (2009)
7. Pensa, R.: Particle filters: un'applicazione per la previsione di guasti di elettrodomestici. Master thesis. Politecnico di Milano (2015)
8. Pitt, M.K., Shephard, N.: Filtering via simulation: Auxiliary particle filters. Journal of the American statistical association, **94**, 590-599 (1999)
9. Wu, H., Meeker, W.Q.: Early detection of reliability problems using information from warranty databases. Technometrics, **44**, 120-133 (2002)

Does Airbnb affect the real estate market? A spatial dependence analysis

Il fenomeno di Airbnb influenza il mercato immobiliare? Un'analisi di dipendenza spaziale

Mariangela Guidolin and Mauro Bernardi

Abstract The problem of evaluating and forecasting the price variation of houses is a traditional one in economic statistics, and the literature dealing with it is very rich. Part of this literature has focused on spatial statistics models in order to account for the structure of spatial dependence among house prices, and studied the relationship between prices and house features, such as dimension, position and type of building. In this paper, we try to extend this approach by considering the effect of exogenous variables, that may exert a significant impact on price dynamics, namely the level of crime and the Airbnb phenomenon. In particular, to our knowledge, the evaluation of the Airbnb activity on the real estate market is still in its infancy, but we expect an increasing role of it. In doing so, we considered the case of New York city, for which this information is fully available as *open data*, and employed spatial autoregressive and spatial error models, in order to study the impact of these variables along with typical house features on the real estate market for each district of the city.

Key words: Bayesian methods, Spike–and–Slab prior, Spatial dependence, Open data, Forecasting.

1 Introduction

A traditional problem in the economic statistics literature has to do with the dynamics of the real estate market and the factors affecting it. An extensive stream of literature has devoted special attention to studying the price variation of houses. Part of this literature has employed spatial statistics models in order to account for the structure of spatial dependence among house prices, and studied the relationship

Mariangela Guidolin
Department of Statistical Sciences, Via Battisti 241, 35121, Padua, e-mail: guidolin@stat.unipd.it

Mauro Bernardi
Department of Statistical Sciences, Via Battisti 241, 35121, Padua e-mail: mbernardi@stat.unipd.it

between prices and house features, such as dimension, position and type of building. In this paper, we try to extend this approach by considering the effect of exogenous variables, that may exert a significant impact on price dynamics, namely the level of crime, the Airbnb phenomenon and the distance from a metro station. In particular, to our knowledge, the evaluation of the Airbnb activity on the real estate market is still in its infancy, but we expect an increasing role of it. In doing so, we considered the case of New York city, for which this information is fully available as *Open Data*, and employed spatial autoregressive and spatial error models, in order to study the impact of these variables along with typical house features on the real estate market for each district of the city. The obtained results confirm our hypothesis on the impact of such variables, opening new perspectives on spatial modelling in the real estate context. The rest of the paper is organised as follows. Section 2 describes the dataset that motivates our empirical analysis and methodological developments, Section 3 briefly review the class of spatial models that we vill employ to model our data while Section 4 deals with the problem of selecting the relevant regressors. Section 5 concludes presenting our main findings.

2 Data set description

The case study here analyzed aims to study the real estate market in New York city and test how some variables directly available as *Open Data* can have an impact on house prices within the different city districts, namely Bronx, Brooklyn, Manhattan, Queens and Staten Island. The main data set contains all the information about house sales in New York for the period 2014–2016, namely: district, type of building, address, size (m^2), year of construction, price, date of sale. Moreover, we considered as potentially significant variables the level of crime of each district, the proximity of a house to a metro station, and the presence of Airbnb in each district. Specifically, taking a temporal window of 6 to 24 months backward with respect to the date of sale, we considered the number of crimes committed, the number of announcements on Airbnb website, the number of positive reviews, the number of host subscriptions, the average price of houses on Airbnb, and the minimum distance to a metro station.

3 Spatial models

Because the data set considered has a spatial cross-section nature, it is necessary to account for spatial dependence between observations in our modelling. To this end, we may follow a wide accepted stream of literature in economics and urban studies, which ascribes such spatial dependence either to a *spillover* phenomenon, implying that prices of spatially close observations will be correlated, or to the *omission* of a variable, which is important for the model but difficult to measure or identify. Since both interpretations of spatial dependence appear plausible, we choose to employ two different modelling approaches incorporating the two, namely the Spatial Autoregressive Model, SAR, and the Spatial Error Model, SEM. The SAR model has the following structure

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \alpha \iota_n + \mathbf{X}\beta + \varepsilon, \tag{1}$$

where $\mathbf{y}$ is the response variables, ρ is a spatial autocorrelation coefficient, defining the intensity of spatial correlation, $\mathbf{W}$ is a $(n \times n)$ matrix of spatial weights, $\mathbf{W}$ is a $(n \times n)$ of explanatory variables, α denotes the intercept of the model, ι_n denotes a unit column vector of dimension n, β is a $(k \times 1)$ vector of regression coefficients associated to the $(n \times k)$ matrix $\mathbf{X}$, $\varepsilon \sim \mathsf{N}(0, \sigma^2 \mathbf{I}_n)$ is the error term. The SEM model incorporates spatial dependence in the error term, has the following structure

$$\mathbf{y} = \alpha \iota_n + \mathbf{X}\beta + \xi \tag{2}$$

$$\xi = \lambda \mathbf{W} \xi + \varepsilon, \tag{3}$$

where λ is the spatial autoregressive coefficient, measuring the effect of omitted variables and model misspecification, and $\varepsilon \sim \mathsf{N}(0, \sigma_\varepsilon^2 \mathbf{I}_n)$ is an error term. The next section will deal with the selection of the relevant regressors for both the SAR and SEM specifications. To this aim let us introduce the transformed variables $\tilde{\mathbf{y}}(\rho) = (\mathbf{I}_n - \rho \mathbf{W})\mathbf{y} = \mathbf{P}_\rho \mathbf{y}$ and $\xi = (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \varepsilon = \mathbf{P}_\lambda^{-1} \varepsilon$ that allows to formulate the SAR and SEM specifications in the following compact way

$$\tilde{\mathbf{y}}(\rho) = \alpha \iota_n + \mathbf{X}\beta + \varepsilon \tag{4}$$

$$\mathbf{y} = \alpha \iota_n + \mathbf{X}\beta + (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \varepsilon. \tag{5}$$

4 Spatial variable selection

Model selection is performed by extending the Stochastic Search Variable Selection (SSVS) algorithm of George and McCulloch (1993). Specifically we propose a SSVS algorithm based on dirac spike and slab Lasso prior specifically tailored to select relevant covariates in the spatial regression context here considered. As in Hans (2009) the main characteristic of the proposed method is that it does not rely on the stochastic representation of the Lasso prior as scale mixture of Gaussians and the associated Gibbs sampling approach. Before considering the Spike–and–Slab approach we introduce the Bayesian version of the Lasso regression problem. In what follows, we refer to the spatial SAR and SEM models defined in the previous Section.

4.1 Likelihood and prior

Assuming the Laplace prior structure specification as in Park and Casella (2008) for the vector of spatial regression parameters β, we get the following representation of the linear model

$$\pi\left(\mathbf{y} \mid \mathbf{X},\mathbf{W},\alpha,\beta,\rho,\sigma^2\right) = \mathsf{N}\left(\tilde{\mathbf{y}}\left(\rho\right) \mid \alpha\iota_n + \mathbf{X}\beta, \sigma^2\mathbf{I}_n\right) \tag{6}$$

$$\pi\left(\mathbf{y} \mid \mathbf{X},\mathbf{W},\alpha,\beta,\lambda,\sigma^2\right) = \mathsf{N}\left(\mathbf{y} \mid \alpha\iota_n + \mathbf{X}\beta, \sigma^2\mathbf{P}_\lambda^{-1}\right) \tag{7}$$

$$\pi\left(\alpha,\beta \mid \delta,\sigma^2\right) \propto \mathsf{L}\left(\alpha \mid \delta,\sigma\right)\prod_{j=1}^{k}\mathsf{L}\left(\beta_j \mid \delta,\sigma^2\right), \tag{8}$$

where equations (6) and (7) refer to the SAR and SEM specifications, respectively, and the Laplace prior specified in equation (8) has probability density function

$$\mathsf{L}\left(x \mid \delta,\sigma^2\right) = \frac{\delta}{2\sigma^2}\exp\left\{-\frac{\delta|x|}{\sigma^2}\right\}\mathbb{1}_{(-\infty,\infty)}\left(x\right), \tag{9}$$

which depends on the penalisation $\delta \in \mathbb{R}^+$ and scale $\sigma \in \mathbb{R}^+$ parameters. Due to its characteristics, the Laplace distribution is the Bayesian counterpart of the Lasso penalisation methodology introduced by Tibshirani (1996) to achieve sparsity within the classical regression framework. The original Bayesian Lasso, see also, e.g., Park and Casella (2008) and Hans (2009), introduces a univariate independent Laplace prior distribution for each regression parameters. The prior specification is completed by assigning a distribution to the hyper parameters $\left(\sigma^2,\delta\right)$ which controls for the scale and the Lasso penalty term. Specifically we assume an Inverse Gamma distribution for the scale parameter σ^2 and a Gamma distribution for the penalty parameter δ

$$\sigma^2 \sim \mathsf{IG}\left(\sigma^2 \mid \xi_\sigma,\eta_\sigma\right) \tag{10}$$

$$\delta \sim \mathsf{G}\left(\delta \mid \xi_\delta,\eta_\delta\right), \tag{11}$$

where $\xi_\sigma,\eta_\sigma,\xi_\delta,\eta_\delta > 0$ are given parameters. A direct characterisation of the full conditional distribution of the regression parameters of the SAR and SEM model specifications $\pi\left(\tilde{\beta} \mid \tilde{\mathbf{y}}\left(\rho\right),\mathbf{X},\mathbf{W},\rho,\sigma,\delta\right)$ and $\pi\left(\tilde{\beta} \mid \mathbf{y},\mathbf{X},\mathbf{W},\lambda,\sigma,\delta\right)$, where $\tilde{\beta} = \left(\alpha,\beta'\right)'$ that does not require the inclusion of latent variables is constructed as follows. Let $\mathsf{Z} = \{-1,1\}^{q+1}$ represent the set of all $2\left(q+1\right)$ possible $\left(q+1\right)$–vectors whose elements are ±1. For any vector $z \in \mathsf{Z}$, let $\mathsf{O} \cup \mathbb{R}^{q+1}$ represent the corresponding orthant: if $\tilde{\beta} \in \mathsf{O}_z$, then $\beta_j \geq 0$, if $z_j = 1$ and $\beta_j < 0$ if $z_j = -1$, for all $j = 1,2,\ldots,q+1$. Write the density function for the orthant–truncated Normal distribution and its associated orthant integrals as

$$\mathsf{N}^{[z]}\left(\tilde{\beta} \mid \mathbf{m},\mathbf{S}\right) = \frac{\mathsf{N}\left(\tilde{\beta} \mid \mathbf{m},\mathbf{S}\right)}{\mathsf{P}\left(z,\mathbf{m},\mathbf{S}\right)} \mathbb{1}_{\mathsf{O}_z}\left(\tilde{\beta}\right) \tag{12}$$

$$\mathsf{P}\left(z,\mathbf{m},\mathbf{S}\right) = \int_{\mathsf{O}_z} \mathsf{N}\left(\mathbf{t} \mid \mathbf{m},\mathbf{S}\right) d\mathbf{t}. \tag{13}$$

Having this notation in mind, we can characterise the full conditional distribution of the spatial regression parameters $\tilde{\beta}$ by exploiting the conjugacy between the augmented likelihood function in equations (6)–(7) and the ℓ_1–prior in equation (8) generalising Hans (2009) to the spatial regression framework.

Proposition 1. *Applying Bayes' theorem to the Lasso regression model defined in equations* (6)–(8)*, the posterior distribution in orthant–wise Normal*

$$\pi\left(\tilde{\beta}_a \mid \tilde{\mathbf{y}}\left(\rho\right), \mathbf{X}, \mathbf{W}, \rho, \sigma, \delta\right) = \sum_{z\in\mathsf{Z}} \varpi_z^s \mathsf{N}^{[z]}\left(\tilde{\beta}_a \mid \tilde{\hat{\beta}}_a^z, \Sigma_a\right) \tag{14}$$

$$\pi\left(\tilde{\beta}_e \mid \mathbf{y}, \mathbf{X}, \mathbf{W}, \lambda, \sigma, \delta\right) = \sum_{z\in\mathsf{Z}} \varpi_z^s \mathsf{N}^{[z]}\left(\tilde{\beta}_e \mid \tilde{\hat{\beta}}_e^z, \Sigma_e\right), \tag{15}$$

i.e., a finite mixture of 2^{q+1} *different truncated Normal distributions that are each restricted to a different orthant, where* $\tilde{\hat{\beta}}_s^z = \hat{\tilde{\beta}}_s - \delta\sigma^{-2}\Sigma_s\mathbf{z}$*, with* $s = \{a, e\}$ *for the SAR and SEM specifications, respectively, and*

$$\hat{\tilde{\beta}}_a = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}'\tilde{\mathbf{y}}\left(\rho\right) \tag{16}$$

$$\hat{\tilde{\beta}}_e = \left(\tilde{\mathbf{X}}'\mathbf{P}_\lambda\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}'\mathbf{P}_\lambda\mathbf{y}, \tag{17}$$

with $\Sigma_a = \sigma^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}$, $\Sigma_e = \sigma^2\left(\tilde{\mathbf{X}}'\mathbf{P}_\lambda\tilde{\mathbf{X}}\right)^{-1}$, $\tilde{\mathbf{X}} = \left[\iota_n\ \mathbf{X}\right]$, $\beta = (\alpha, \beta')'$ *and* $\varpi_z^s = \frac{\frac{\mathbb{P}\left(z,\tilde{\hat{\beta}}_s^z,\Sigma_s\right)}{\phi_{q+1}\left(0|\tilde{\hat{\beta}}_s^z,\Sigma_s\right)}}{\sum_{z\in\mathsf{Z}}\frac{\mathbb{P}\left(z,\tilde{\hat{\beta}}_s^z,\Sigma_s\right)}{\phi_{q+1}\left(0|\tilde{\hat{\beta}}_s^z,\Sigma_s\right)}}$ *where* $\phi_{q+1}\left(0 \mid \mu_z, \Sigma\right)$ *denotes the pdf of the multivariate Normal distribution with mean* μ_z *and variance–covariance matrix* Σ *evaluated at* 0*, and* $\mathbf{z} = \left(z_1, z_2, \ldots, z_n\right)'$.

4.2 Regressors selection using dirac spike–and–slab ℓ_1–prior

Using standard notation, let $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)$ be the q–dimensional vector where $\gamma_j = 1$ if the j–th covariate $\mathbf{x}_j = \left(x_{j,1}, x_{j,2}, \ldots, x_{j,n}\right)'$, for $j = 1, 2, \ldots, q$ is included as explanatory variable in the regression model and $\gamma_j = 0$, otherwise. Assuming that $\gamma_j \mid \pi_0 \sim \mathsf{Ber}\left(\pi_0\right)$, the prior distribution for β_j can be written as the mixture

$$\pi\left(\beta_j \mid \delta, \sigma^2, \pi_0\right) = \left(1 - \pi_0\right)\delta_0\left(\beta_j\right) + \pi_0\mathsf{L}\left(\beta_j \mid \delta, \sigma^2\right), \tag{18}$$

for $j=1,2,\dots,q$, where $\delta_0\left(\beta_j\right)$ is a point mass at zero and $\mathsf{L}\left(\beta_j\mid\delta,\sigma^2\right)$ denotes the Laplace density defined in equation (9). Under the spike and slab prior in equation (18), an iteration of the Gibbs sampling algorithm cycles through the full conditional distribution $\beta_j\mid\mathbf{y},\mathbf{X},\mathbf{W},\alpha,\beta_{-j},\delta,\sigma^2,\pi_0$, where β_{-j} denotes the vector of regression parameters without the j–th element, i.e., $\beta_{-j}=\left(\beta_1,\dots,\beta_{j-1},\beta_{j+1},\dots,\beta_q\right)$, for $j=1,2,\dots,q$. The next proposition provides analytical expression for the full conditional distribution of β_j, for $j=1,2,\dots,q$.

Proposition 2. *Applying Bayes' theorem to the spatial regression models defined in equations* (6)–(7) *with the Spike–and–Slab Lasso prior in equation* (18)*, the full conditional distributions of β_j, for $j=1,2,\dots,q$, is*

$$\begin{aligned}\pi\left(\beta_{j,a}\mid\mathbf{y},\mathbf{X},\mathbf{W},\alpha,\beta_{-j,a},\rho,\sigma^2,\delta,\pi_0\right)&=\varpi^0_{j,a}\left(\mathbf{y},\mathbf{X},\mathbf{W},\alpha,\beta_{-j,a},\rho,\delta,\sigma^2,\pi_0\right)\delta_0\left(\beta_{j,a}\right)\\&\quad+\left(1-\varpi^0_{j,a}\left(\mathbf{y},\mathbf{X},\mathbf{W},\alpha,\beta_{-j,a},\rho,\delta,\sigma^2,\pi_0\right)\right)\\&\quad\times\left[\left(1-\varpi_{j,a}\right)\frac{\phi\left(\beta_{j,a}\mid\hat{\beta}^-_{j,a},\sigma^2_{j,a}\right)}{\Phi_1\left(-\frac{\hat{\beta}^-_{j,a}}{\sigma_{j,a}}\right)}\mathbb{1}_{(-\infty,0)}\left(\beta_{j,a}\right)\right.\\&\quad\left.+\varpi_{j,a}\frac{\phi\left(\beta_{j,a}\mid\hat{\beta}^+_{j,a},\sigma^2_{j,a}\right)}{\Phi_1\left(\frac{\hat{\beta}^+_{j,a}}{\sigma_{j,a}}\right)}\mathbb{1}_{[0,\infty)}\left(\beta_{j,a}\right)\right],\end{aligned}\tag{19}$$

where $\sigma^2_{j,s}=\sigma^2\left(\mathbf{x}_j'\mathbf{A}_s\mathbf{x}_j\right)^{-1}$, $\mathbf{A}_a=\mathbf{I}_n$, $\mathbf{A}_e=\mathbf{P}_\lambda$, $\tilde{\varepsilon}_{j,a}=\tilde{\mathbf{y}}\left(\rho\right)-\alpha_a-\mathbf{X}_{-j}\beta_{-j,a}$, $\tilde{\varepsilon}_{j,e}=\mathbf{y}-\alpha_e-\mathbf{X}_{-j}\beta_{-j,e}$ *and*

$$\hat{\beta}^-_{j,s}=\left(\mathbf{x}_j'\mathbf{A}_s\mathbf{x}_j\right)^{-1}\left[\mathbf{x}_j'\mathbf{A}_s\tilde{\varepsilon}_{j,s}+\delta\right]\tag{20}$$

$$\hat{\beta}^+_{j}=\left(\mathbf{x}_j'\mathbf{A}_s\mathbf{x}_j\right)^{-1}\left[\mathbf{x}_j'\mathbf{A}_s\tilde{\varepsilon}_{j,s}-\delta\right]\tag{21}$$

$$\varpi_{j,s}=\frac{\varpi^+_{j,s}}{\varpi^-_{j,s}+\varpi^+_{j,s}},\tag{22}$$

with $\tilde{\mathbf{X}}_j=\left[\iota_n\ \mathbf{X}_{-j}\right]$, $\beta_{-j}=\left(\alpha,\beta_{-j}'\right)'$ *and*

$$\tilde{\varpi}_j^0 \left(\mathbf{y}, \mathbf{X}, \mathbf{W}, \alpha, \beta_{-j}, \delta, \sigma, \pi_0\right) = \left[1 + \frac{\pi_0}{(1-\pi_0)} \frac{\delta}{2\sigma} \times \left(\frac{\mathit{\Phi}\left(-\frac{\hat{\beta}_{j,s}^-}{\sigma_{j,s}}\right)}{\phi\left(0 \mid \hat{\beta}_{j,s}^-, \sigma_{j,s}^2\right)} + \frac{\mathit{\Phi}\left(\frac{\hat{\beta}_{j,s}^+}{\sigma_{j,s}}\right)}{\phi\left(0 \mid \hat{\beta}_{j,s}^+, \sigma_{j,s}^2\right)} \right)\right]^{-1}, \tag{23}$$

$$\begin{aligned}
\varpi_{j,s}^- &\equiv \varpi_{j,s}^- \left(\mathbf{y}, \mathbf{X}, \mathbf{W}, \alpha_s, \beta_{-j,s}, \delta, \sigma^2, \pi_0\right) \\
&= \int_{-\infty}^0 \pi\left(\mathbf{y} \mid \mathbf{X}, \mathbf{W}, \alpha_s, \beta_s, \sigma^2, \delta\right) \pi\left(\beta_{j,s} \mid \sigma^2, \delta\right) d\beta_{j,s} && (24) \\
&= \frac{\delta}{2\sigma^{2-p}} \frac{\mathit{\Phi}_1\left(-\frac{\hat{\beta}_{j,s}^-}{\sigma_{j,s}}\right) \phi\left(\tilde{\tilde{\mathbf{y}}}_s \mid 0, \sigma^2 \mathbf{A}_s\right)}{\exp\left\{-\frac{\tilde{\tilde{\mathbf{y}}}_s' \mathbf{A}_s \tilde{\mathbf{X}}_{-j} \tilde{\beta}_{-j,s}}{\sigma^2}\right\} \phi\left(0 \mid \hat{\beta}_{j,s}^-, \sigma_{j,s}^2\right)} && (25) \\
&\quad \times \frac{\phi_{p-1}\left(0 \mid \tilde{\beta}_{-j,s}, \sigma^2 \left(\tilde{\mathbf{X}}'_{-j} \mathbf{A}_s \tilde{\mathbf{X}}_{-j}\right)^{-1}\right)}{|\tilde{\mathbf{X}}'_{-j} \mathbf{A}_s \tilde{\mathbf{X}}_{-j}|^{\frac{1}{2}}} && (26) \\
\varpi_{j,s}^+ &\equiv \varpi_{j,s}^+ \left(\mathbf{y}, \mathbf{X}, \mathbf{W}, \alpha, \beta_{-j,s}, \delta, \sigma^2, \pi_0\right) \\
&= \int_0^{\infty} \pi\left(\mathbf{y} \mid \mathbf{X}, \mathbf{W}, \alpha, \beta_s, \sigma^2, \delta\right) \pi\left(\beta_{j,s} \mid \sigma^2, \delta\right) d\beta_{j,s} && (27) \\
&= \frac{\delta}{2\sigma^{2-p}} \frac{\mathit{\Phi}_1\left(\frac{\hat{\beta}_{j,s}^+}{\sigma_{j,s}}\right) \phi\left(\tilde{\tilde{\mathbf{y}}}_s \mid 0, \sigma^2 \mathbf{A}_s\right)}{\exp\left\{-\frac{\tilde{\tilde{\mathbf{y}}}_s' \mathbf{A}_s \tilde{\mathbf{X}}_{-j} \tilde{\beta}_{-j}}{\sigma^2}\right\} \phi\left(0 \mid \hat{\beta}_{j,s}^+, \sigma_{j,s}^2\right)} && (28) \\
&\quad \times \frac{\phi_{p-1}\left(0 \mid \tilde{\beta}_{-j,s}, \sigma^2 \left(\tilde{\mathbf{X}}'_{-j} \mathbf{A}_s \tilde{\mathbf{X}}_{-j}\right)^{-1}\right)}{|\tilde{\mathbf{X}}'_{-j} \mathbf{A}_s \tilde{\mathbf{X}}_{-j}|^{\frac{1}{2}}}, && (29)
\end{aligned}$$

and $\tilde{\tilde{\mathbf{y}}}_a = \tilde{\mathbf{y}}(\rho)$, $\tilde{\tilde{\mathbf{y}}}_e = \mathbf{y}$.

5 Results and conlusion

Overall, the obtained results show that the inclusion of the exogenous factors derived as Open Data variables is significant in all the city districts. With respect to these exogenous factors, the results show, as expected, a negative relationship between level of crime and level of prices, so that an increase in crime rate leads to a decrease in house prices in all districts. Interestingly, in the case of Airbnb diffusion, the relationship with prices is positive in Manhattan, Staten Island, Queens and Brooklyn, and negative in the Bronx district. This finding suggests that, depending on the district considered, Airbnb may be seen as a threat or an opportunity for the market. Indeed, the significantly negative relationship in Bronx may be taken as a sign that the presence of Airbnb can lead to an upgrading of the entire district, through a more accessible real estate market.

References

George, E. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Bayesian Quantile Trees for Sales Management

Mauro Bernardi and Paola Stolfi

Abstract Sales management is a fundamental issue in retail commerce, being one of the main ingredient to maximise stores revenue. The huge amount of data accessible nowadays makes this task challenging. This paper proposes a new nonparametric method for predicting sales quantiles at different confidence levels, conditional on several explanatory variables such as the type of store, the location of the store, the type of product and its the price, etc, thereby providing a complete picture of the relation between the response and the covariates. Moreover, predicting extreme sales quantiles provide valuable information for building automatic stock management systems and for the sales monitoring. As concerns the methodology, we propose to approximate the conditional quantile at level $\tau \in (0,1)$ of the response variable using bayesian additive nonparametric regression trees. Decision trees and their additive counterparts are promising alternatives to linear regression methods because of their superior ability to characterise nonlinear relationships and interactions among explanatory variables that is of fundamental relevance to get accurate predictions.

Key words: Regression tree, bayesian methods, quantile regression, prediction.

1 Introduction

In empirical studies, researchers are often interested in analysing the behaviour of a response variable given the information on a set of covariates. The typical answer is to specify a linear regression model where unknown parameters are estimated using the Ordinary Least Squares (OLS). The OLS method estimates unknown parameters by minimising the sum of squared errors leading to the approximation of the mean function of the conditional distribution of the response variable. Although the mean represents the average behaviour of the response variable, it provides little or no information about the behaviour of the conditional distribution on the tails. As far as the entire distribution is concerned, quantile regression methods [Koenker and Bassett, 1978] adequately characterise the behaviour of the response variable at different con-

Mauro Bernardi
Department of Statistical Sciences, University of Padova, Via Cesare Battisti, 241, 35121, Padova, e-mail: mauro.bernardi@unipd.it

Paola Stolfi
Institute for applied mathematics "Mauro Picone" (IAC) - CNR, Rome, Italy, e-mail: p.stolfi@iac.cnr.it

fidence levels. Moreover, the quantile analysis is particularly suitable when the conditional distribution is heterogeneous, non–Gaussian, skewed or fat–tailed, see, e.g., [Lum and Gelfand, 2012] and [Koenker, 2005].

Quantile models admitting a linear representation have been extensively applied in different areas, see, e.g., [Yu et al., 2003], such as, finance, as direct approach to estimate the Value–at–Risk, i.e., the loss–level a financial institution may suffer with a given confidence [Bassett Jr. and Chen, 2001], economics and social sciences [Hendricks and Koenker, 1992], medicine [Heagerty and Pepe, 1999], survival analysis [Koenker and Geling, 2001] and environmetrics [Pandey and Nguyen, 1999]. Futhermore, linear quantile models have been theoretically investigated from both a Bayesian, [Sriram et al., 2013] and a frequentist point of view, and the properties of the resulting estimates has been deeply studied. See [Koenker et al., 2017] and [Davino et al., 2014] for an extensive and up to date review latest theoretical results on quantile methods and their interesting applications. However, despite their relevance and widespread application in empirical studies, linear quantile models provide only a rough "first order" approximation of the relationship between the τ–level quantile of the response variable and the covariates. Indeed, as first recognised by [Koenker, 2005], quantiles are linear functions only within a Gaussian world, thereby stimulating many recent attempts to overcome this limitation. [Chen et al., 2009], [Faugeras, 2009], [De Backer et al., 2017] and [Kraus and Czado, 2017], for example, consider the copula–based approach to formalise nonlinear and parametric conditional quantile relationships. The copula approach, although quite flexible in fitting marginal data, forget to consider nonlinear interactions among the covariates. This paper try to overcome the traditional limitations of linear quantile methods by extending the quantile approach to the promising field of decision trees. Decision trees are regression techniques that are very popular within the machine learning community that try to mitigate the relevant problem of parsimoniously modelling interactions among covariates. Indeed, decision trees partition the space of relevant covariates into pieces, usually hyper–rectangles, where observations are homogeneous. For example, when the objective is to model the average response, observations with the same unconditional variance are clustered together, while for classification problems observations are partitioned according to the Gini index. Since their introduction several theoretical and applied papers have contributed to the diffusion of such idea in different contexts. Both the machine learners and the statistics communities contributed to the development of tools and methods. An up to date and comprehensive review of the methods developed in the machine learning literature can be found for example in [Loh, 2014a], [Strobl, 2014], [Ciampi, 2014], [Ahn, 2014], [Song and Zhang, 2014], [Rusch and Zeileis, 2014], [Loh, 2014b]. The main drawback of decision trees is related to the high variance of the resulting forecasts. The most promising alternative approach [Breiman, 2001], namely, the random forest, is an ensemble of bootstrap decision trees that reduces the variance and provides also a straightforward way to assess the relevant covariates. On the likelihood–based side, the Bayesian estimation of decision trees have been considered in [Chipman et al., 1998], [Denison et al., 1998], [Sha, 2002] and [Wu et al., 2007] and extended to additive trees by [Chipman et al., 2010]. The main novelty of this latter approach relies on exploiting the likelihood of parametric models where regressors splitting rules play the role of hard thresholding operators that partition the overall model into local models.

In this paper, we consider the Bayesian approach, namely, we extend the Bayesian Additive Regression Tree (BART) of [Chipman et al., 2010] to model the quantile of the response variable. The Bayesian Additive Quantile Regression Trees (BAQRT) exploits the data augmentation approach that relies on the Asymmetric Laplace working likelihood, see [Bernardi et al., 2015], to provide a Metropolis–within–Gibbs sampling method that efficiently explores the regressors space. The data augmentation approach allows to effectively marginalise out the leaf parameters of the trees when changing the tree structures. Section 2 formalises the likelihood function and the prior structure of the BAQRT method. Quantile random forest (QRF) methods have been previously introduced in the machine learning literature by [Meinshausen, 2006]. [Meinshausen, 2006] exploits the original version of random forest for modelling the conditional mean to infer the structure of the tree, while assigning the empirical quantile of the observations falling into each terminal leaf instead of the mean value. Therefore, the QRF algorithm is not highly flexible to adapt the structure of the generated trees according to the modelled quantiles.

Given their appealing perspective of discriminating observations below and above a given quantile threshold, the quantile approach is a promising method for solving many statistical problems usually encountered in business and industry. In Section 3 we apply BAQRT to strategic sales management in retail stores. Sales management is one of the main issues in retail commerce. Indeed, it is one of the main ingredient to maximise the income of stores. The huge amount of data accessible nowadays makes this task challenging. Indeed, there are many potentially predictors that could be useful to predict and monitor sales but there is not any model to reach this task. This is the usual situation in which machine learning algorithms represent a powerful instruments to extract insight from such heterogeneous data. We consider the dataset provided by BigMart, an international brand with both free home delivery services and outlet store of food and grocery, to show how quantile regression trees could be useful in selecting the most relevant variables and analyse their impact both for predicting and monitoring tasks.

2 Quantile regression tree

The linear quantile regression framework for independent and identically distributed data models the conditional τ–level quantile of the response variable Y, with $\tau \in (0,1)$, as a linear function of the vector of dimension $(q \times 1)$ of exogenous covariates $\mathbf{X}$, i.e., $\mathcal{Q}_\tau(Y \mid \mathbf{X} = \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, thereby avoiding any explicit assumptions about the conditional distribution of $Y \mid \mathbf{X} = \mathbf{x}$. From a frequentist perspective, within a likelihood–based framework, this is equivalent to assume an additive stochastic error term ε for the conditional regression function $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ to be independent and identically distributed with zero τ–th quantile, i.e, $\mathcal{Q}_\tau(\varepsilon \mid \mathbf{x}) = 0$, and constant variance. Following [Yu and Moyeed, 2001] and [Bernardi et al., 2015], the previous condition is implicitly satisfied by assuming that the conditional distribution of the response variable Y follows an Asymmetric Laplace (AL) distribution located at the true regression function $\mu(\mathbf{x})$, with constant scale $\sigma > 0$ and shape parameter τ, i.e., $\varepsilon \sim \mathsf{AL}(\tau, \mu(\mathbf{x}), \sigma)$, with probability density function

$$\mathsf{AL}(Y \mid \mathbf{X}, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\frac{1}{\sigma}\rho_\tau(Y - \mu(\mathbf{x}))\right\} \boldsymbol{I}_{(-\infty,\infty)}(Y), \tag{1}$$

where $\mu(\mathbf{x})$ is the regression function and $\rho_\tau(u) = u\left(\tau - \boldsymbol{I}_{(-\infty,0)}(u)\right)$ denotes the quantile check function at level τ. The quantile regression model postulated in equation (1) assumes the AL distribution as a misspecified working likelihood that correctly identify the conditional quantile function.

Similarly to the Bayesian Additive Regression Tree approach of [Chipman et al., 2010] for modelling the conditional mean of the response variable, the quantile regression tree approach extends the linear quantile model defined in equation (1) by assuming a sum–of–trees ensemble for the regression function $\mu(\mathbf{x})$. Specifically, the Bayesian Additive Quantile Regression Tree (BAQRT) model can be expressed as

$$\begin{align} Y &= \mu(\mathbf{x}) + \varepsilon \tag{2}\\ &\approx \mathscr{T}_1^{\mathscr{M}}(\mathbf{x}) + \mathscr{T}_2^{\mathscr{M}}(\mathbf{x}) + \cdots + \mathscr{T}_m^{\mathscr{M}}(\mathbf{x}) + \varepsilon, \tag{3}\end{align}$$

where $\varepsilon \sim \mathsf{AL}(\tau, 0, \sigma)$. The assumption about the error term in equation (3) implies that $\mu(\mathbf{x}) = \mathcal{Q}_\tau(Y \mid \mathbf{X} = \mathbf{x})$. Furthermore, in equation (3) we assume that the quantile of the response variable is an additive function of

$m \geq q$ regression trees, each composed by a tree structure, denoted by $\mathscr{T}$, and the parameters of the terminal nodes (also called leaves), denoted by $\mathscr{M}$. Therefore, the j–th tree for $j = 1,2,\dots,m$, denoted by $\mathscr{T}_j^{\mathscr{M}}$, represents a specific combination of tree structure $\mathscr{T}_j$ and tree parameters $\mathscr{M}$, i.e., the regression parameters associated to its terminal nodes. The tree structure $\mathscr{T}_j$ contains information on how any observation y_i, in a set of n independent and identically distributed observations $\mathbf{y} = (y_1, y_2, \dots, y_n)$, recurses down the tree specifying a splitting rule for each non–terminal (internal) node. The splitting rule has the form $x_k \leq c$ and consists of the splitting variable x_k and the splitting value $c \in \mathbb{R}$. The observation y_i is assigned to the left child if the splitting rule is satisfied and to the right child, otherwise, until a terminal node is reached and the value of the leaf of that terminal node is assigned as its predicted value. Therefore, the quantile prediction corresponding to y_i assigned by the sum of regression tree specified in equation (3) is the sum of the m leaf values. Hereafter, we denote by $\mathscr{M}_j = \left\{\mu_{j,1}, \mu_{j,2}, \dots, \mu_{j,b_j}\right\}$ the set of parameters associated to the b_j terminal nodes of the j–th tree, where $\mu_{j,l}$, for $l = 1,2,\dots,b_l$ denotes the conditional quantile predicted by the model.

The additive quantile regression tree specified in equation (3) provides a natural framework for likelihood–based inference on the set of quantile regression parameters, i.e., the location parameters associated to the terminal nodes of each tree belonging to the ensemble. However, additional prior information should be imposed in order to infer the structure of the each tree. The next Section discusses the likelihood and the prior structure for both the model parameters and the trees.

2.1 Likelihood and prior

As discussed in [Yu and Moyeed, 2001], due to the complexity of the quantile likelihood function in equation (1), the resulting posterior density for the regression parameters does not admit a closed form representation for the full conditional distributions, and needs to be sampled by using MCMC–based algorithms. Following [Kozumi and Kobayashi, 2011] and [Bernardi et al., 2015], we instead adopt the well–known representation (see, e.g., [Kotz et al., 2001] and [Park and Casella, 2008]) of $\varepsilon \sim \mathsf{L}(\tau, 0, \sigma)$ as a location–scale mixture of Gaussian distributions:

$$\varepsilon = \zeta\omega + \varsigma\sqrt{\sigma\omega}\epsilon, \tag{4}$$

where $\omega \sim \mathsf{Exp}\left(\sigma^{-1}\right)$ and $\epsilon \sim \mathsf{N}(0,1)$ are independent random variables and $\mathsf{Exp}(\cdot)$ denotes the Exponential distribution. Moreover, the parameters ζ and ς^2 are fixed equal to

$$\zeta = \frac{1-2\tau}{\tau(1-\tau)}, \qquad \varsigma^2 = \frac{2}{\tau(1-\tau)}, \tag{5}$$

in order to ensure that the τ–th quantile of ε is equal to zero. The previous representation in equation (4) allows us to use a Gibbs sampler algorithm for sampling the trees parameters $\boldsymbol{\mu}_j$ of tree $j = 1,2,\dots,m$, detailed in the next subsection. Exploiting the augmented data structure defined in equation (4), the additive quantile regression tree in (3) admits, conditionally on the latent factor ω, the following Gaussian representation:

$$\begin{aligned} Y \mid \omega &= \mu(\mathbf{x}) + \zeta\omega + \varsigma\sqrt{\sigma\omega}\epsilon, && (6)\\ &\approx \mathscr{T}_1^{\mathscr{M}}(\mathbf{x}) + \mathscr{T}_2^{\mathscr{M}}(\mathbf{x}) + \cdots + \mathscr{T}_m^{\mathscr{M}}(\mathbf{x}) + \zeta\omega + \varsigma\sqrt{\sigma\omega}\epsilon && (7)\\ \omega &\sim \mathsf{Exp}\left(\sigma^{-1}\right). && (8)\end{aligned}$$

The hierarchical model representation in equations (6)–(8) has the advantage of being conditionally Gaussian, leading to a conjugate Bayesian analysis for the parameters associated to the terminal nodes and the scale parameter as well.

The Bayesian inferential procedure requires the specification of the prior distribution for the unknown vector of model parameters $(\boldsymbol{\mu},\sigma)$ and the structure of the tree. In principle, as discussed in the seminal paper of [Yu and Moyeed, 2001], non informative priors can be specified for the vector of regression parameters, i.e., $\pi(\boldsymbol{\mu}) \propto 1$. Alternatively, as in [Bernardi et al., 2015], the usual Normal–Inverse Gamma prior can be specified for regression and scale parameters, respectively, i.e.,

$$\mu_i \sim \mathsf{N}_1\left(\mu_0, \sigma_\mu^2\right) \tag{9}$$

$$\sigma \sim \mathsf{IG}\left(\frac{\eta_\sigma}{2}, \frac{\eta_\sigma \lambda_\sigma}{2}\right) \tag{10}$$

$$\mathbb{P}(\mathscr{T}) \propto \alpha(1+d)^{-\beta}, \tag{11}$$

where $\alpha \in (0,1)$ and $\beta \in [0,\infty)$ and d is the depth of the tree defined as the distance from the root. Here, N_1 denotes the univariate Normal density while IG is the Inverse Gamma distribution and $\left(\mu_0, \sigma_\mu^2, \eta_\sigma, \lambda_\sigma\right)$ are fixed hyperparameters, with $\eta_\sigma > 0$ and $\lambda_\sigma > 0$.

Now, let $\mathbf{y} = (y_1, y_2, \dots, y_n)$ be the vector of observations on the response variable Y and let $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ be the associated matrix of covariates of dimension $(n \times q)$, then the joint prior distribution of the tree structure and tree parameters augmented by the latent factor ω, can be factorised as follows:

$$\begin{aligned}
\mathbb{P}\left(\mathscr{T}_1^{\mathscr{M}}, \mathscr{T}_1^{\mathscr{M}}, \dots, \mathscr{T}_1^{\mathscr{M}}, \boldsymbol{\omega}, \sigma\right) &= \left[\prod_{j=1}^{m} \mathbb{P}\left(\mathscr{T}_j^{\mathscr{M}} \mid \boldsymbol{\omega}\right)\right] \mathbb{P}(\sigma) \\
&= \left[\prod_{j=1}^{m} \mathbb{P}(\mathscr{M}_j \mid \mathscr{T}_j, \boldsymbol{\omega}) \mathbb{P}(\mathscr{T}_j \mid \boldsymbol{\omega})\right] \mathbb{P}(\boldsymbol{\omega}) \mathbb{P}(\sigma) \\
&= \left[\prod_{j=1}^{m} \prod_{l=1}^{b_j} \mathbb{P}\left(\mu_{j,\ell} \mid \mathscr{T}_j, \boldsymbol{\omega}\right) \mathbb{P}(\mathscr{T}_j \mid \boldsymbol{\omega})\right] \mathbb{P}(\boldsymbol{\omega}) \mathbb{P}(\sigma) \\
&= \left[\prod_{j=1}^{m} \prod_{l=1}^{b_j} \phi\left(\mu_{j,\ell} \mid \mathscr{T}_j, \boldsymbol{\omega}\right) \mathbb{P}(\mathscr{T}_j \mid \boldsymbol{\omega})\right] \mathbb{P}(\boldsymbol{\omega}) \mathbb{P}(\sigma),
\end{aligned} \tag{12}$$

where $\mu_{j,\ell}$ denotes the parameter (conditional quantile) associated to the ℓ–th terminal nodes of tree $j = 1, 2, \dots, m$, for $\ell = 1, 2, \dots, b$, $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_n, \sigma)$ is the vector of auxiliary variables $\omega_i \sim \mathsf{Exp}\left(\sigma^{-1}\right)$, $\sigma \in \mathbb{R}$ is the scale parameter which is common to all the trees, and $\phi(\cdot)$ denotes the gaussian probability density function.

3 Application

Retail stores invest much effort in high level strategy to maximise their income. The type of store, its location, the furnitures and the product proposal are some of the main ingredients driving strategic decisions. Sales product prediction is therefore one of the most challenging problem in retail commerce, fundamental for

Variable Name	Description
Item Identifier	Unique product ID
Item Weight	Weight of product
Item Fat Content	Whether the product is low fat or not
Item Visibility	The % of total display area of all products in a store allocated to the particular product
Item Type	The category to which the product belongs
Item MRP	Maximum Retail Price of the product (list price)
Outlet Identifier	Unique store ID
Outlet Estabilishment Year	The year in which store was estabilished
Outlet Size	The size of the store in terms of ground area covered
Outlet Location Type	The type of city in which the store is located
Outlet Type	Whether the outlet is just a grocery store or some sort of supermarket
Item Outlet Sales	Sales of the product in the particular store. This is the outcome variable.

Table 1: Variables and their description for the BigMart dataset.

instance for commercialising new products, opening new stores or monitor the income performance of the stores.

In this work we apply BAQRT to analyse data coming from BigMart. It is an international brand with both free home delivery services and outlet store of food and grocery. Data scientists at BigMart created a dataset containing sales for 1559 products in 10 different stores located in several cities. They also reported many features related both to products and stores, in table 1 we provide a list of all the variables together with a description. In particular, there are eleven predictors and a scalar response function that is the "Outlet Sales". Eight of the eleven predictors are categorical while the remaining are continuous. The main interest consists in the identification of the variables that mostly influence the sales and if the relevance of these features changes by considering the different quantiles of the response variable. Indeed, the tail behaviour of the sales is also useful in stock management.

The variables considered in this example are a small subset of all the possible variables that could be analysed by retail stores. The importance of the variables are however based on some hypothesis and there is not any model that can be used both for sales prediction and monitoring tasks. This motivates the application of our method to investigate the type of relations occurring between the predictors and the response variable.

In figure 1 we report the predictor importance for the confidence levels $\tau = (0.1, 0.5, 0.9)$. The first two quantiles show similar variables importance ranking, in particular, the most relevant variables are "Item MRP", "Item Type", "Item Visibility" and "Item Weight". The first variable represents the price, its relevance supports the fact that promotional offers makes the customers much more inclined to buy products. The second one is the "Item Type" that represents the category to which the product belongs to. This finding supports the idea that one of the main indicator of product's sale is its utility, that is daily use products have much more sales rate than others. The third one, the "Item Visibility", confirms the fact that the position of the products in the stores is fundamental for their sales. Finally, the last "Item Weight" supports the hypothesis that lighter (and often smaller) products are easier to carry and so people are more inclined to buy them even if they are not needed.

The higher quantile instead show a different situations for the predictors' importance. Indeed, the "Item Visibility" becomes the most relevant variable, while all the others have a quite similar importance.

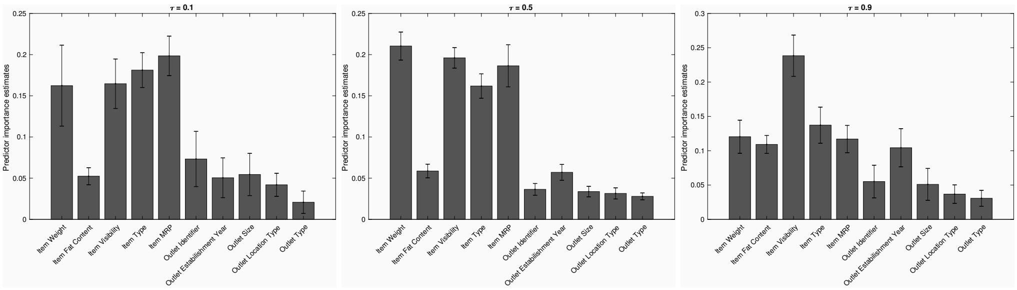


Fig. 1: Predictors' importance for different quantiles. The bottom figure in the left column refers to the mean behaviour.

References

[Ahn, 2014] Ahn, H. (2014). Discussion: "Fifty years of classification and regression trees" [mr3280974]. *Int. Stat. Rev.*, 82(3):357–359.

[Bassett Jr. and Chen, 2001] Bassett Jr., G. W. and Chen, H.-L. (2001). Portfolio style: Return-based attribution using quantile regression. *Empirical Economics*, 26(1):293–305.

[Bernardi et al., 2015] Bernardi, M., Gayraud, G., and Petrella, L. (2015). Bayesian tail risk interdependence using quantile regression. *Bayesian Anal.*, 10(3):553–603.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Chen et al., 2009] Chen, X., Koenker, R., and Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *Econometrics Journal*, 12:S50–S67.

[Chipman et al., 1998] Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948.

[Chipman et al., 2010] Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298.

[Ciampi, 2014] Ciampi, A. (2014). Discussion: "Fifty years of classification and regression trees" [mr3280974]. *Int. Stat. Rev.*, 82(3):352–357.

[Davino et al., 2014] Davino, C., Furno, M., and Vistocco, D. (2014). *Quantile regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. Theory and applications.

[De Backer et al., 2017] De Backer, M., El Ghouch, A., and Van Keilegom, I. (2017). Semiparametric copula quantile regression for complete or censored data. *Electron. J. Statist.*, 11(1):1660–1698.

[Denison et al., 1998] Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika*, 85(2):363–377.

[Faugeras, 2009] Faugeras, O. P. (2009). A quantile-copula approach to conditional density estimation. *Journal of Multivariate Analysis*, 100(9):2083 – 2099.

[Heagerty and Pepe, 1999] Heagerty, P. J. and Pepe, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):533–551.

[Hendricks and Koenker, 1992] Hendricks, W. and Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, 87(417):58–68.

[Koenker, 2005] Koenker, B. (2005). *Quantile Regression*. Cambridge University Press, Cambridge.

[Koenker and Bassett, 1978] Koenker, R. and Bassett, Jr., G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

[Koenker et al., 2017] Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017). *Handbook of Quantile Regression*. Chapman & Hall / CRC Handbooks of Modern Statistical Methods. CRC Press, Taylor & Francis Group.

[Koenker and Geling, 2001] Koenker, R. and Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *J. Amer. Statist. Assoc.*, 96(454):458–468.

[Kotz et al., 2001] Kotz, S., Kozubowski, T., and Podgórski, K. (2001). *The Laplace distribution and generalizations. A revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhäuser.

[Kozumi and Kobayashi, 2011] Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81:1565–1578.

[Kraus and Czado, 2017] Kraus, D. and Czado, C. (2017). D-vine copula based quantile regression. *Computational Statistics & Data Analysis*, 110:1 – 18.

[Loh, 2014a] Loh, W.-Y. (2014a). Fifty years of classification and regression trees. *Int. Stat. Rev.*, 82(3):329–348.

[Loh, 2014b] Loh, W.-Y. (2014b). Rejoinder: "Fifty years of classification and regression trees" [mr3280975; mr3280976; mr3280977; mr3280978; mr3280979; mr3280974]. *Int. Stat. Rev.*, 82(3):367–370.

[Lum and Gelfand, 2012] Lum, K. and Gelfand, A. (2012). Spatial quantile multiple regression using the asymmetric laplace process. *Bayesian Analysis*, 7:235–258.

[Meinshausen, 2006] Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

[Pandey and Nguyen, 1999] Pandey, G. and Nguyen, V.-T.-V. (1999). A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, 225(1):92 – 101.

[Park and Casella, 2008] Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686.

[Rusch and Zeileis, 2014] Rusch, T. and Zeileis, A. (2014). Discussion: "Fifty years of classification and regression trees" [mr3280974]. *Int. Stat. Rev.*, 82(3):361–367.

[Sha, 2002] Sha, N. (2002). *Bolstering CART and Bayesian variable selection methods for classification*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Texas A&M University.

[Song and Zhang, 2014] Song, C. and Zhang, H. (2014). Discussion: "Fifty years of classification and regression trees" [mr3280974]. *Int. Stat. Rev.*, 82(3):359–361.

[Sriram et al., 2013] Sriram, K., Ramamoorthi, R., and Ghosh, P. (2013). Posterior consistency of bayesian quantile regression based on the misspecified asymmetric laplace density. *Bayesian Analysis*, 8:479–504.

[Strobl, 2014] Strobl, C. (2014). Discussion: "Fifty years of classification and regression trees" [mr3280974]. *Int. Stat. Rev.*, 82(3):349–352.

[Wu et al., 2007] Wu, Y., Tjelmeland, H. k., and West, M. (2007). Bayesian CART: prior specification and posterior simulation. *J. Comput. Graph. Statist.*, 16(1):44–66.

[Yu et al., 2003] Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. *The Statistician*, 52(3):331–350.

[Yu and Moyeed, 2001] Yu, K. and Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54:437–447.

Discrimination in machine learning algorithms

Discriminazione negli algoritmi di apprendimento automatico

Roberta Pappadà and Francesco Pauli

Abstract Machine learning algorithms are routinely used for business decisions which may directly affect individuals: for example, because a credit scoring algorithm refuses them a loan. It is then relevant from an ethical (and legal) point of view to ensure that these algorithms do not discriminate based on sensitive attributes (sex, race), which may occur unwittingly and unknowingly by the operator and the management. Statistical tools and methods are then required to detect and eliminate such potential biases.

Abstract *Sempre più decisioni, nei campi più vari, sono prese impiegando algoritmi a supporto o in sostituzione dell'intervento umano. Queste decisioni possono avere un effetto sulle persone che le subiscono, ad esempio quando un algoritmo di valutazione di solvibilit decide di rifiutare un prestito. Diventa quindi rilevante eticamente (e legalmente) assicurarsi che questi algoritmi non basino la loro decisione anche su caratteristiche sensibili, cosa che può avvenire senza intento e consapevolezza da parte del responsabile. Si apre dunque la necessità di studiare strumenti e metodi statistici per individuare e eventualmente eliminare queste potenziali distorsioni.*

Key words: machine learning; CEM; protected categories; sensible attribute

1 Introduction

The kind of discrimination we refer to consists in treating a person or a group depending on some sensitive attribute (s.a., *S*) such as race (skin color), sex, religious orientation, etc.

A human may discriminate either because of irrational prejudice induced by ignorance and stereotypes or based on statistical generalization: lacking specific in-

Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy, e-mail: rpappada@units.it, francesco.pauli@deams.units.it

formation on an individual, he is assigned the characteristics which are prevalent in the sensitive attribute category he belongs to (for example in the US lacking information on education, a black person may be assumed to have relatively low level since this is the case in general for black people in the country) [7]. When a statistical/machine learning algorithm is used in the decision process, its behavior with respect to discrimination depends on the information it is given. In particular, if the sensitive attribute is available to the algorithm (that is, it is included in the learning data and can be used for predictions), it may discriminate either because the data it is taught by contain irrational prejudice (Fig. 1*(a)*) and because the sensitive attribute is associated to an unobserved attribute which is relevant for prediction of Y, the outcome of interest (Fig. 1*(b)*). An example of the former may occur if an algorithm is sought for screening job applicants and the learning data consist of selections made in the past by humans who, in some instances, based their decisions on irrational prejudice, consciously or not; as an example of the latter consider a credit scoring algorithm where the education of applicant is unknown but the race is, then due to the fact that education is relevant to assess reliability and is associated to the race the latter may be used in the decision. For an actual example consider car insurance (RCA), where the sex of the insured used to be a relevant factor in pricing likely because it is related to the number of km driven per year which is an unobserved variable strongly related to the outcome (liabilities). We note in pass that, although this is a rare circumstance, it may also be the case that the sensitive attribute is directly related to the outcome not because of human irrational prejudice, as it happens in life insurance for the gender attribute.

If the algorithm is constrained not to use the s.a. in the final rule (it is excluded from the data from which the algorithm learns) discrimination may still occur indirectly, either because a variable which is related to the outcome is also related to S (Fig. 1*(c)*) or because a variable included in the data which is unrelated to the outcome is related to S which is related to the outcome (either directly as in Fig. 1*(d)* or mediated as in Fig. 1*(e)* or in both ways as in Fig. 1*(f)*). In the first case the collective possessing the s.a. may experience the desired outcome less frequently than the rest whenever it differs from the population in the distribution of some relevant features included in the data. In the second case the collective possessing the s.a. may experience the desired outcome less frequently than the rest even if it were equal to the population as far as the other relevant features included in the data are concerned; in this second situation the discrimination may be due to a difference in the distribution of some relevant feature not included in the data but related to S and some variable in the sample.

Whether the situations described above are instances of undesired discrimination or not depends on how the concept is defined from a legal/ethical standpoint. In particular it depends on whether we require that the groups identified by S have the same treatment (rate of positive outcome) unconditionally or that they have the same treatment conditionally on some relevant characteristics different than S and which are deemed lawful to use to discriminate.

Which definition is more appropriate is a matter of ethic/law: a requirement of unconditional equal treatment means that the groups must be given equal treatment even if they are not equal (demographic parity, disparate impact), which may seem unjust from an individual point of view [2]. On the other hand, admitting the groups to be treated differently because one of them possesses desirable characteristics possibly amounts to perpetuating past unfair (S based) discrimination which may have lead the groups to be different. In fact, the first requirement may be an instance of affirmative action since it goes in the direction of eliminating the differences between groups, differences which are at least partly admissible within the second requirement. If conditional equal treatment is sought, one must decide which characteristics other than S but possibly correlated with it are ethical/lawful to use, which may be partly dictated by law, partly uncertain (for example in a civil law system such as in the US whether a characteristic is lawful to discriminate on may be for a jury to decide).

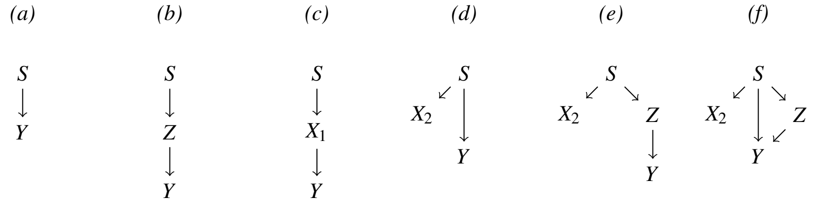


Fig. 1 Y is the outcome, S is the sensitive attribute, X_i denotes observed variables, Z denotes an unobserved variable. Example: S: race, Y: restitution of a loan, X_1: socioeconomic status, X_2: zip code residence, Z availability of family financial support.

In order to give precise definitions, from now on assume both S and Y dichotomous, $Y = 1$ be the desired outcome and $S = 1$ the belonging to a protected category. If unconditional equal treatment (demographic parity) is desired, data are apt if it is not possible to predict S from Y [1], that is, data are compatible with

$$P(Y = 1|S = 1) = P(Y = 1|S = 0).$$

If equal treatment should be conditional on X_1 then the requirement becomes

$$P(Y = 1|S = 1, X_1 = x_1) = P(Y = 1|S = 0, X_1 = x_1).$$

An extreme version of conditional equal treatment is advocated in [2], a rule is non discriminatory if the prediction errors are the same regardless of S

$$P(\hat{Y} = 1|S = 0, Y = y) = P(\hat{Y} = 1|S = 1, Y = y), \ \forall y;$$

or, which is the same, prediction and sensitive attribute are independent conditional on the outcome.

2 Measuring and avoiding discrimination by causal inference

The different strategies which can be used to avoid discrimination from an algorithm can be distinguished depending on the level at which action is taken: 1) learning data can be modified to ensure they do not imply discrimination; 2) the learning algorithm can be integrated with a non discrimination objective; 3) the final algorithm can be tampered with after it has been fitted; 4) the final predictions can be changed. The precise action to be taken depends on the definition of discrimination adopted; as outlined in the previous section, we may seek unconditional equality or conditional equality.

We focus on data preprocessing techniques. In a nutshell, this entails first establishing whether and to what extent the available data are discriminatory. If discrimination is detected data are modified in order to make them discrimination free before and then used with a standard algorithm [5]. Various ways of modifying data have been proposed including: 1) removing attributes (correlated with S); 2) changing the labels of some units; 3) weighing units; 4) resampling units We note that the precise implementation for both steps depends on the definition of discrimination which is adopted, presumably on legal/ethical grounds (see section 1).

In order to measure discrimination within a dataset it has been proposed to use causal inference techniques. Causal inference techniques aim at estimating the causal effect of a treatment–in this context, the belonging to a protected category–on an outcome. Generally speaking, causal inference methods aim at assessing the effect of a treatment based on observational data by matching treated units to untreated units which are similar with respect to their observed characteristics: a balanced dataset, suitable to draw causal inference, is built by restricting the original dataset to treated and untreated units which have been matched. Comparing a protected units outcome with the outcome of unprotected units which are similar with respect to the other observables is also a reasonable way to detect whether the unit has been discriminated (conditionally).

Following this idea, Luong *et al.* [6] propose to measure discrimination for unit (individual) i of the protected category ($S_i = 1$) as

$$\Delta_i = \frac{1}{k}\#\{j|j \neq i, x_j \in U_i^{k,1}, y_j = y_i\} - \frac{1}{k}\#\{j|j \neq i, x_j \in U_i^{k,0}, y_j = y_i\} \qquad (1)$$

where $U_i^{k,s}$ is the set of the k nearest neighbours of x_i within those units for which $S = s$, according to a Gower type distance (possibly, other measure of the difference between the two frequencies are used such as the ratio or the odds). Note that a positive Δ indicates discrimination against the protected category if y_i is the undesired outcome or discrimination in favour of the protected category if y_i is the desired outcome. The authors suggest fixing a threshold $\tau \in [0,1]$ to declare the individual i discriminated if $\Delta_i \geq \tau$, where τ. To prevent discrimination the dataset is changed by altering the value of y_i for those units fo which $\Delta_i > \tau$. τ is then a tuning parameter which regulates the trade off between residual discrimination and accuracy.

In what follows we focus on discrimination against the protected group, and so we compute a different version of the measure Δ:

$$\delta_i = \frac{1}{k}\#\{j|j \neq i, x_j \in U_i^{k,1}, y_j = 0\} - \frac{1}{k}\#\{j|j \neq i, x_j \in U_i^{k,0}, y_j = 0\} \qquad (2)$$

3 CEM based discrimination measure

Coarsened Exact Matching (CEM, [4, 3]) is based on coarsening continuous variables and match a treated unit to those untreated units which are equal with respect to the coarsened continuous variables and the categorical ones. In order to obtain a reliable estimate of the causal effect, CEM algorithm discards those units which can not be matched. Here, the objective is different in that we are not interested in a global estimate of the effect of the S, but rather whether unit i has experienced a different outcome because of it possesses the S: a comparison of unit i outcome with the outcome of units matched by CEM is a suitable measure of discrimination. In particular, let $\bar{y}_i^{(S=0)}$ be the relative frequency of positive outcomes among units matched with unit i (which possesses the s.a.) and not possessing the s.a., then $D_i = y_i - \bar{y}_i^{(S=0)}$ is a measure of discrimination which takes negative values when a unit is discriminated against and positive values when the unit is favoured (positive discrimination), so that the value of D_i is bounded between -1 and 1. However, in a standard implementation of CEM algorithm not all units are matched, which makes it unsuitable for our purpose. A possible strategy to exploit CEM technique is to apply it sequentially as follows. Suppose that all units are matched using k variables, it is possible that once an additional variable is considered in matching some units remain unmatched, we then measure the discrimination for those units based on the matching with k variables. Starting with an initial matching on 0 variables, that is, $D_i^{(0)} = y_i - \bar{y}^{(S=0)}$ where $\bar{y}^{(S=0)}$ is the relative frequency of the positive outcome for all units in the dataset not possessing the s.a., the sequential CEM allows to obtain a discrimination measure for all units. Clearly, the discrimination measures D_i depend on the order of addition of the variables, so the procedure is repeated for different (random) orders of addition and the final result is the average (See Fig. 2).

An alternative measure of discrimination based on CEM stratification which is more similar to (2), $\bar{D}_i$ in what follows is obtained by computing $\bar{D}_i = \bar{y}_i^{(S=1)} - \bar{y}_i^{(S=0)}$ instead of $\bar{D}_i$ (with obvious adaptations in the algorithm of Fig. 2). Note that, unlike D_i, $\bar{D}_i$ takes positive values when the unit is discriminated against (similar to δ).

4 Simulation experiment

We tested the procedure using the *adult* dataset as in [6]. The outcome is having an income greater than 50 000 USD, the sensitive attribute is being non white, other

```
Let
  • $x_1, \dots, x_K$ be the available variables;
  • let $\mathscr{M}^{(j_1,\dots,j_h)}$ be the set of units matched by CEM performed using variables $x_{j_1}, \dots, x_{j_h}$
    and let $C_i^{(j_1,\dots,j_h)}$ be the set of units belonging to the same CEM cell of unit $i$.
Repeat $M$ times

$-1$: select a random permutation $i_1, \dots, i_K$ of $1, \dots, K$;
  0: for all units $i$ let $D_i^{(0)} = y_i - \bar{y}$ where $\bar{y} = \hat{P}\{Y = 1|S = 0\} = \#\{i|y_i = 1, s_i = 0\}/\#\{i|s_i = 0\}$;
  $k$:   • for all $i \notin \mathscr{M}^{(i_1,\dots,i_k)}$ let $D_i^{(k)} = D_i^{(k-1)}$;
         • for all $i \in \mathscr{M}^{(i_1,\dots,i_k)}$ let $D_i^{(k)} = y_i - \bar{y}_i$ where $\bar{y}_i = \hat{P}\{Y = 1|S = 0, x_{i_1}, \dots, x_{i_k}\} =$
           $\#\{i|y_i = 1, s_i = 0, i \in C_i^{(i_1,\dots,i_k)}\}/\#\{i|s_i = 0, i \in C_i^{(i_1,\dots,i_k)}\}$;
$K+1$: $D_i^{(i_1,\dots,i_k)} = D_i^{(K)}$.

Set the final discrimination scores as $D_i = \frac{1}{M} \sum_{i_1,\dots,i_K} D_i^{(i_1,\dots,i_k)}$.
```

Fig. 2 Pseudo-code for repeated sequential implementation of CEM.

variables in the dataset include age, years of education, working hours, type of occupation (sector, type of employer), family status, sex, capital gain and loss, native country (redefined as areas of the world). The *adult* dataset is comprised of 45222 observations, the s.a. is possessed by 6319 units (13.97%), the outcome is favorable in 13008 (28.76%), 31.38% among those not possessing the s.a., 12.68% among those possessing it. For the analysis, the dataset is split in a learning (30162 units) and test (15060 units) subsamples (same as in [6]).

To check the stability of the procedure, we performed it twice for 100 iterations and compared the results, which showed a very good agreement ($\rho > 0.99$).

In order to explore how well the proposed measures detect discrimination against the protected group (i.e. $D < 0$, $\bar{D} > 0$ are relevant) under different circumstances we considered three different scenarios: *(a)* presence of a variable which is related to S but unrelated to Y (Fig. 1*(e)*); *(b)* discrimination free data (simulated); *(c)* discriminating data (simulated).

It is expected that if a variable we condition on is related to the s.a., then the level of conditional discrimination be lower. We added to the data a variable having a mild correlation with S (independent of the outcome). In Fig. 3*(a)* we compare the discrimination scores D and $\bar{D}$ (first and second row respectively) estimated by the procedure with and without the correlated variable: as expected, the discrimination scores are higher if the variable is omitted ($D_i^{(-p)}$, y-axis).

Further, we modified the dataset by adding and removing discrimination against the protected class to assess the sensitivity of the discrimination scores. Starting from a classification tree estimate of the probability of a positive outcome based on all variables but the S, we built a discrimination free dataset by simulating the outcome according to the estimated probabilities and a strong discrimination dataset by changing the outcome of 2200 units for which the probability of positive outcome is between 0.3 and 0.7 (in particular, we changed the outcome to negative for 200 units having the S and a positive outcome, and we changed the outcome to positive

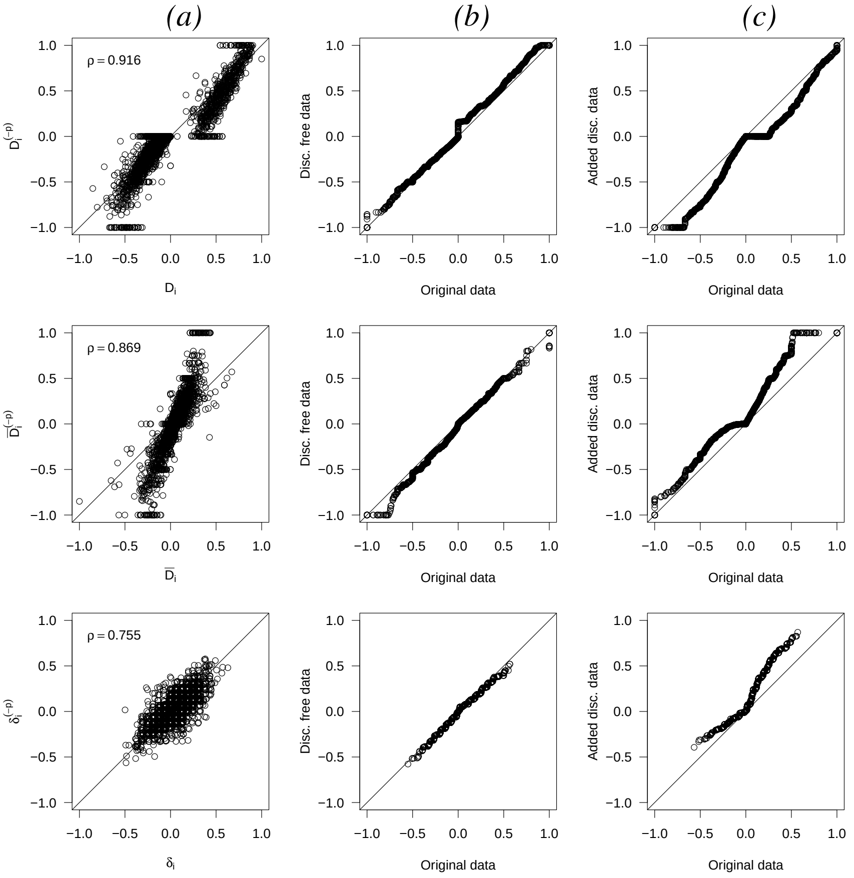


Fig. 3 Discrimination measures D_i, $\bar{D}_i$, δ_i (top, middle, bottom row respectively) performances: *(a)* scatter plots comparing discrimination measures obtained by including or omitting a conditioning variable simulated independently of the outcome to be correlated to S; *qq*-plots comparing the distributions of discrimination measures when discrimination is eliminated *(b)* and added *(c)*.

for 2000 units not having the S and a negative outcome). In Fig. 3*(b)* and *(c)* we compare the distributions of the discrimination measures using *qq*-plots.

Table 1 Classification accuracy after discrimination reductions.

% of S obs. changed	accuracy on testing data	accuracy on non S	accuracy on S
0	85.76	85.06	90.10
5	85.90	85.17	90.43
10	85.55	84.92	89.47
15	85.06	84.55	88.23
25	84.81	84.51	86.65

For comparison purposes we computed the measure δ of [6] (eq. (2) on the same datasets using 8, 16 and 32 as k values. Results are reported in Fig. 3 (bottom row) for $k = 16$, the other cases give qualitatively similar outputs. With respect to D, and to a minor extent to $\bar{D}$, the measure δ shows less variability since it is always based on groups of fixed size k, while very small (CEM) strata may contribute to the values of D ($\bar{D}$). Hence δ may be less sensitive to detect discrimination. This can be seen in scenario *(b)*, where only the D measure distribution differentiates between original and discrimination free data. Moreover, note that, contrary to D and $\bar{D}$, δ is weakly affected by the presence of a variable correlated to S.

The above discrimination measures can then be used to build a discrimination free dataset by changing the outcome of units with discrimination above a certain threshold. In Table 1 we report the performance of the classification (by a classification tree) after changing different numbers of units.

5 Conclusions

In order to detect inequality of treatment against protected classes in historical data it has been proposed to use the methods of causal inference to compare the treatment of statistical units belonging to the protected class to units not in the protected class which are similar with respect to the other observed characteristics. Similarity may be defined based on propensity scores or a distance metric. We argue that CEM stratification could be more apt to this task since it allows matching of units only if their characteristics are equal (possibly after coarsening of numerical variables), rather than relying on an overall distance. Preliminary results appear to confirm these expectations.

References

1. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268. ACM (2015)
2. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp. 3315–3323 (2016)
3. Iacus, S., King, G., Porro, G., et al.: CEM: software for coarsened exact matching. Journal of Statistical Software **30**(13), 1–27 (2009)
4. Iacus, S.M., King, G., Porro, G.: Multivariate matching methods that are monotonic imbalance bounding. Journal of the American Statistical Association **106**(493), 345–361 (2011)
5. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012)
6. Luong, B.T., Ruggieri, S., Turini, F.: k-nn as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 502–510. ACM (2011)
7. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review **29**(5), 582–638 (2014)

Statistical models for sports dat

The study of relationship between financial performance and points achieved by Italian football championship clubs *via* GEE and diagnostic measures

Lo studio della relazione tra risultati finanziari e punti realizzati delle squadre di calcio di serie A tramite GEE e misure di diagnostica

Anna Crisci, Sarnacchiaro Pasquale e Luigi D'Ambra

Abstract
Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. The main goal of any Football Championship club is to achieve sport results. The study of the relationship between sport and economic results attracts the interest of many scholars belonging to different disciplines. Very informative is considered the connection, over short or long periods of time, between the points in the championship and the resource allocation strategies. The aim of this paper is to give a interpretation of this last link using the Generalized Estimating Equation (GEE) for longitudinal data. Some diagnostic measures and graphical plots for checking the adequacy of GEE method will be presented and used.

Abstract
Il calcio in Italia è un fenomeno sociale che coinvolge intere comunità e continua ad aumentare il suo valore sociale ed economico. Lo studio della relazione tra i risultati sportivi ed economici riscuote l'interesse di tantissimi studiosi appartenenti a diverse discipline. Particolarmente stimolante è risultato il dibattito che lega, per ciascuna squadra di calcio, i punti in classifica alle capacità imprenditoriali del management sportivo in termini di allocazione delle risorse finanziarie e sportive. Obiettivo del presente lavoro è quello di dare un contributo in termini di interpretazione di quest'ultimo legame attraverso l'utilizzo delle Equazioni di Stima Generalizzate (GEE) per dati longitudinali. Alcune misure diagnostiche e metodi grafici per testare l'adeguatezza del metodo GEE saranno illustrati e utilizzati.

Key words: Italian Football championship clubs, sports and economic results, generalized estimating equations, Regression diagnostics

[1] Crisci Anna, Pegaso Telematic University, anna.crisci@unipegaso.it
Sarnacchiaro Pasquale, Univ. of Rome Unitelma Sapienza, pasquale.sarnacchiaro@unitelmasapienza.it
Luigi D'Ambra, University of Naples Federico II, dambra@unina.it

1. Introduction

Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. The main goal of any Football Championship club is to achieve sport results. Nevertheless, football has also become one of the most profitable industries, with a significant economic impact in infrastructure development, sponsorships, TV rights and transfers of players. Very informative is considered the connection between the points in the championship and the resource allocation strategies.

The Generalized Estimating Equation (GEE) methodology has been introduced to extend the application of generalized linear models to handle correlated data. For repeated measures, nowadays GEE represents a method based on a quasi-likelihood function and provides the population-averaged estimates of the parameters.

The aim of this paper is to give an interpretation of the link between the points in the championship and the resource allocation strategies using the GEE. In particular, we analyze the impact that some variables, including Income statement, Net equity and Team value, have on points made by football teams participating in the series A championship (2010-2015), by GEE for count data.

There are six sections in this study. A summary of GEE has been introduced in the second section. Section 3 deals with how to choose a working correlation structure and model selection in GEE. The Influential observation, Leverage and Outlier in GEE have been discussed in section 4. In section 5 the case study has been presented. Next, concluding remarks are presented in the final section of the paper.

2. Summary of the Generalized Estimating Equation method (GEE)

Let $\boldsymbol{y_i} = (y_{i1}, \dots, y_{it_i})'$ be a vector of responses value and let $\boldsymbol{X_i} = (\boldsymbol{X}'_1, \dots, \boldsymbol{X}'_n)'$ be a $t_i \times K$ matrix of covariates, with $\boldsymbol{x_{it}} = (x_{it1}, \dots, x_{itK})'$, i = 1,2, … ,n and t = 1,2, …,T. To simplify notation, let $t_i = t$ without loss of generality.

The expected value and variance of measurement y_{it} can be expressed using a generalized linear model:

$$E(y_{it}|\boldsymbol{x_{it}}) = \mu_{it}$$

Suppose that the regression model is $\eta_{it} = g(\mu_{it}) = \boldsymbol{x}_{it}^T\boldsymbol{\beta}$ where g is a link function and $\boldsymbol{\beta}$ is an unknown $K \times 1$ vector of regression coefficients with the true value as $\boldsymbol{\beta}_0$. The $Var(y_{it}|\boldsymbol{x_{it}}) = v(\mu_{it})\phi$, where v is a known variance function of μ_{it} and ϕ is a scale parameter which may need to be estimated. Mostly, v and ϕ depend on the distributions of outcomes. For instance, if y_{it} is continuous, $v(\mu_{it})$ is specified as 1, and ϕ represents the error variance; if y_{it} is count, $v(\mu_{it}) = \mu_{it}$ and ϕ is equal to 1.

Also, the variance-covariance matrix for $\boldsymbol{y_i}$ is noted by $\boldsymbol{V}_i = \phi \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R_i}(\boldsymbol{\alpha})$, $\boldsymbol{A}_i = diag\{v(\mu_{i1}), \dots, v(\mu_{iT})\}$ and the so-called "working" correlation structure $\boldsymbol{R_i}(\boldsymbol{\alpha})$ describes the pattern of measures within the subjects, which is of size $T \times T$ and depends on a vector of association parameters denoted by $\boldsymbol{\alpha}$. An iterative algorithm is applied for estimating α using the Pearson residuals $rp_{it} = \frac{y_{it}-\hat{\mu}_{it}}{\sqrt{v(\mu_{it})}}$ calculated from the current value of $\boldsymbol{\beta}$ (see section 4). Also, the scale parameter ϕ can be estimated by: $\hat{\phi} = \frac{1}{n-K}\sum_{i=1}^{n}\sum_{t=1}^{T} rp_{it}^{2}$. The parameters β are estimated by solving: $U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i'[V(\hat{\boldsymbol{\alpha}})]^{-1}\boldsymbol{s}_i = 0$ where $\boldsymbol{s}_i = (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$ with $\hat{\boldsymbol{\mu}}_i = (\mu_1, \dots \dots \mu_{iT})'$and $(\hat{\boldsymbol{\alpha}})$ is a consistent estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{D}_i' = \boldsymbol{X}_i'\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Lambda}_i = diag\ (\partial_{\mu_{i1}}/\partial_{\eta_{i1}} \dots \dots, \partial_{\mu_{it}}/\partial_{\eta_{it}})$. Under mildregularity conditions $\hat{\boldsymbol{\beta}}$ is asymptotically distributed with a mean $\boldsymbol{\beta}_0$ and covariance matrix estimated based on the sandwich estimator:

$$\hat{\boldsymbol{V}}_i^{\boldsymbol{R}} = (\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i)^{-1} \sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{s}_i\boldsymbol{s}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i)^{-1}\ (\mathbf{1})$$

In GEE models, if the mean is correctly specified, but the variance and correlation structure are incorrectly specified, then GEE models provide consistent estimates of the parameters and also the mean function, while consistent estimates of the standard errors can be obtained via a robust “sandwich” estimator. Similarly, if the mean and variance are correctly specified but the correlation structure is incorrectly specified, the parameters can be estimated consistently and the standard errors can be estimated consistently with the sandwich estimator.

3. Criteria for choosing a working correlation structure and model selection

Unlike the GLM method, which is based on the maximum likelihood theory for independent observations, the GEE method is based on the quasi-likelihood theory and no assumption is made about the distribution of response observations. Therefore, AIC (Akaike's Information Criterion), a widely used method for model selection in GLM, is not directly applicable to GEE.

The QIC (Quasilikelihood under the Independence model Criterion) statistic proposed by Pan [8], and further discussed by Hardin and Hilbe [7], is analogous to the familiar AIC statistic used for comparing models fit with likelihood-based methods:

$$QIC = -2Q(\hat{\boldsymbol{\mu}}; \boldsymbol{I}) + \boldsymbol{2trace}(\hat{\boldsymbol{\Omega}}_{\boldsymbol{I}}^{-1}\widehat{\boldsymbol{V}_i^{\boldsymbol{R}}})$$

where $\boldsymbol{I}$ represents the independent covariance structure used to calculate the quasi-likelihood, $\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{x}_{it}\hat{\boldsymbol{\beta}})$. The coefficient estimates $\hat{\boldsymbol{\beta}}$ and robust variance (estimator $\hat{\boldsymbol{V}}_i^{\boldsymbol{R}}$ are obtained from a general working covariance structure. Another variance estimator $\hat{\boldsymbol{\Omega}}_I$ is obtained under the assumption of an independence correlation structure.

QIC can be used to find an acceptable working correlation structure for a given model. When trace $\hat{\boldsymbol{\Omega}}_{\boldsymbol{I}}^{-1}\hat{\boldsymbol{V}}_i^{\boldsymbol{R}} \approx trace\ (\boldsymbol{I}) = K$, there is a simplified version of QIC, called QIC_u [8]: $QIC_u = -2Q(\hat{\boldsymbol{\mu}}; \boldsymbol{I}) + 2K$. QIC and the related QIC_u statistics can be used to compare GEE models and aid model selection. QIC_u approximates QIC when the

GEE model is correctly specified. When using QIC and related QIC_u to compare two models, the model with the smaller statistic is preferred.

4. Regression diagnostics: Residuals, Influential and leverage points

Model checking is an important aspect of regression analysis with independent observation [9]. Unusual data may substantially alter the fit of the regression model, and regression diagnostics identify subjects which might influence the regression relation substantially. Therefore, GEE approach also needs diagnostic procedures for checking the model's adequacy and for detecting outliers and influential observations. Graphical diagnostic plots can be useful for detecting and examining anomalous features in the fit of a model to data.

Regression diagnostic techniques that are used in the linear model [3] or in GLM [4] have been generalized to GEE. Venezuela *et al.* [10] described measures of local influence for generalized estimating equations. Here, we extend such diagnostic measures of the regression model in GEE approach (Table 1). The diagnostic measures are numerous and can be classified into five groups:

a) Measures based on the prediction matrix

In GEE the Hat matrix is $\boldsymbol{H} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{\frac{1}{2}}$ where $\boldsymbol{W} = diag(\boldsymbol{W}_1, \dots, \boldsymbol{W}_n)$ is a block diagonal weight matrix whose ith block corresponds to the ith subject. The leverage h_{it} as the ith diagonal element of the Hat matrix. Thus, h_{it} represent the high-leverage of ith observation y_i in determinig its own predicted value. It ignores the information contained in y. High-leverage can be rewritten by considering the Mahalonobis Distance (MD):

$h_{it} = MD_{it}^2 + \frac{1}{N}$ where $MD_{it} = \sqrt{(w_{it}^{1/2}\boldsymbol{x}_{it} - \bar{x})'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{x}_{it}'w_{it}^{1/2} - \bar{x})}$ and $\bar{x}$ is weighted mean. Cut off point: $2K/N$.

b) Measures based on residual

One method of detecting model failures is examining the residuals. There are many ways to compute residuals. The Pearson residual is a simple residual scaled by standard deviation of y_{it}. Pearson standardized residuals have been computed in order to have unit asymptotic variance. Anscombe residual have been introduced to make the distribution of the residuals as normal as possible [2].

c) Measures based on Influence Function

In this group we find CD [3, 4]; $\mathrm{SIC}_{\mathrm{it}}$ and $\mathrm{SC}_{\mathrm{it}}$.

d) Measures based on the volume of confidence ellipsoids

A measure of the influence of the ith observation on the estimated regression coefficients can be based on the change in volume of confidence ellipsoids with and without the ith observation. We consider:

a) The Andrews –Pregibon Statistic (AP) [1]: AP measures the volume of the confidence ellipsoid. It provides considerable information not only on outlying and influential observations but also on the remoteness of observations in the parameter space. We extend the AP in GEE approach. It is worth noting that this statistics does not assume the linear model and it therefore has a more general applicability. Indeed, it may be used not only as confirmative but also an exploratory tool and thus it may be extended to any data set independently of a model hypothesis.

b) The quasi-likelihood distance: Let $Q(\hat{\beta})$ is the quasi-likelihood estimate of the GEE parameters β using all response values and $Q(\hat{\beta}_{(it)})$ is the corrisponding estimate evaluate with the y_{it} observation deleted. A measure of the influence of the ith observation on $\widehat{\boldsymbol{\beta}}$ can be based on the distance between $Q(\widehat{\boldsymbol{\beta}})$ and $Q(\widehat{\boldsymbol{\beta}}_{(it)})$: $QD = 2[Q(\widehat{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}}_{(it)})]$.

e) Measures based on total influence.

The overall influence [6] is based on the simple fact that potentially influential observations are outliers as either X-outliers, y-outliers, or both (see Table 1). Hadi recommends using "mean(HD_{it}^2)+ c$\sqrt{var(HD_{it}^2)}$" as a cut-off point for Hadi's measure, where c is an appropriately chosen constant such as 2 or 3.

5. Case study

The data used for our case study was obtained from the financial statements filed by the Serie A football teams. The period of study concerned the championship from season 2010/2011 up to 2014/2015. The focus of the analysis is to verify the impact that the income statement, Net equity and Team value variables have on the points achieved by football teams. We have started by previous paper where we selected the best model through *Cp* Mallows [5]. The independent variables considered in the final model are: Depreciation Expense of multi-annual player contracts(DEM); Revenue net of player capital gain (RNC); Net Equity (NE). Later, for this model, considering the diagnostic measures presented in section 4, we can note that the teams, Roma, Udinese and Genoa, exceed the cut of value of some measures (see table 2). In particular, the Roma team to the championship 2013-14 exceeds the cut off values related to the measures in table 2. For this reason, we consider a new GEE model with exchangeable work correlation structure (α = 0.612), without the observation related to Roma team to the championship 2013-14. The results are described in table 3.

Table 2: Diagnostic Measures

Teams	***Pearson***	***Leverage*** ***(h_{it}> 0,2454)***	***Cook-Distance*** ***(CD_{it}> 0,07272)***	***Hadi***
Roma (2013-14)	2,3079	0,3426	0,3966	0,7505 (HD^2>0,64)
Udinese (2012-13)		0,3083		0,4757 (HD^2>0,45)
Genoa (2010-11)		0,2558		

Table 3: The Poisson GEE Population-Averaged Model with Exchangeable Structure

Points	***Coef***	***St.err.***	***Z***	***P >| z|***
DEM	-0,0720	0,039	-1,84	0,066
RNC	0,3589	0,059	6,08	0,000**
NE	0,0539	0,017	3,17	0,002**
Cons	-2,2083	0,665	-3,32	0,009
Wald Stat.	93,56	Prob>chi^2 0,0000		

** significant at 5%.

Finally, we can note that the minimal working residual, computed by using correlation matrix, is obtained when we delete Roma Team. Concluding, we have discussed and reviewed the various measures which have been presented for studying outliers, high leverage points, and influential observations in the context of GEE approach. As an illustration, a data set about impact that some budget variables have on points achieved by football teams in the Serie A championship [5], has been presented, applying the methods developed in Section 4.

References

1. Andrews, D.F. and Pregibon, D. (1978). Finding the ouliers that matter. *J. Roy. Statist. Soc. Ser. B* **40** 85-93.
2. Anscombe, F. J. (1953). Contribution to the discussion of H. Hotelling's paper. *J. Roy. Statist. Soc. Ser. B* **15** 229-230.
3. Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**,15–18.
4. Cook, R. D and Thomas, W. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika*, 76, 741–750.
5. Crisci, A., D'Ambra, L. and Esposito, V. (2018). A Generalized Estimating Equation in Longitudinal Data to Determine an Efficiency Indicator for Football Teams. Social Indicators Research. https://doi.org/10.1007/s11205-018-1891-6
6. Hadi, A.S (1992). A new measure of overall potential influence in linear regression. Computational Statistics and Data Analysis **14**, 1–27.
7. Hardin, J., & Hilbe, J. (2003). *Generalized estimating equations*. London: Chapman and Hall.
8. Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics,* **57**(1), 120–125.
9. Thomas, L.R., 1997. Modern Econometric: An Introduction. Addison-Wesley.
10. Venezuela M.K., Botter D.A., Sandoval M.C. (2007) Diagnostic techniques in generalized estimating equations, Journal of Statistical Computation and Simulation, 77:10, 879-888, DOI: 10.1080/10629360600780488

Table 1: Regression Diagnostics

Measures and Formula	***Interpretation and Cut off point***
a) Hat matrix h_{it} is the *i*-th diagonal element of the Hat matrix $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$	It allows to identify high leverage point $h_{it} \geq 2K/N$
$h_{it} = MD_{it}^2 + \frac{1}{N}$ where $\mathrm{MD_{it}} = \sqrt{(\mathrm{w}_{\mathrm{it}}^{\frac{1}{2}}\mathbf{x}_{\mathrm{it}} - \bar{\mathrm{x}})'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{x}_{\mathrm{it}}'\mathrm{w}_{\mathrm{it}}^{\frac{1}{2}} - \bar{\mathrm{x}})}$	High-leverage point can be computed by using the Mahalonobis Distance (MD) $h_{it} \geq 2K/N$
b) Residuals	
Square Pearson residuals $rp_{it}{}^2 = \sum_{i=1}^{n}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)'\boldsymbol{\Lambda}_i{}^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)$	rp_{it} is a simple residual scaled by standard deviation of y_{it} . Residuals are evalueted at the current value of $\boldsymbol{\beta}$. The matrix $\boldsymbol{\Lambda}_i{}^{-1}$ can be replaced by $\boldsymbol{V}_i$ and $\boldsymbol{R}_i$ in order to consider the correlation within subjects (working residual)
Pearson standardized residuals $(rpsd)_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{v(\mu_{it})(1 - h_{it})}}$	$(rpsd)_{it}$ is standardized in order to have unit asymptotic variance
Anscombe residuals for poisson distribution $r_{it}^A = \frac{\frac{3}{2}(y_{it}^{2/3} - \hat{\mu}_{it}^{2/3})}{\hat{\mu}_{it}^{1/6}}$	r_{it}^A: Anscombe (1953), proposed a residual using the function $G(y)$ in place of y where $G(\cdot)$ is chosen to make the distribution of as normal as possible. For univariate generalized linear models $G(\cdot)$ is given by: $G(\cdot) = \int \frac{1}{V^{\frac{1}{3}}(\mu)}\partial\mu$
c) Influence function	
Cook distance (CD) $(CD)_{it} = rpsd_{it}^2 \frac{h_{it}}{K(1 - h_{it})}$	$(CD)_{it}$ is a measure to detect clusters with a strong influence on parameter estimates $(CD)_{it} > 4/N$
$SIC_{it} = (\boldsymbol{N} - \boldsymbol{1})(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{it}}'(\boldsymbol{y}_{\boldsymbol{it}} - \boldsymbol{x}_{\boldsymbol{it}}\widehat{\boldsymbol{\beta}}_{(\boldsymbol{it})})$ $SC_{it} = N(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{x}_{it}'\frac{r_{it}}{1 - h_{it}}$ where $r_{it} = (\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)$	SIC_{it} and $SC_{it}$are easier to interpret; they are proportianal to the distance between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(\mathrm{it})}$
d) Volume of confidence ellipsoid	
Andrews/Pregibon Statistic (AP) $AP_{it} = \frac{\lvert \boldsymbol{X}_{(it)}^{*\prime}\boldsymbol{X}_{(it)}^{*}\rvert}{\lvert \boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}\rvert}$	It provides considerable information not only on outlying and influential observations but also on the remoteness of observations in the parameter space. Small value of AP_{it} calls for special attention

The Quasi likelihood distance (QD): $QD = 2\left[Q\left(\widehat{\boldsymbol{\beta}}\right) - Q(\widehat{\boldsymbol{\beta}}_{(\boldsymbol{it})})\right]$	
e) Overall influence; $HD^2 = \frac{K}{(1-h_{it})}\frac{d_{it}^2}{(1-d_{it}^2)} + \frac{h_{it}}{(1-h_{it})}$ where $d_{it}^2 = \frac{\boldsymbol{r_{it}^2}}{\boldsymbol{r'r}}$	HADI is based on the simple fact that potentially influential observations are outliers as either X-outliers, y-outliers, or both. $HD^2 >$ mean(HD_{it}^2)+ c$\sqrt{var(HD_{it}^2)}$

Exploring the Kaggle European Soccer database with Bayesian Networks: the case of the Italian League Serie A

Maurizio Carpita and Silvia Golia

Abstract In the last decade, the application of statistical techniques to the sport field significantly increased. One of the most famous sports in the world is the soccer (football), and the present work deals with it, using data referred to the seasons from 2009/2010 to 2015/2016 of the Italian League Serie A extracted from the Kaggle European Soccer database. The players overall performance indicators, obtained on the basis of the players position or role (forward, midfielder, defender and goalkeeper), are used to predict the result of the matches by applying the Bayesian Networks as well as the Naive Bayes and the Binomial Logistic Regression models, considered as their competitors.

Abstract *Nell'ultima decade, l'applicazione di tecniche statistiche in ambito sportivo ha avuto un incremento significativo. In questo lavoro che riguarda il calcio, uno degli sport più famosi al mondo, si usano i dati relativi alle stagioni 2009/2010-2015/2016 del Campionato Italiano di Serie A estratti dal database Kaggle European Soccer. Gli indicatori di performance complessiva dei giocatori, ottenuti sulla base della loro posizione ovvero del loro ruolo (attaccante, centrocampista, difensore e portiere), sono utilizzati per prevedere il risultato delle partite utilizzando le Reti Bayesiane così come i modelli Naive Bayes e Logistico, considerati loro competitori.*

Key words: Kaggle European Soccer database, Bayesian Networks, Naive Bayes, Binomial Logistic Regression Model, Italian League Serie A

Maurizio Carpita
University of Brescia, Department of Economics and Management, C.da S.Chiara, 50 - 25122 Brescia, Italy, e-mail: maurizio.carpita@unibs.it

Silvia Golia
University of Brescia, Department of Economics and Management, C.da S.Chiara, 50 - 25122 Brescia, Italy, e-mail: silvia.golia@unibs.it

1 The Kaggle European Soccer database

In the last decade, the application of statistical techniques to the sport field and in particular to the European soccer (football) significantly increased, especially to predict matches' results [1, 15] using for example players performance statistics [10, 13] also for the Italian League [3, 4]. In the big data era, many databases are constructed and used to develop predictive models. On-line platforms for predictive modeling and analytics competitions, as Kaggle (www.kaggle.com), emerged and developed a meeting point for data scientists. This study uses data from the Kaggle European Soccer (KES) database [12], that contains data about 28,000 players and about 21,000 matches of the championship leagues of 10 countries and 7 seasons from 2009/2010 to 2015/2016, resulting in one of the biggest open database devoted to the soccer leagues of European countries, with data about players, teams and matches for several seasons [5].

The *Player Attributes table* of the KES database contains 33 variables, that represent player's performance indicators on the 0-100 scale with respect to overall and different abilities of the soccer play (power, mentality, skill, movement, attacking, goalkeeping), built with the experts classification of the EA Sports FIFA videogame. To develop the models of this study, only the players overall performance indicators and match result (in terms of goals scored by home and away teams) for the seasons from 2009/2010 to 2015/2016 of the Italian League Serie A are used. Given the presence of some missing values, the number of available matches is 2,587 instead of 2,660. The players overall performance indicator has been averaged based on the players' position or role, that is forward (FOR), midfielder (MID), defender (DEF) and goalkeeper (GOK), in each match according to the coach decisions before the match takes place. As explained in [5], the players' role is obtained from the *Match table* of the KES database, which contains X and Y coordinates representing the positions of the 22 players on the soccer pitch.

The aim of the study is twofold. First, one wants to explore these data using the Bayesian Networks (BN) as main model, to be compared to Naive Bayes (NB) and Binomial Logistic Regression (BLR) models, second, one wants to evaluate the power of overall performance indicators of the four roles in a soccer team (goalkeeper, defender, midfielder and forward role) in predicting the matches' results.

The paper is organized as follows. Sect. 2 contains a brief description of the theory underlying BN, NB and BLR models, whereas Sect. 3 reports the main results of the application of the three models to the data under study. Conclusions and ideas for future research follow in Sect. 4.

2 The three statistical models

Probabilistic networks are graphical models that explicit through a graph, the interactions among a set of variables represented as vertices or nodes of the graph. Any pair of unconnected nodes of the graph indicates (conditional) independence between the variables represented by these nodes under certain conditions that can be read from the graph itself. Hence, a probabilistic network captures a set of (conditional) dependence and independence properties associated with the variables represented in the network [8]. BN belong to the class of probabilistic networks. The underlined graph is called Directed Acyclic Graphs (DAG). A DAG $\mathscr{G}$ is a pair $\mathscr{G} = (\mathbf{V}, E)$, where $\mathbf{V}$ is a finite set of distinct vertices, $\mathbf{V}= \{V_i\}_{i=1}^k$, which correspond to a set of random variables $\mathscr{X} = \{X_{V_i}\}_{i=1}^k$ indexed by $\mathbf{V}$, $E \subseteq \mathbf{V} \times \mathbf{V}$ is the set of directed links (or edges) between pairs of nodes in $\mathbf{V}$. An ordered pair $(V_i, V_j) \in E$ denotes a directed edge ($\rightarrow$) from node V_i to node V_j; V_i is said to be a parent of V_j whereas V_j a child of V_i. The set of parents of a node V shall be denoted by $pa(V)$. A Bayesian Network (BN) over the variables $\mathscr{X}$ is defined as the triplet $\mathscr{N} = (\mathscr{X}, \mathscr{G}, \mathscr{P})$, where $\mathscr{G}$ is a DAG and $\mathscr{P}$ is a set of conditional probability distributions containing one distribution $P(X_v|X_{pa(v)})$ for each random variable $X_v \in \mathscr{X}$, where $X_{pa(v)}$ denotes the set of parent variables of variable X_v. The joint probability distribution $P(\mathscr{X})$ over the set of variables $\mathscr{X}$ is factorized as

$$P(\mathscr{X}) = \prod_{v \in \mathbf{V}} P(X_v|X_{pa(v)}). \tag{1}$$

So, a BN can be described in terms of a qualitative component, that is the DAG, and a quantitative component, consisting of the joint probability distribution (1).

The construction of a BN runs in two steps. First, one identifies the interactions among the variables generating a DAG, then the joint probability distribution has to be specified in terms of the set of conditional probability distributions $P(X_v|X_{pa(v)})$.

The DAG can be derived either manually or automatically from data. In order to automatically find the BN structure, several algorithms have been proposed in the literature, falling under three broad categories: constraint-based, score-based, and hybrid algorithms. Constraint-based algorithms make use of conditional independence tests focusing on the presence of individual arcs, score-based algorithms assign scores to evaluate DAGs as a whole, whereas hybrid algorithms combine constraint-based and score-based algorithms. The method used in this paper to derive the DAG is the Hill Climbing (HC), implemented in the R package `bnlearn` provided by Scutari [16]. It belongs to the class of the score-based algorithms and it consists in exploring the search space starting from an empty DAG and adding, deleting or reversing one arc at a time until the score considered can no longer be improved [17].

Given a BN, it can be used to answer questions (queries) related to the domain of the data that goes beyond the description of its behavior; for BN the process of answering questions is also known as probabilistic reasoning or belief updating. Basically, it focuses on the calculus of the posterior probabilities given a new piece of information called evidence. In fact, suppose to have learned a BN, it is used to compute the effect of new evidence Ev on one or more target variables X' using the knowledge encoded in the BN, that is to compute posterior distribution $P(X'|Ev)$. The procedure used in the paper makes use of the junction tree representation of a BN which is composed by cliques (a clique is a maximal complete subgraph of a moralization of the DAG $\mathscr{G}$). Belief updates can be performed efficiently using Kim and Pearls Message Passing algorithm [9].

Even if all the variables play the same role in the construction and usage of the BN, in the present paper one of the variables will be considered as target variable, similar to the variable Y used in the definition of the two competing models that follows.

In order to compare the performance of BN with the performances of competing models, the NB and BLR models are taken into account.

The Naive Bayes (NB) is one of the simplest restricted probabilistic graphical models [6]. It is characterized by a structure where one single class variable is parent of the remaining attribute variables, which are conditionally independent given the class variable. So, let Y be the class variable and let $\{X_i\}_{i=1}^{p}$ be the attribute variables, their joint probability is factorized as $P(Y,X_1,\cdots,X_p) = P(Y)\cdot\prod_{i=1}^{p} P(X_i|Y)$. From the definition of conditional probability, $P(Y|X_1,\cdots,X_p)$ is given by the following expression:

$$P(Y|X_1,\cdots,X_p) = \alpha\cdot P(Y)\cdot\prod_{i=1}^{p} P(X_i|Y), \tag{2}$$

where α is a normalization constant. Equation (2) is the common definition of a NB.

The Binomial Logistic Regression (BLR) model belongs to the family of the generalized linear models and it is used to estimate the probabilities of the two categories of a binary dependent variable Y using a set of covariates or predictors $\mathbf{X} = \{X_i\}_{i=1}^{p}$ [7]. The model assumes that the binary response variable is distributed as a Bernoulli with probability π and a logit link function. It can be expressed through the *logit transformation* as:

$$logit[\pi(\mathbf{X})] = log\left[\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}\right] = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

where $\pi(\mathbf{X}) = P(Y = 1|\mathbf{X})$ and $\boldsymbol{\beta}$ is the vector of coefficients involved in the linear predictor $\mathbf{X}\boldsymbol{\beta}$. The BLR model was already applied to predict the result of soccer matches in other studies [2, 5, 11].

3 Statistical evidences for the Italian League Serie A

As explained in Sect. 1, the dataset used in this paper was obtained from the KES database and contains the overall performance indicators for the four roles in a soccer team (goalkeeper, defender, midfielder and forward role), used as predictors, and the matches' results reported in terms of goals scored by the home and away teams. The number of overall performance indicators is eight and corresponds to p for the NB and BLR models. From the goals scored by the two teams of the match, it is possible to determine the outcome of the match from the home team point of view, *result*, classified as win, draw and loss. Preliminary analysis, using this classification for the *result* variable, has shown a high difficulty of the BN in predicting the draw with respect to win and loss; similar performances were observed in previous studies [3, 4]. This finding could be due to the fact that a draw is an outcome characterized by a higher degree of uncertainty, as well as to class imbalance [14] as in this case, where percentage of wins is 46.1%, whereas percentages of loss and draw are 26.4% and 27.5% respectively. In order to partially overcome this problem, the *result* variable was dichotomized into two categories: WIN and NOWIN (that is loss or draw) of the home team. Clearly, a model that takes into account three categories for the outcome with satisfactory prediction accuracy should be preferred to a model that involves less categories. Nevertheless, a model based on a binary variable, as the dichotomized *result*, can be of interest for the home team (obviously interested to the WIN probability) as well as the away team (interested to the NOWIN probability of the home team, that is the NOLOSS probability of the away team). The *result* variable represent the target variable denoted by Y in the NB and BLR models. So, the variables used in all the analyses are nine and this number corresponds to k for BN; the link between p and k is the following: $k = p + 1$.

Given that the dataset under study contains continuous variables (averages by players role of the overall performance indicator on 0-100 scale) and a discrete variable (outcome of the match), it is necessary to discretize the continuous variables. In applying BN it is possible to manage hybrid database like the one under study, but it is necessary to constraint the relation parent-child, imposing that a discrete variable may only have discrete parents [8]. In the context of this paper, this kind of constraint implies that the overall performance indicators can not be parents of the outcome of the match and this constraint appears unrealistic.

The method used to discretize the players' performance indicators, is the Equal Frequency Discretizer that involves the quantiles of the variable's distribution. The number of categories was fixed to four.

In order to identify the best combination of score-based algorithm and score that gives a DAG with highest score, a cross-validation has been performed. The scores under evaluation were the Bayesian Information criterion (BIC) and the Bayesian Dirichlet Equivalent (BDE) uniform posterior probability of the DAG associated with a uniform prior over both the space of the DAGs and of the parameters [17]. The selected score was the BDE with imaginary sample size (iss) equal to 100; iss determines how much weight is assigned to the prior distribution compared to the data when computing the posterior. Fig. 1 shows the obtained DAG. It can be seen that the home and away teams are well separated and with a coherent structure of links. In fact, the goalkeeper role is connected with the defender and midfielder roles, the defender role is linked to the midfielder role and the forward role is related to the midfielder and defender roles. Moreover, the structure of relations between the performance indicators of both the home and away teams is the same, whereas the variables with a direct link with the variable *result* are different. For the home team the role directed linked to the variable *result* is the midfielder role whereas the one of the away team is the defender role. This finding is coherent with the match strategies chosen by the two coaches: generally, the home team goal is to win the match, so the midfielder role is important, whereas the away team goal is not to lose the match and in this case, the defender role is the strategic one.

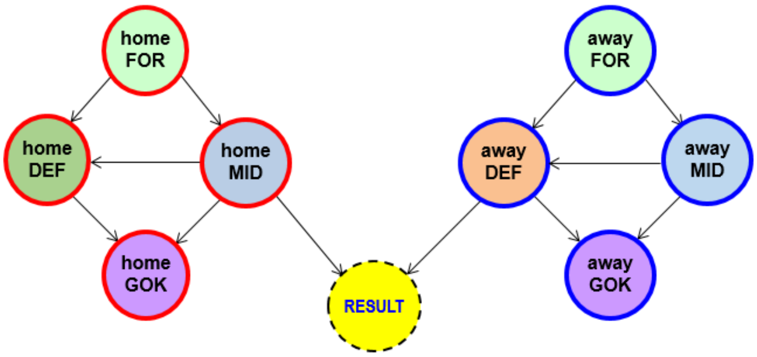


Fig. 1 DAG for the Italian League Serie A, Seasons 2009/2010-2015/2016

In order to compare the performance of the BN applied to the data, NB and BLR models were estimated. Table 1 reports the estimates of the parameters involved in the BLR model with the corresponding z-statistic. The performance indicators are maintained categorical in order to use the same data as the BN; same comments arise from the analysis of the results of the BLR model with continuous performance indicators. The most significant variables for the BLR model are *home_MID* and *away_DEF*, which are the ones with directed links to *result* in BN.

Table 1 Parameter estimates and z-statistics for the BLR Model

Perf. Indicator	$\hat{\beta}$	z-stat	Perf. Indicator	$\hat{\beta}$	z-stat
home_FOR2	-0.023	-0.181	away_FOR2	-0.094	-0.793
home_FOR3	0.029	0.206	away_FOR3	-0.021	-1.157
home_FOR4	0.319	1.959	away_FOR4	-0.289	-1.758
home_MID2	0.155	1.231	away_MID2	0.057	0.463
home_MID3	0.555	3.817	away_MID3	0.077	0.527
home_MID4	0.776	4.173	away_MID4	-0.466	-2.430
home_DEF2	0.086	0.693	away_DEF2	-0.260	-2.165
home_DEF3	0.336	2.286	away_DEF3	-0.420	-2.788
home_DEF4	0.478	2.689	away_DEF4	-0.672	-3.692
home_GOK2	-0.035	-0.288	away_GOK2	-0.078	-0.649
home_GOK3	-0.007	-0.060	away_GOK3	-0.153	-1.309
home_GOK4	-0.013	-0.087	away_GOK4	0.010	0.072

All the methods considered calculate the probability of win given new information. In order to predict the result of the match in terms of WIN-NOWIN of the home team, the simple majority rule is the most popular way to convert a probability into a predicted result; this is the rule used in the paper. Precision of model predictions can be evaluated by computing some indexes such as the *Accuracy*, which expresses how effectively the model predicts matches' results, the *Sensitivity*, which expresses how effectively the model predicts WIN matches and the *Specificity*, which expresses how effectively the model predicts NOWIN matches. Moreover, the accuracy of a model can be compared with the so called *null accuracy*, which measures the accuracy obtained without models and corresponds to the highest observed frequency of the two possible results of the match. Table 2 reports the prediction performances of BN, NB and BLR considering 500 random samples of 500 matches form the 2,587 available. For each sample of 500 matches, the DAG is the one reported in Fig. 1, whereas in applying NB and BLR, all the overall performance indicators are taken into account as predicting variables of *result*. All the conditional probabilities involved in BN and NB and the coefficients of BLR are computed making use of the remaining 2,087 matches.

Table 2 Mean predictive capability of the models based on the prediction of 500 matches' results randomly sampled 500 times (standard errors are in parenthesis)

Model	Accuracy	Sensitivity	Specificity
BN	0.623 (0.020)	0.550 (0.055)	0.688 (0.052)
NB	0.632 (0.021)	0.577 (0.032)	0.679 (0.030)
BLR	0.630 (0.021)	0.525 (0.031)	0.721 (0.031)

All methods have similar accuracy (about 63%), which is about ten percentage points greater than the null accuracy (53.9%). Moreover, BN and NB are more accurate in the WIN prediction (sensitivity) than BLR, which is more accurate in predicting NOWIN (specificity). Finally, standard errors for sensitivity and specificity of BN are greater than those of NB and BLR of about two percentage points.

4 Conclusions and future research

The paper shows some results obtained analyzing data regarding the Italian League Serie A extracted from the KES database; the aim was to explore these data using the BN and evaluate the power of overall performance indicators of the four roles in a soccer team (goalkeeper, defender, midfielder and forward role) in predicting the matches' results. A preliminary analysis suggested the necessity to collapse the match's results loss and draw in a unique category. Moreover, due to the fact that BN works with discrete variables, the overall performance indicators were discretized. In addition to BN, NB and BLR models were considered as competitors. All the three models have similar accuracy (around 0.63) significantly greater than null accuracy (0.54), so it is possible to conclude that the variables considered have some predictive power. Regarding the application of BN, the DAG expressed by the data shows coherent structure of links between the roles and maintains the home and away teams indicators well separated. Moreover the variables with a direct link with the variable *result* were the midfielder role for the home team and the defense role for the away team. This finding is coherent with the match strategies chosen by the two coaches: generally, the home team goal is to win the match, so the midfielder role is important, whereas the away team goal is not to lose the match and in this case, the defender role is the strategic one.

This study can be extended in various directions. For example, an advantage in using BN with respect of other models is the possibility to use partial information to predict a match's result, and this will be argument of future research. Moreover, another interesting development of this paper will be to extend the analysis to leagues of other countries in Europe in order to verify similarities in the results.

References

1. Albert, J., Glickman, M.E., Swartz, T.B., and Koning, R.H.: Handbook of Statistical Methods and Analyses in Sports. CRC Press (2017)
2. Alves, A.M., Mello, J.C.C.B.S., Ramos, T.G., Sant'Anna, A.P., et al.: Logit models for the probability of winning football games. Pesquisa Operacional **31(3)**, 459–465 (2011) doi: 10.1590/S0101-74382011000300003
3. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Football mining with R. In: Yanchang, Z., Yonghua, C. (eds.) Data Mining Applications with R, pp. 397-433. Springer (2014)
4. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Discovering the drivers of football match outcomes with data mining. Quality Technology & Quantitative Management **12(4)**, 561–577 (2015) doi: 10.1080/16843703.2015.11673436
5. Carpita M., Ciavolino E., Pasca P.: Exploring and modelling team performances of the Kaggle European Soccer Database. Submitted (2018)
6. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning **29**, 131–163 (1997) doi: 10.1023/A:1007465528199
7. Hosmer, D.W.Jr, Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, 3rd ed. John Wiley & Sons (2013)
8. Kjærulff, U.B., Madsen, A.L.: Bayesian networks and influence diagrams: a guide to construction and analysis. Springer, New York (2013)
9. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT Press (2009)
10. Leung, C.K., Joseph, K.W.: Sports data mining: predicting results for the college football games. Procedia Computer Science **35**, 710–719 (2014) doi: 10.1016/j.procs.2014.08.153
11. Magel, R., Melnykov, Y.: Examining influential factors and predicting outcomes in European soccer games. International Journal of Sports Science **4(3)**, 91–96 (2014) doi: 10.5923/j.sports.20140403.03
12. Mathien, H.: European Soccer Database. (2016) Data retrieved from
http://www.kaggle.com/hugomathien/soccer
13. McHale, I.G., Szczepański, L.: A mixed effects model for identifying goal scoring ability of footballers. Journal of the Royal Statistical Society: Series A **177(2)**, 397–417 (2014) doi: 10.1111/rssa.12015
14. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery **28(1)**, 92–122 (2014) doi: 10.1007/s10618-012-0295-5
15. Odachowski, K., Grekow, J.: Using bookmaker odds to predict the final result of football matches. In Graña, M., Toro, C., Howlett, R.J., Jain, L.C., (eds.) Knowledge Engineering, Machine Learning and Lattice Computing with Applications, pp. 196-205. Springer, Berlin Heidelberg (2013)
16. Scutari, M.: Learning bayesian networks with the bnlearn R package. Journal of Statistical Software **35(3)**, 1–22 (2010) doi: 10.18637/jss.v035.i03
17. Scutari, M., Denis J-B.: Bayesian Networks With Examples in R. CRC Press, Taylor & Francis Group (2015)

A data-mining approach to the Parkour discipline

Un approccio data-mining alla disciplina del Parkour

Paola Pasca, Enrico Ciavolino and Ryan L. Boyd

Abstract Parkour is a relatively new discipline. As an uncommon nexus between risk, resistance, and the philosophy of overcoming obstacles, it is continuously gaining interest and popularity. However, the “voice” of parkour practitioners has been explored only qualitatively, on small samples and from a phenomenological point of view. In this work, raw data from the official American Parkour forum (from 2005 to 2013) have been web-scraped and treated with the Meaning Extraction Method (MEM), a simple and flexible technique providing optimal dimension reduction and the identification of broader themes related to the parkour discipline, thus providing a broader vision of a phenomena which is configuring itself as a true life-style.

Abstract *Il parkour è uno sport piuttosto recente: come nesso tra rischio, resistenza e filosofia degli “ostacoli da superare”, è diventata una disciplina sempre più interessante e popolare. Tuttavia, la voce dei praticanti è stata esplorata solo in una ricerca qualitativa, effettuata su un piccolo campione, e solo da un punto di vista fenomenologico. In questo lavoro, per la prima volta, sono stati estratti dati provenienti dall'intero forum American Parkour (dal 2005 al 2013) e sono stati trattati con il Meaning Extraction Method (MEM), una tecnica semplice e flessibile che permette di ridurre la dimensionalità dei dati e di identificare i temi più ampi e caratteristici del parkour, una disciplina che si configura sempre più come uno stile di vita.*

Key words: parkour, online language use, Meaning Extraction Helper

Pasca Paola, e-mail: paola.pasca@unisalento.it
Enrico Ciavolino, e-mail: enrico.ciavolino@unisalento.it
Department of History, Society and Human Studies. University of Salento, via di Valesio, 73100, Lecce (Italy)
Ryan L. Boyd, e-mail: ryanboyd@utexas.edu
Department of Psychology, University of Texas at Austin, 108 E Dean Keeton St, Austin, TX 78712, Stati Uniti

1 An introduction to the Parkour discipline

Parkour is a relatively recent practice which is continuouly gaining interest and popularity, due to a significant presence on both tv and the *web*. Its first manifestations started as a game for its founders David Belle and Sébastian Foucault, in the context of the deprived parisian suburbs [3].

However, the roots of parkour can be dated back prior to the First World War: Georges Hébert, a military trainer of the French Navy, believing in intense physical training as a means for self-development, created a training method that combined physical obstacles with psycho-emotional barriers. The so-called *Natural Method* encouraged training in unconstrained, outdoor settings, where different terrain or obstacles could be faced and overcome. Grounded on ten main groups of exercises (walking, running, jumping, quadrupedal movement, climbing, balancing, throwing, lifting, defending and swimming), the *Natural Method* aims at making mind and body agile and adaptive in any situation: its fast, fluid and forward movements transposed into the urban environment are the essence of parkour [3, 17].

Despite its popularity, parkour is still permeated by prejudice. On a large scale, thoughtlessness, reckless behaviour and even suicide is sometimes attributed to it, even if there are no formal statistics related to deaths caused by parkour; moreover, training at one's own physical level is at the heart of the practice and practice itself starts from a ground level where falls are not life threatening. On a smaller scale, parkour practitioners, better known as *traceurs*, are often discouraged from practicing or banished by authorities, even if they almost always take care of their training spots by keeping them safe, clean and populated [17].

These controversial aspects make parkour an interesting topic for research and indepth exploration. Previous literature shed light on various aspects of the discipline [5, 19, 3, 13]. However, it is mainly focused on a small-group scale, based on qualitative data (interviews, narratives) on which no statistical processing has ever been performed.

This paper approaches the parkour discipline through simple data-mining techniques. It focuses on the entire American Parkour forum and lets the voice of *traceurs* and interested people emerge with the help of recent software developments for the treatment, processing and meaning extraction from linguistic data [10, 7].

A brief description of dataset, procedures and methods is reported in section 2. Some of the results of the natural language processing are reported in section 3.

2 Data and Methods

When it comes to the collection of unstructured data from the *web*, *web scraping* may help [16]. In the field of computer science, the term refers to several methods for data extraction and for their conversion into metadata which can further be processed and analyzed. In this case, one of the simplest, most intuitive software for efficient collection of *web* data has been used: *Helium Scraper* [2].

Through this software, linguistic data were exported from one of the biggest and most active online communities of parkour/freerunning practitioners: the American Parkour Forum [1]. There, *traceurs* and *freerunners* can share experiences, advices and ideas related to the discipline.

2.1 Dataset

The raw dataset covers a period of time ranging from 2005 to 2013. It consists of a .csv file containing *unique Id* of the post, *text in the post, number of responses, dates* and *times, author* of the thread, thread *title*.

On the raw dataset, a consistent amount of threads containing few hundreds of words could be noticed (precisely, 2692 threads contained an amount of words equals or inferior to 640 words).

In order to facilitate the information extraction in the text processing phase, authors decided to exclude from the analysis the threads with a word count lower than the minimum determined for segmentation, that is, 100 words. With respect to the whole corpus of data, this choice prevents from including threads likely to contain repetitions and comment citations, that is, less informative data. Building on these assumptions, we retained threads whose overall word count was above or equal to 5000 words.

Moreover, we excluded one thread which was in arabic language, and removed *links* and *urls* mentioned within the texts. Grouped by thread, the final dataset ($>$5000) included a total of 249 threads and 17275 comments.

2.2 Meaning Extraction Method

The Meaning Extraction Method (MEM), developed by Chung and Pennebaker [11] is based on the assumption that words related to a particular topic will tend to be used together. The mere observation of associated words allows researchers to draw inferences on how much people are talking, about what topics and in what way. This simple assumption makes application and interpretation procedures easier and more efficient.

MEM procedures start with a binary dataset of used-vs-non used words. This first step requires cleaning data from stop words (words that carry little to no meaning, such as ”the”, ”you”, ”did”), uncommon and non-informative words, in order to be sure that the most meaningful themes will be extracted. Then, a common factor analytic approach such as Principal Component Analysis can be applied to the word-by-observation table. Applied to natural language, results will provide word clusters reflecting broader themes emerged from the sample of text.

In order to speed up the front-end procedures used in this (and other) topic modeling approaches, Boyd released the Meaning Extraction Helper (MEH) [8], a free-ware that automates many of the steps described above. Its simplest application requires the user to make few basic selections, to point the software to the location of the .txt files and run the analysis. MEH efficiently converts unstructured linguistic data into structured matrices ready to be analyzed, leaving researchers to apply statistical techniques for meaning extraction [7].

In literature, several applications confirmed MEM as a promising method for uncovering information regarding psychological dimensions [11], personal values [9] and cultural self-schemas [18]. In particular, a recent application [6] proved MEM efficacy in catching how people think and talk online.

Building on previous literature, the American Parkour dataset has been processed with MEH through the following settings:

- upload a list of stop words (both the default one and another stop word list determined by the first author who, as a parkour practitioner, is familiar to parkour language);
- upload conversions (e.g. words such as *compete, competitive, competing* to be coded as *competition*);
- split files into equally sized segments (38 per thread). The number of segments is the ratio between the *mean by thread* value and a pre-determined words-per-segment value, that is 250. As said above, the minimum size (in terms of word count) admitted for a segment to be included in the output is set to 100 words.
- setting the minimum word percentage of word appearance in order to be included in the outputs to 3%.

The rationale for text processing choices is to make the most out of the combination of data descriptive statistics and MEH characteristics. The average number of segments (38) will split small threads into very small segments (in terms of word count) and bigger threads into big ones. However, the 3% required to a word to be included for analysis ensures to take the very essential from the smallest threads and more contents from bigger threads.

MEH uses a dictionary to identify common content words in each segment and systematically assigns a binary score to each word according to presence (”1”), or absence (”0”). After processing each word in each segment, MEH generates an output file that identifies the words and shows in which segments they are used.

Moreover, frequency of each common word across all the text observation is computed. The output file can be uploaded into statistical software (e.g. R, SPSS). A Principal Component Analysis with varimax rotation allows to extract components reflecting broader themes across all texts.

3 Results

Top 50 most frequently used words are reported in table 1.

Table 1 50 most frequent words across 9462 segments

Word	Frequency	Word	Frequency
competition	3357	friend	1311
work	3063	strong	1294
jump	2427	complete	1197
community	2338	technique	1181
body	1980	level	1124
life	1979	ability	1120
flip	1807	fast	1096
freerun	1763	martial	1078
foot	1726	goal	1072
discipline	1673	walk	993
sport	1668	important	975
hard	1651	interest	956
vault	1606	difference	953
change	1516	focus	925
fun	1457	roll	925
mind	1439	hand	914
strength	1434	group	911
kid	1421	climb	908
physical	1417	challenge	905
experience	1400	fight	903
philosophy	1398	improve	901
wall	1379	practitioner	893
skill	1378	environment	890
movement	1350	human	879
love	1344	style	876

The top 50 most frequent words themselves reveal relevant aspects of parkour: words related to it as a concept (*competition, discipline, philosophy, sport*), as well as words reminding to social aspects (*community, friend, group, practitioner*) and more technical words (*jump, flip, vault, wall, roll*).

PCA results along with descriptives are reported in table 2. The parkour corpora can be summarized by 11 themes that globally accounted for the 26% of variance.

Table 2 Corpora descriptives

Component	1	2	3	4	5	6
Eigenvalue	2.95	1.81	1.69	1.62	1.59	1.47
Variance %	4.27	2.62	2.45	2.36	2.31	2.13
M (*SD*)	.08 (.27)	.08 (.27)	.08 (.26)	.08 (.26)	.08 (.26)	.09 (.28)
Component	**7**	**8**	**9**	**10**	**11**	
Eigenvalue	1.45	1.43	1.42	1.36	1.32	
Variance %	2.1	2.07	2.05	1.97	1.91	
M (*SD*)	.08 (.26)	.09 (.28)	.08 (.27)	.07 (.25)	.08 (.26)	

Notes. The Kaiser-Meyer-Olkin measure was .79, above the recommended value of .6 [14, 15], and Bartlett's Test of Sphericity [4] reached statistical significance, thus confirming the correlation matrix of components.

Table 3 illustrates the 11 themes emerged from PCA analysis, their meaning and the loadings.

Table 3 Component information and loadings

Component	1 - Being Practitioner	2 - Values	3 - Outdoor	4 - Technique
	ability .559	discipline .582	climb .679	land .68
	strength .546	practitioner .482	wall .613	roll .65
	physical .56	philosophy .361	vault .461	technique .353
	improve .465	sport .329	jump .432	jump .35
	method .451	community .323	building .358	body .32
	skill .398	freerun .315	small .178	hurt .282
	challenge .395	martial .29	environment .176	vault .268
	strong .375	experience .278	efficient .176	basic .235
	body .318	purpose .274		ground .213
	life .314	competition .27		strength .187
	goal .296	respect .247		strong .17
	movement .277	physical .233		
	environment .271	involve .218		
	important .238	important .18		
	limit .236	environment .179		
	face .229			
	focus .221			
	work .217			
	hard .213			
	technique .212			
	basic .21			
	experience .2			
	practitioner .178			
	mind .169			

Component	5 - Freerunning	6 - Meetings	7 - Mindset	8 - Interactions
	flip .616	jam .454	mind .53	kid .464
	trick .589	group .438	open .453	fun .418
	freerun .536	community .45	human .362	friend .372
	movement .278	together .43	body .328	life .357
	efficient .231	work .33	change .33	love .318
	style .24	basic .254	community .245	sport .252
	fun .189	small .229	close .24	honest .24
	competition .167	experience .229	competition .229	money .29
	experience -.206	month .24	turn .214	hard .186
		human -.252	grow .199	hurt .179
		body -.244	involve .185	philosophy .177
			goal .179	
			purpose .17	

Component	9 - Competition	10 - Martial Art	11 - Injury
	push .588	martial .593	face .55
	limit .43	fight .534	head .379
	competition .418	style .477	ground .374
	sport .31	technique .291	fall .371
	hard .262	focus .245	fight .351
	focus .255	movement .212	hand .288
	money .26	efficient .212	turn .248
	challenge .176	hand .183	hurt .17
	efficient -.217		sport -.207

Values with an absolute value potentially rounded to .2 were used in order to identify PCA components [11, 9]. The themes emerged have been evaluated (in terms of meaning) by both the first author and an italian parkour instructor (see the *acknowlegdement* section for more info). They cover many facets of the parkour discipline, related both to the global status of being a practitioner (*component 1*), but also to more specific dimensions: the idea of parkour as discipine and philosophy (*component 2*) is largely reported by the most conservative practitioners on a smaller scale [12, 19].

The *outdoor* and *technique* components mention specific parkour moves (*vault, jump, wall, roll*) and environment words (*building, ground*). It is interesting to note that the word *efficient* loads positively within the *outdoor* component, whereas it loads negatively in the *competition* component suggesting, in agreement with literature [3, 12], that when it comes to moving in real environments, efficiency matters more than spectacularity or aesthetic . This is also coherent with Georges Hébert's principle that inspired parkour: *be strong to be useful* [19].

Also, references to parkour *jams*, that is, intensive training meetings, as well as a strong social component appeared: in the context of an open-minded attitude towards discipline and life, a precious contribution to personal and physical development comes from confrontation and advise of other *traceurs*, and from sharing both sweat and goals.

4 Conclusions

In this paper, Meaning Extraction Method was used to web-scrape and analyse data from the official American Parkour forum. The method provided an optimal dimensionality reduction and identify some latent themes that underlay the vision of a Parkour discipline.

The obtained results showed that the Parkour discipline confirmed and extended the vision of a phenomenon involving global existence and perceived as a way of life.

Acknowledgements A special thanks to Ryan L. Boyd for providing the first author with the largest parkour dataset so far and introducing her to MEM.

A special thanks to Alessandro Pennella, one of the best certified parkour trainers in Italy, for examining and commenting the themes emerged from the analysis.

References

[1] APK, American Parkour Forum. http://americanparkour.com/smf/. Accessed: 2013-09-10.

[2] Helium Scraper. Extract data from any website. http://www.heliumscraper.com/en/index.php?p=home. Accessed: 2013-09-10.

[3] M. Atkinson. Parkour, anarcho-environmentalism, and poiesis. *Journal of sport and social issues*, 33(2):169–194, 2009.

[4] M. S. Bartlett. A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 296–298, 1954.

[5] N. Bavinton. From obstacle to opportunity: Parkour, leisure, and the reinterpretation of constraints. *Annals of leisure research*, 10(3-4):391–412, 2007.

[6] K. G. Blackburn, G. Yilmaz, and R. L. Boyd. Food for thought: Exploring how people think and talk about food online. *Appetite*, 123:390–401, 2018.

[7] R. Boyd and J. Pennebaker. A way with words: Using language for psychological science in the modern era, 10 2015.

[8] R. L. Boyd. Meaning Extraction Helper (Version 1.2.65)[Software], 2015. URL http://meh.ryanb.cc.

[9] R. L. Boyd, S. R. Wilson, J. W. Pennebaker, M. Kosinski, D. J. Stillwell, and R. Mihalcea. Values in words: Using language to evaluate and understand personal values. 2015.

[10] G. G. Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[11] C. K. Chung and J. W. Pennebaker. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality*, 42(1):96–132, 2008.

[12] J. L. Clegg and T. M. Butryn. An existential phenomenological examination of parkour and freerunning. *Qualitative research in sport, exercise and health*, 4(3):320–340, 2012.

[13] P. Gilchrist and B. Wheaton. Lifestyle sport, public policy and youth engagement: examining the emergence of parkour. *International journal of sport policy and politics*, 3(1):109–131, 2011.

[14] H. F. Kaiser. A second generation little jiffy. *Psychometrika*, 35(4):401–415, 1970.

[15] H. F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, 1974.

[16] D. Murthy, A. Gross, A. Takata, and S. Bond. Evaluation and development of data mining tools for social network analysis. In *Mining Social Networks and Security Informatics*, pages 183–202. Springer, 2013.

[17] A. O'grady. Tracing the city–parkour training, play and the practice of collaborative learning. *Theatre, dance and performance training*, 3(2):145–162, 2012.

[18] N. Ramírez-Esparza, C. K. Chung, G. Sierra-Otero, and J. W. Pennebaker. Cross-cultural constructions of self-schemas: Americans and mexicans. *Journal of Cross-Cultural Psychology*, 43(2):233–250, 2012.

[19] S. J. Saville. Playing with fear: Parkour and the mobility of emotion. *Social & cultural geography*, 9(8):891–914, 2008.

Players Movements and Team Shooting Performance: a Data Mining approach for Basketball.

Movimenti dei giocatori e performance di squadra nella pallacanestro con l'uso di tecniche di Data Mining.

Rodolfo Metulini

Abstract In the domain of Sport Analytics, Global Positioning Systems devices are intensively used as they permit to retrieve players' movements. Team sports' managers and coaches are interested on the relation between players' patterns of movements and team performance, in order to better manage their team. In this paper we propose a Cluster Analysis and Multidimensional Scaling approach to find and describe separate patterns of players movements. Using real data of multiple professional basketball teams, we find, consistently over different case studies, that in the defensive clusters players are close one to another while the transition cluster are characterized by a large space among them. Moreover, we find the pattern of players' positioning that produce the best shooting performance.

Abstract *N*el dominio attinente Sport Analytics, i dispositivi di posizionamento globale vengono utilizzati intensivamente poiche' consentono di raccogliere e analizzare i movimenti dei giocatori. Managers e allenatori sono interessati a conoscere la relazione tra i movimenti dei propri giocatori e le prestazioni della squadra, al fine di gestire al meglio la loro squadra. In questo articolo proponiamo un approccio che utilizza la Cluster Analysis e il Multidimensional Scaling con l'obiettivo di identificare e descrivere specifiche dinamiche di movimento. Usando dati reali di più squadre di basket professionistiche, troviamo, consistentemente su diversi casi di studio, che le azioni di difesa si caratterizzano per avere giocatori tra loro vicini, mentre le azioni di transizione presentano un' ampia spaziatura tra di essi. Inoltre è stato trovato il posizionamento in campo che meglio si associa con una buona performance di tiro.

Key words: Sport Statistics; Basketball; Team Performance; Sensor Data; Data Mining; Cluster Analysis

Rodolfo Metulini
Department of Economic and Management, University of Brescia, Contrada Santa Chiara, 50, 25122 Brescia, e-mail: rodolfo.metulini@unibs.it

1 Introduction

Studying the interaction between players in the court, in relation to team performance, is one of the most important issue in Sport Science, as team sports' Managers, more and more in recent years, are becoming aware of the potential of Data Analytics in order to better manage their team. Recent years make it possible, thanks to the advent of Information Technology Systems (ITS), that permits to collect, store, manipulate and process a large amount of data. On the one hand, a sequence of relevant events of the match, such as passes, shots and fouls (player-specific) and time-outs (team-specific) takes the name of play-by-play. On the other hand, information on the movement of players on the court has been captured with the use of appropriate Geographical Positioning Systems (GPS) devices, for example the accelerometer, a device that measures proper velocity and positioning. Analysing players' interaction, however, is a complex task, as the trajectory of a single player depends on a large amount of factors related, among others, to coaches, single players and the whole team. The trajectory of a player depends on the trajectories of all the other players in the court, both teammates and opponents. Players interactions have been mainly studied in the new domain of ecological dynamics [1, 2]. Typically, there are certain role definitions in a sports team that influence movements. Predefined strategies are used by the coach to achieve specific objectives. A common method to approach with this complexity in team sport analysis consists on segmenting a match into phases, as it facilitates the retrieval of significant moments of the game. For example, Perin et al. [3] developed a system for visual exploration of phases in football, while, to the same goal, Metulini [4] propose motion charts. Cluster analysis methodology is widely used in team sports literature. To name a few, Sampaio and Janeira [5] applied a cluster analysis to investigate the discriminatory power of game statistics between winning and losing teams in the Portuguese Professional Basketball League, by using game final score differences, Ross [6] uses cluster analysis to segment team sport spectators identifying potential similarities according to demographic variables. Csataljay et al. [7] used cluster approach to the purpose of identifying those critical performance indicators that most distinguish between winning and losing performances. However, differently from the aforementioned papers, to the aim of segmenting game into phases, in this paper we cluster time instants. In doing so, we use GPS tracked data. In this regard, Goncalvez [8] applied a two-step cluster to classify the regularity in teammates dyads' positioning. Metulini, Manisera and Zuccolotto [9] used cluster analysis to an amatorial basketball game in order to split the match in a number of separate time-periods, each identifying homogeneous spatial relations among players in the court. They also adopt a Multidimensional Scaling to visually characterize clusters and analysed the switch from *defense* to *offense* clusters, by mean of transition probabilities. This paper aims to fill the gap in Metulini et al., by extending the analysis to multiple matches. Moreover: i) we apply our cluster analysis procedure to professional basketball games, ii) we use the data generated by the algorithm proposed in Metulini [10] in order to consider active game moments only, iii) we use a more detailed labelling scheme introducing *transition* moments, which permits a better interpretation of the transition probabilities. Last, we characterize clusters in term of team performance, by retrieving shooting events throughout a video analysis.

2 Data and Methods

Basketball is a sport generally played by two teams of five players each on a rectangular court. The objective is to shoot a ball through a hoop 46 centimeters in diameter and mounted at a height of 3.05 meters to backboards at each end of the court. According to FIBA rules, the match lasts 40 minutes, divided into four periods of 10 minutes each. There is a 2-minutes break after the first quarter and after the third quarter of the match. After the first half, there is a 10 to 20 minutes half-time break. In this paper we use tracked data from three games played by Italian professional basketball teams, at the Italian Basketball Cup Final Eight. MYagonism (https://www.myagonism.com/) was in charge to set up a system to capture these data during the games, trough accelerometer devices. Each player worn a microchip that, having been connected with machines built around the court, collected the player's position (in pixels of 1 cm^2 size) in the x-axis (court length), the y-axis (court width), and in the z-axis (height). Data, filtered with a Kalman approach, has been detected at a millisecond level. Available data contain information on players' positioning, velocity and acceleration during the full game length. Throughout the text we will call the three games case study 1 (CS1), case study 2 (CS2) and case study 3 (CS3). As the initial dataset is provided to us considering the full game length, we cleaned it by dropping the pre-match, the quarter- and the half-time intervals and the post match periods, as well as the time-outs and the moments when a player is shooting a free-throw. More information on this filtering procedure can be found in Metulini [10]. The final dataset for CS1 counts for $206,332$ total rows, each identifying the milliseconds in which the system captured at least one player. CS2 dataset counts for $232,544$ rows, while CS3 counts for a total of $201,651$ rows.

We apply a k-means Cluster Analysis in order to group a set of objects. Cluster analysis is a method of grouping a set of objects in such a way the objects in the same group (clusters) are more similar to each other than to those in other groups. In our case, the objects are represented by the time instants, expressed in milliseconds, while the similarity is expressed in terms of distance between players' dyads. In the analyses that follows we only consider moments when a particular lineup is on the court. More specifically, we only consider lineups that played for at least 5 minutes. According to this criteria, we consider two lineups (*p1, p3, p6, p7, p8* and *p1, p4, p5, p7, p10*) for CS1, two (*p1, p2, p4, p5, p6* and *p1, p2, p5, p6, p8*) for CS2, and one lineup for CS3 (*p2, p5, p6, p9, p10*, *p* stays for player). We chose number of clusters based on the value of the between deviance (BD) / total deviance (TD) ratio and the increments of this value by increasing the number of clusters by one. We consistently, and surprisingly, find k=6 (BD/TD= around 45% along the different lineups, and relatively low increments for increasing k, for $k \geq 6$) for almost all the lineups considered. Specifically, increasing the number of clusters from 5 to

6, BD/TD increments by around 11-12 % in all the five lineups, while increasing from 6 to 7, BD/TD increments by around 6-7 %.

3 Results

In this section we describe clusters for their dimension and their characteristics in term of pattern of player's positioning and team performance, along the five lineups. According to the first lineup of CS1, the first cluster (C1) embeds 13.31% of the observations (i.e. 13.31% of the total game time), the other clusters, named C2, ..., C6, have size of 19.76%, 3.40%, 29.80%, 6.41% and 27.31% of the total sample size, respectively. Consistently for all the five lineups, we find a couple of small clusters, with less than 10% of the total observations, and 2-3 larger ones, containing at least 20% of the observations.

Cluster profile plots have been used to better interpret the players' spacing structure in each group. Figure 1 reports profile plot for the first lineup of CS1, to characterize groups in terms of average distances among players. In this case, we find the smaller cluster (C3, 3.4% of observations) displaying large average distances among players (horizontal lines in Figure 1 represent the average value along the game time played by that lineup). On the contrary, the larger cluster (C4, 29.8%) displays all the average distances below the game average. These two facts are confirmed in the second lineup of CS1, as it presents the larger cluster (C5, 40.4%, which is not reported for the sake of space saving) displaying really small average distances, while its smaller cluster (C6, 3.2%) reports large average distances. Same evidences have been found in other case studies.

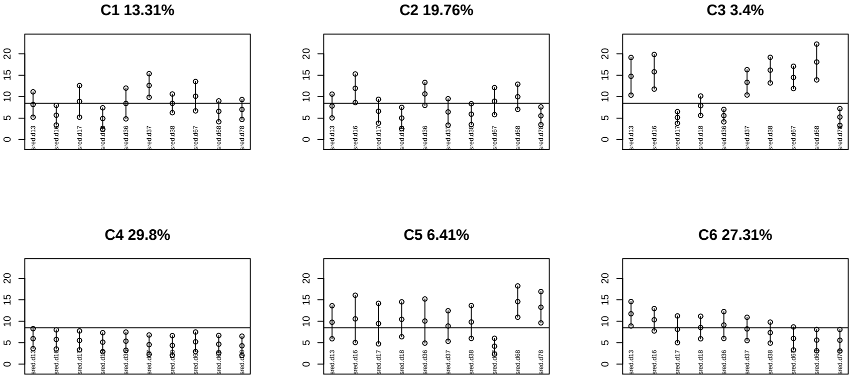


Fig. 1: Profile plots representing, for each of the 6 clusters, the average distance among players' dyads.

To the aim of producing further visual evidences, we used Multidimensional Scaling (MDS), which plots the differences between the groups in terms of positioning in the court. With MDS algorithm we aim to place each player in N-dimensional space such that the between-player average distances are preserved as well as possible. Each player is then assigned coordinates in each of the N dimensions. We choose N=2 and we draw the related scatterplots. Figure 2 reports the scatterplot for the first lineup of CS1. We observe strong differences between the positioning pattern among groups. The figure highlights large space among players in CS3, as also highlighted by the average distances in the profile plot. Moreover, moments in C4 are characterized by close to each others players. Despite not reported here, other lineups display similar MDS results: smaller clusters are characterized by large average distances and by a large space among players, while larger clusters by small average distances and by close to each others players.

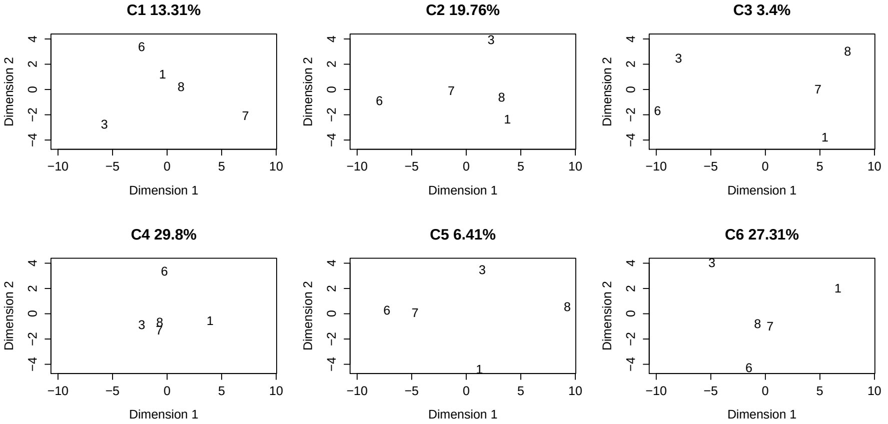


Fig. 2: Map representing, for each of the 6 clusters, the average position in the $x-y$ axes of the five players in the court, using MDS.

The filtered datasets label each moment as *offense*, *defense* or *transition* by mean of looking to the average x-axis positioning of the five players on the court. A moment is labelled as *transition* when the average x-axis is in within the interval [-4,+4], where 0 corresponds to the half court line. Throughout this information, we associate each cluster to offense, defense or to transition, according how many time instants in a specific cluster corresponds to the specific label.

Table 1 reports related percentages for the first lineup in CS1. Clusters C1, C2 and C6 mainly correspond to offense (respectively, for the 68.85%, 67.97% and 71.52% of the times), C3 and C5 correspond to defensive actions (82.11% and 54.49% of the times, respectively), while C4 corresponds to defense (70.48%). It emerges that large clusters with small average distances among players contains defensive moments. Moreover, the small cluster with large distances corresponds to transition. This result is consistent in all the five considered lineups. For example, in the sec-

Table 1 Percentages of time instants classified in Transition (TR) Defense (D) or Offense (O), for each cluster.

Cluster	**1**	**2**	**3**	**4**	**5**	**6**
TR	8.41	21.76	**82.11**	7.08	**54.49**	10.53
D	22.74	10.28	6.6	**70.48**	23.98	17.95
O	**68.85**	**67.97**	11.29	22.45	21.53	**71.52**
Total	100.00	100.00	100.00	100.00	100.00	100.00

ond lineup of CS1, the small cluster (C6) corresponds to transition moments for the 80.76% of the time. The large cluster with corresponding small distances (C5) contains moments classified as defense the 72.99% of the times.

Table 2 shows the transition matrix for the first lineup of CS1, which reports the relative frequency in which subsequent moments in time report a switch from a cluster to a different one. Main diagonal values of the matrix have been set to zero, so that each column percentages sum to 100% without including subsequent moments in which there isn't a switch.

Table 2 Transition matrix reporting the relative frequency subsequent moments $(t, t+1)$ report a switch from a group to a different one.

Cluster label	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
C1	0.00	11.27	10	8.45	15	10.34
C2	31.03	0.00	10	23.94	15	35.34
C3	0.00	1.41	0	0.00	0	7.76
C4	34.48	21.13	0	0.00	25	35.34
C5	3.45	4.23	0	4.23	0	11.21
C6	31.03	61.97	80	63.38	45	0.00

It emerges that, for the 34.48% of the times C1 switches to a new cluster, it switches to C4. It also switch to C2 and C6, respectively for the 31.03% and for the 31.03% of the times. We can note that C2, C4 and C6 are the three largest clusters. C3, marked as *Transition*, switches 80% of the times to C6 (a offensive cluster). Moreover, C2 switches most of the times (61.97%) to C6. C2 and C6 are both marked as offensive clusters. Since the total number of switches for this lineup is equal to 309, and this lineup played for a total of 8 minutes and 21 seconds, on average we have a switch every 2 seconds. For this reason we have frequent cluster switches during the same action. Switch from C2 to C6 is an example: players change patterns of positioning during the same action. Table 4 highlights that offensive clusters often switch to another offensive cluster (beside C2, C1 switches for the 31.03% of the times to C2 and for the 31.03% of the times to C6, C6 switches for the 35.34% of the times to C2). This evidence is confirmed in the other case studies, since, in the second lineup of CS1, we have three offensive clusters (C1, C2 and C3); C1 switches to C2 for the 33.33% of the times, C3 switches to C2 for the 40.91% of the times. In the first lineup of C2, we find three offensive clusters (C3, C4 and C6); C4 switches to C6 for the 75.86% of the times.

With this in mind, we can not associate a cluster with a whole action played with a particular tactic, instead, we have to interpret offensive clusters as subsequent players' positioning configurations, to the aim of finding the best positioning for a good shot.

In light of this, we collect the shooting events of the match. Since play-by-play data are not available for this tournament, we collect such events by watching the video of the game. Zuccolotto, Manisera, Sandri [11] analysed the shooting performance under pressure. Here we study shooting performance with respect to different players' positioning patterns, by associating shots to the cluster in which the team was (at the moment of the shot). We take into consideration only shots from the court, disregarding free throws. During the 8 minutes and 21 seconds players *p1, p3, p6, p7* and *p8* were in the court together, the team made 15 shots from the court, with 7 made shots and 8 missed, for a percentage of 46.67%. We find that most of these shots (8) has been attempted in moments that belongs to cluster C6. During this cluster, the team scored 5 out of 8 total attempts, with a really high percentage of 62.5%, ways higher than the average of 46.67%. Moreover, the team attempted only 4 shots (2 of them made) during cluster C1, only 2 shots (both missed) during cluster C2 and it missed a shot during cluster C5. So, 14 out of 15 shots have been attempted during the clusters labelled as offensive (i.e. C1, C2 and C6) while only one during a transition cluster (C5). Looking to bottom-right chart in Figure 2 (C6), we find player 3 far away from the others. We could suppose that the tactic of the team was to leave that player free to shot on the weaker side of the court. Results support the idea that C6 represents the cluster of (good) shooting moments. Furthermore, the other offensive (C1, C2) and transition (C3, C5) clusters often switch to cluster C6, which support our hypothesis of subsequent game configurations to the aim of finding the best positioning for a good shot: the best positioning to shot is that in C6 moments.

4 Conclusions

In recent years, the availability of ‘big data" in Sport Science increased the possibility to extract insights from the games that are useful for managers and coaches, as they are interested to improve their team's performances. In particular, with the advent of Information Technology Systems, the availability of players' trajectories permits to analyse the space-time patterns with a variety of approaches. With this paper we pursue the points raised by Metulini et al. [9] as suggestions for future research, by analyzing multiple professional games and relate clusters with team shooting performance. We segmented the game into phases of play and we characterized each phase in terms of spacing structure among players, relative distances among them and whether they represent an offensive, a defensive or a transition play, finding substantial differences among different phases. Moreover, we analysed this structure in terms of shooting performance, finding the cluster corresponding to the best shooting performance. These results shed light on the potentiality of data-mining methods for players' movement analysis in team sports. In future research we aim to better explain the relation between players' positioning and team performance, adding more play-by-play data and analysing this relationship for a larger amount of time and for multiple matches.

Acknowledgements Research carried out in collaboration with the Big&Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016, title Big Data Analytics in Sports, www.bodai.unibs.it/BDSports/), granted by Fondazione Cariplo and Regione Lombardia. Authors would like to thank MYagonism (https://www.myagonism.com/) for having provided the data and Paola Zuccolotto (University of Brescia) for the useful discussions.

References

1. Travassos, B., Davids, K., Araujo, D., Esteves, P. T.: Performance analysis in team sports: Advances from an Ecological Dynamics approach. International Journal of Performance Analysis in Sport 13.1 (2013): 83-95.
2. Passos, P., Araujo, D., Volossovitch, A.: Performance Analysis in Team Sports. Routledge (2016).
3. Perin, C., Vuillemot, R., Fekete, J. D.: SoccerStories: A kick-off for visual soccer analysis. IEEE transactions on visualization and computer graphics 19.12 (2013): 2506-2515.
4. Metulini, R.: Spatio-Temporal Movements in Team Sports: A Visualization approach using Motion Charts. Electronic Journal of Applied Statistical Analysis Vol. 10.3 (2017): 809-831.
5. Sampaio, J., Janeira, M.: Statistical analyses of basketball team performance: understanding teams' wins and losses according to a different index of ball possessions. International Journal of Performance Analysis in Sport 3.1 (2003): 40-49.
6. Ross, S. D.: Segmenting sport fans using brand associations: A cluster analysis. Sport Marketing Quarterly, 16.1 (2007): 15.
7. Csataljay, G., O'Donoghue, P., Hughes, M., Dancs, H.: Performance indicators that distinguish winning and losing teams in basketball. International Journal of Performance Analysis in Sport 9.1 (2009): 60-66.
8. Gonçalves, B. S. V.: Collective movement behaviour in association football. UTAD Universidade de Tras-os-Montes e Alto Douro (2018)
9. Metulini, R., Marisera, M., Zuccolotto, P.: Space-Time Analysis of Movements in Basketball using Sensor Data. Statistics and Data Science: new challenges, new generations SIS2017 proceeding. Firenze Uiversity Press. eISBN: 978-88-6453-521-0 (2017).
10. Metulini, R.: Filtering procedures for sensor data in basketball. Statistics&Applications 2 (2017).
11. Zuccolotto, P., Manisera, M., Sandri, M.: Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. International Journal of Sports Science & Coaching (2017): 1747954117737492.

Supporting Regional Policies through Small Area Statistical Methods

Survey-weighted Unit-Level Small Area Estimation

Jan Pablo Burgard and Patricia Dörr

Abstract For evidence-based regional policy making, geographically differentiated estimates of socio-economic indicators are the basis. However, national surveys are often conducted under a complex sampling design due to diverse reasons. Often small sample sizes result within regions of interest leading to too inefficient classical design-based estimators for policy making. In this case, the methodology of small area estimation (SAE) is applicable. Classical SAE relies on the assumption of a multi-level regression model underlying the population data and presumes the sample design to be non-informative. These assumptions are hard to verify in practice. Under an informative sample design, estimated regression parameters are biased and the model-consistency of SAE gets lost. We correct for the sample informativeness in the parameter estimates, and construct design- and model-consistent estimates for regional indicators. Besides the estimation procedure we also propose a MSE estimator. In a simulation study, we illustrate the necessity of survey weights under the violation of typical SAE assumptions. Furthermore, we show that the proposed method is also applicable to generalized linear mixed model settings, allowing also for non-continuous dependent variables.

Key words: Small area estimation, generalized linear mixed models, survey-weighting

Jan Pablo Burgard
Research Institute for Official and Survey Statistics (RIFOSS), Trier University, Universitätsring 15, D-54295 Trier, e-mail: burgardj@uni-trier.de

Patricia Dörr
Research Institute for Official and Survey Statistics (RIFOSS), Trier University, Universitätsring 15, D-54295 Trier e-mail: doerr@uni-trier.de

1 Introduction

National Statistical Institutes often conduct surveys using a complex survey design, either due to costs or due to optimality considerations at the national level. This may lead to small sample sizes in certain geographic regions for which an estimate can be of interest, though. Small sample sizes lead to high variances of the classical design-based estimators and lead to the tradional set-up of the small area estimation (SAE) framework. In SAE borrowing strength across small domains and thus an increased efficiency is attained using a regression model. The inclusion of a random effects term - whose realization is area-specific - the regression model is called mixed. SAEs are usually composite of such random effects predictions and the realized sample's estimate. However, when the model - that is estimated on the realized sample - does not correspond to the population model for some reason, these procedure returns biased estimates. Complex survey designs or model misspecification may contribute to such non-correspondence between the sample and population model. In general, survey weights contain information about the sampling design and thus, their inclusion in the regression model component can reduce possible model bias.

We consider the case where the mixed model is estimated on the sampling unit, i.e. unit-level SAE in the sense of [2]. Usually, estimation is done using the sample log-likelihood, which requires integration over unit-likelihoods in a given area such that unit-specific weighting is not straight forward (cf, for example [13]). Existing proposals require units sharing one random effect to have the same weight, because the likelihood is expressed as a nested integral and elements within an integral cannot be weighted differently. This implies that the random effects structure reflects the sampling clustering and thus are nested [12], [13]. This is quite restrictive and furthermore, access to sampling stage specific inclusion probabilities is seldom for final data users. Therefore, a more general estimation procedure that allows for crossed and nested random effects and that requires only final survey weights is needed. The Expectation-Maximization (EM) methodology ([5]) that is applicable to mixed effects models in general ([7]) provides a framework that is applicable to these needs. However, as one could also think of dependent variables stemming from other exponential family distributions such as binary or count data, we employ a Monte-Carlo version (Monte-Carlo EM algorithm, MCEM) that replaces the E-step by a Monte-Carlo approximation ([10], [3]). The specific survey-weighting application is outlined in [4].

Section 2 introduces the algorithm and turns on problems of the MC-integration. Section 3 handles consistency considerations and Mean Squared Error (MSE) estimation. Afterwards, a simulation study demonstrates the possible gains of the survey-weighted SAE estimator. The final section discusses possible further research and concludes.

2 Proposed Estimators

2.1 Likelihood Set-up

We use the Monte-Carlo EM-algorithm adapted to survey-weighted Generalized Linear Mixed Models (GLMMs) as proposed in [4]. Here, we give a brief review of the set-up from which the SAE point estimator result. Consider as data generating process (DGP) a GLMM described through

$$\eta_i = \mathbf{x}_i^T \beta + \mathbf{z}_i^T \gamma \tag{1}$$

$$\mu_i = g(\eta_i) \tag{2}$$

$$Y_i \sim F(\mu_i, \varphi) \tag{3}$$

$$G \sim N(\mathbf{0}, \Sigma) \quad , \tag{4}$$

where γ is a realization of the mulitvariate normal random variable G (with normal density function $\phi(\cdot|\sigma)$), $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ and $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^T$ being matrices of explanatory variables in $\mathbb{R}^{N \times p}$ and $\mathbb{R}^{N \times q}$ respectively, β a superpopulation parameter vector of fixed effects and F a distribution from the exponential family. Consequently, g denotes the inverse link function between expectation μ_i and linear predictor η_i. We consider here the canonical link functions under which

$$\log f(y_i, \eta_i) = \frac{y_i \eta_i - b(\eta_i)}{a(\varphi)} + c(y_i, \varphi) \tag{5}$$

is the logarithmic density function. The covariance matrix Σ only depends on a parameter vector σ of length $q^\star$. Define the vector of model parameters by $\psi = (\beta^T, \sigma^T)^T$. Let $\mathscr{U}$, $|\mathscr{U}| = N$ be an index set and for $i \in \mathscr{U}$, let y_i be a realization of Y_i. Then, the population log-likelihood is

$$\mathscr{L}\mathscr{L}(\mathbf{y}, \psi, \gamma) = \sum_{i \in \mathscr{U}} \log f(y_i|\gamma, \beta) + \log \phi(\gamma|\sigma) \quad , \tag{6}$$

and if the random effect vector γ was known, a natural survey-weighted version of (6) would be the Horvitz-Thompson estimator for totals ([6]). However, this is not the case. But a γ simulated from the correct distribution may serve as a plug-in and then (6) can be estimated through a HT-estimator. Repeating the simulation and averaging yields then the (model) expectation of the HT-estimator (denoted by $\mathrm{E}(\widehat{\mathscr{L}\mathscr{L}_{\mathscr{S}}})$), that is maximized in the M-step. Computational and conceptual difficulties that must be taken into account with that procedure are discussed in [4]. As the EM-algorithm applied on the expectation of the estimand of (6) converges to a saddlepoint of the likelihood and the latter is the weighted sum of strictly convex unit-likelihood: The first summand is a generalized linear model component, whose maximum likelihood (ML) estimator is unique for the canonical link (cf. [16]) and the second summand in (6) is a normal log-density, whose ML is unique, too. The proposed procedure thus converges to the maximum likelihood (ML) estimator of the population likelihood if the survey-weights reflect the selection process into the sample. In contrast, an unweighted sample log-likelihood rather converges to the ML of the *sample's* data generating process, composed of the population DGP and the sample randomization. Thus, the survey weighting may protect against some violations of the SAE assumptions (cf. [14]): A non-ignorable sample design.

2.2 SAE Estimator

Having the estimates of the model components, $\hat{\psi}$, we can define our estimator, a weighted unit-level estimator (wUL) of the finite population mean of variable Y. For the pairwise disjoint subpopulations $\mathscr{U}_d$, $|\mathscr{U}_d| = N_d$ and $\mathscr{U} = \uplus_{d=1}^{D} \mathscr{U}_d$, we propose three alternative estimators for the finite population mean in domain d, $\bar{y}_d = N_d^{-1} \sum_{i \in \mathscr{U}_d} y_i$. The notation $\bar{y}_d$ is chosen in order to differentiate between the expectation under the superpopulation model, μ_d, and the finite population realization, which is the focus here. The first alternative is

$$\hat{\mu}_d^{wUL} = N_d^{-1} \sum_{i \in \mathscr{U}_d} \hat{\mu}_i, \tag{7}$$

which is similar to [14, eq 5.3.7] in the linear case and [14, eq 9.4.20] for binary data. This version does not include any finite population correction, which however, might be negligible in a small area setting where the sampling fractions are rather small. Another version that incorporates the sampled observations y_i is

$$\hat{\mu}_d^{wUL} = N_d^{-1} \left(\sum_{i \in \mathscr{U}_d} \hat{\mu}_i + \sum_{i \in \mathscr{S} \cap \mathscr{U}_d} (y_i - \hat{\mu}_i) \right). \tag{8}$$

In both equations, $\hat{\mu}_i$ is the prediction of individual i's variable Y_i expectation under the estimated vector $\hat{\psi}$ and the mode $\hat{\gamma}$ of the weighted sample likelihood under $\hat{\psi}$

$$\hat{\mu}_i = g(\mathbf{x}_i^T \hat{\beta} + \mathbf{z}_i^T \hat{\gamma}), \tag{9}$$

where the random effects only can be predicted for those areas that have at least one observation in the sample. In the LMM setting, (8) is a generalization of the classical unit-level estimator introduced in [2], which only incorporates a random intercept and does not include any additional survey information. That means, in [2], the survey weights are considered to be equal across areas and units.

Another option for SAE would be a model-assisted version

$$\hat{\mu}_d^{wUL} = N_d^{-1} \left(\sum_{i \in \mathscr{U}_d} \hat{\mu}_i + \sum_{i \in \mathscr{S} \cap \mathscr{U}_d} w_i (y_i - \hat{\mu}_i) \right). \tag{10}$$

Estimator (10) is similar to the Generalized Linear Regression Estimator proposed in [15] and [8] in the logistic regression setting. However, note that the version (10) is based on a GLMM in lieu of GLM and incorporates area-specific information through the random effect prediction $\hat{\gamma}$, which makes this model-assisted estimator especially applicable to SAE settings.

2.3 MSE Estimation

As our proposed estimators are smooth functions of ψ and γ, asymptotic MSE estimation fits into the framework of [11] who even deal with non-smooth SAE estimators for poverty analysis. Therefore, under the (common) regularity conditions given in [11] (that we also partially assume in the previous sections in order to establish design-consistency of the point estimators), an asymptotic MSE of (10) consists of the design variance of the model residuals in the domain under consideration. For the model-based estimators (7) and (8), however, MSE estimation is more difficult. We suggest a first-order Taylor approximation of the predictions $\hat{\mu}_i$ at the population parameters $(\beta^{pop}, \gamma^{\star})^T$ - $\gamma^{\star}$ being the population likelihood mode and then calculating the variance of the linearized predictions. This requires variance estimators for the fixed effects estimates and the random effect predictions. [9] gives a formula on how to approximate the Hessian of the observed data matrix in the EM-algorithm and its application for the fixed effects parameter is also proposed in [3]. A lower bound of the prediction error of $\hat{\gamma}$, on the other hand, is the inverse Hessian of the log-likelihood evaluated at the ML-estimates of $\hat{\beta}$. An approximation of the inverse Hessian is readily available from the specific MCEM-estimation algorithm discussed in [4]. However, note that this suggestion is only a lower bound for MSE estimation and hold only asymptotically.

3 Simulation Study

We present a small simulation study in order to demonstrate the necessity of survey-weighting when the non-informativeness assumption is violated. We therefore generate a fixed population $\mathscr{U}$, $|\mathscr{U}| = 3000$ under the following superpopulation model where the population is made up of 50 pairwise disjoint domains $\mathscr{U} = \dot{\cup}_{d=1}^{50} \mathscr{U}_d$, $|\mathscr{U}_d| = 60$, and $X_1 \sim N(6, 3^2)$ and $X_2 \sim \text{Exp}(3)$. The DGP is

$$\eta_i = 30 - 3x_{1,i} - 8x_{2,i} + \gamma_d, \quad i \in \mathscr{U}_d \tag{11}$$

$$\gamma_d \sim N(0, 2^2) \tag{12}$$

$$\mu_i = g(\eta_i), \qquad g \in \{\text{id}, \text{logit}^{-1}\} \tag{13}$$

$$Y_i \sim \begin{cases} N\left(\mu_i, (2.3)^2\right) \\ \text{Ber}(\mu_i) \end{cases} . \tag{14}$$

We draw $B = 1500$ (equally allocated) stratified samples $\mathscr{S}$ of size $n = 200$ of a finite population generated in this way. Consequently, $\mathscr{S} \cap \mathscr{U}_d = \mathscr{S}_d$ and $|\mathscr{S}_d| = n_d = 4$. This is a relatively easy set-up and if the sampling design is non-informative, the estimation conditions for the BHF ([2]) are optimal.

We contrast a πps design (which is under the correct model specification non-informative) where the inclusion probability π_i for a unit in domain d equals

$$\pi_i = \max\left\{\frac{x_{2,i}}{\sum_{j\in\mathscr{U}_d} x_{2,j}} \cdot n_d, 1\right\} \tag{15}$$

and an informative design where the inclusion probability of unit i in domain d is calculated in three steps:

$$e_i = y_i - \mu_i \tag{16}$$

$$q_i = \begin{cases} 0.1 \text{ if } e_i \text{ is below the 0.25 quantile} \\ 0.2 \text{ if } e_i \text{ is between the 0.25 and 0.5 quantile} \\ 0.4 \text{ if } e_i \text{ is above the 0.5 quantile} \end{cases} \tag{17}$$

$$\tilde{\pi}_i = \max\left\{\frac{q_i}{\sum_{j\in\mathscr{U}_d} q_j} \cdot n_d, 1\right\} \quad . \tag{18}$$

As observations with a bigger residual tend thus to be oversampled, the intercept estimator of the unweighted model estimation is expected to be overestimated which in return should yield biased predictions.

We compare the proposed estimator (8) (that has conceptually the closest similarity to the traditional BHF) to the Generalized Regression estimator (GREG), the BHF estimator and another survey-weighted SAE estimator, the You-Rao estimator (YR, cf. [17]). In the case of the binary outcome, we consider the logistic regression estimator (LGREG, cf. [8]), too, and the BHF is estimated like (8), but the underlying regression is a generalized linear mixed model estimated with the R-Package `lme4` ([1]). Due to construction, the GREG and the YR estimate a linear model for the binary outcome, too.

The quality criteria that we assess are the relativeempirical bias and the relative empirical mean squared error (MSE) of an estimator $\hat{\mu}$ over all domains:

$$\text{relBias} \quad = \tfrac{1}{50}\sum_{d=1}^{50} \tfrac{1}{1500}\sum_{b=1}^{1500} \tfrac{\hat{\mu}_{d,b}-\mu_d}{\mu_d} \tag{19}$$

$$\text{relMSE} = \tfrac{1}{50}\sum_{d=1}^{50} \tfrac{1}{1500}\sum_{b=1}^{1500} \left(\tfrac{\hat{\mu}_{d,b}-\mu_d}{\mu_d}\right)^2 \tag{20}$$

Results are listed in table (1). The results under the non-informative design and the gaussian outcome variable are standard: Survey-weights are not needed for consistent estimation and inflate the estimators. Consequently, the BHF is the most efficient estimator with respect to relative mean squared error. However, we note that the loss of efficiency when applying survey weights is low and thus might be recommended as the model assumptions are usually not verifiable. Furthermore, we find that the proposed estimator wML can compete with YR.

Under a binary variable of interest, we find that results are similar to the linear mixed model case and both GREG and YR perform well though they employ a linear regression model. Nonetheless, the LGREG is the less biased estimator and has a lower relative MSE than the GREG.

Table 1 Simulation Results

Sampling design	Outcome Variable	Estimator	relBias	relMSE
Non-informative	Normal	GREG	0.00457	0.10959
		wML	0.00974	0.0248
		YR	0.0212	0.02504
		BHF	0.01264	0.01759
	Binary	GREG	0.00466	0.07273
		LGREG	0.00242	0.02196
		wML	0.00682	0.00456
		YR	0.01076	0.00752
		BHF	0.00647	0.00454
Informative	Normal	GREG	0.00312	0.04289
		wML	0.04555	0.02394
		YR	0.05136	0.02402
		BHF	0.12068	0.03189
	Binary	GREG	0.00594	0.04658
		LGREG	0.00742	0.02256
		wML	0.02744	0.00564
		YR	0.05692	0.01108
		BHF	0.06463	0.01033

Under the informative sampling design, results change remarkably, though. As expected, the unweighted traditional BHF has an unacceptable high relative bias, although the relative MSE may compete with the GREG. In the continuous dependent variable case, wML and YR return comparable results. But when the dependent variable is binary, the suggested method outperforms BHF due to the inclusion of survey weights and the YR due to the GLMM framework employed. Thus, wML gives any of the four presented cases a good balance between bias and variance.

4 Conclusion

In this paper, we propose the use of a GLMM estimation framework that allows the inclusion of unit-specific survey weights in the estimation process in order to protect unit-level based SAE estimators against sampling informativeness. A linearization and/ or residual based MSE estimation of the suggested estimators is discussed. Finally, a simulation study demonstrates the necessity of survey weights in the model estimation step in order to reduce the SAE estimators' bias, both in a continuous variable set-up as well as for a binary variable of interest. We find that the loss in efficiency when survey weights need not be included in the model estimation but are nonetheless, is marginal. In contrast there are important gains when the sampling is informative. To conclude, we would like to note that the informativeness of the survey design is hard to verify in real world application and may also depend on the analyzed variable. Thus, we highly recommend the inclusion of survey weights in SAE analysis.

References

1. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823 (2014)
2. Battese, G. E., Harter, R. M., Fuller, W. A.: An error-components model for prediction of county crop areas using survey and satellite data. J. Am. Stat. Assoc. **83**(401), 28–36 (1988)
3. Booth, J. G., Hobert, J. P.: Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. Royal Stat. Soc., Series B (Methodological), **61**(1), 265–285 (1999)
4. Burgard, J.P., Dörr, P.: Survey-weighted Generalized Linear Mixed Models. Research Papers in Economics 2018-01, University of Trier, Department of Economics (2018)
5. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., Series B (Methodological), 1–38 (1977)
6. Horvitz, D. G., Thompson, D. J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc., **47**(260), 663–685 (1952)
7. Laird, N. M., Ware, J. H.: Random-effects models for longitudinal data. Biometrics, 963–974 (1982)
8. Lehtonen, R., Veijanen, A.: Logistic generalized regression estimators. Surv. Methodol., **24**, 51–56 (1998)
9. Louis, T. A.: Finding the observed information matrix when using the EM algorithm. J. Royal Stat. Soc., Series B (Methodological), 226–233 (1982)
10. McCulloch, C. E.: Maximum likelihood algorithms for generalized linear mixed models. J. Am. Stat. Assoc., **92**(437), 162–170 (1997)
11. Morales, D., del Mar Rueda, M., Esteban, D.: Model-Assisted Estimation of Small Area Poverty Measures: An Application within the Valencia Region in Spain. Soc. Indic. Res., 1–28 (2017)
12. Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., Rasbash, J.: Weighting for unequal selection probabilities in multilevel models. J. Royal Stat. Soc., Series B (Methodological), **60**(1), 34–40 (1998)
13. Rabe–Hesketh, S., Skrondal, A.: Multilevel modelling of complex survey data. J. Royal Stat. Soc., Series A (Statistics in Society), **169**(4), 805–827 (2006)
14. Rao, J. N. K.: Small Area Estimation. Wiley, New York (2003)
15. Rondon, L. M., Vanegas, L. H., Ferraz, C.: Finite population estimation under generalized linear model assistance. Comput. Stat. Data Anal., **56**(3), 680–697 (2012)
16. Wedderburn, R. W. M. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika, **63**(1), 27–32 (1976)
17. You, Y., Rao, J. N. K.: A pseudoempirical best linear unbiased prediction approach to small area estimation using survey weights. Can. J. Stat., **30**(3), 431–439 (2002)

Robust and model-assisted small area estimation methods: an application to the Banca d'Italia Survey of Industrial and Service Firms

Stimatori robusti e assistiti dal modello nelle piccole aree: un'applicazione ai dati dell'indagine Banca d'Italia Survey sulle Imprese industriali e dei servizi

Bottone Marco, Casciano Maria Cristina, Fabrizi Enrico, Filiberti Salvatore, Neri Andrea and Salvati Nicola

Abstract The Banca d'Italia survey of Industrial and Service Firms gathers information many relevant economic variables relating to Italian industrial and service firms with 20 or more employees. In paper we study small area estimation of parameters describing sub-populations of this survey. We consider small area models based on specifying a regression model for the relationship between the target and the auxiliary variable at the unit (firm) level. As normality is likely to fail because of skewness and the presence of outliers we focus on robust methods and specifically on M-quantile regression methods as they allow for the estimation of pseudo-area effects without recourse to any distributional assumptions. Our approch is model assisted in the sense that although motivated by regression models are evaluated with respect to design-based properties.

Abstract *L'Indagine della Banca d'Italia sulle imprese industriali e dei servizi raccoglie informazioni su un gran numero di rilevanti variabili relative alla vita delle imprese con almeno 20 addetti. In questo lavoro studiamo la stima per piccole aree di parametri descrittivi di sotto-popolazioni di questa indagine. Specificatamente, consideriamo modelli di regressione lineare per piccole aree specificati a livello di*

Bottone Marco
Banca d'Italia, Via Nazionale 91, 00184 Roma e-mail: Marco.Bottone@bancaditalia.it

Casciano Maria Cristina
Italian National Institute of Statistics, Via Cesare Balbo 16, 00184 Roma e-mail: casciano@istat.it

Enrico Fabrizi
DISES, Universit Cattolica del S. Cuore, Via Emilia Parmense 84, Piacenza e-mail: enrico.fabrizi@unicatt.it

Filiberti Salvatore
Italian National Institute of Statistics, Via Cesare Balbo 16, 00184 Roma e-mail: filiberti@istat.it

Neri Andrea
Banca d'Italia, Via Nazionale 91, 00184 Roma e-mail: Andrea.Neri@bancaditalia.it

Salvati Nicola
Universit di Pisa, Via Ridolfi 10, 56124 Pisa e-mail: nicola.salvati@unipi.it

unit. Poich la normalit per i residui irrealistica a causa di asimmetria e outliers ci focalizziamo su metodi robusti e in particolare sulla regressione M-quantilica. La nostra impostazione disegno-assistita: ancorch motivati da un modello, le propriet degli stimatori vengono valutate rispetto al disegno.

Key words: quantile regression, GREG, design consistency, sub-regional estimation, bias correction

1 Introduction

The accurate estimation of aggregates such as total or average turnover, investments, value added for subsets (domains) of the population of small and medium enterprises is very relevant to to the economic debate and policies.

Estimates obtained from the Banca d'Italia survey of Industrial and Service Firms are of special relevance because of the Bank institutional role. Estimates produced in the framework of this survey are reliable for the nation as a whole or large domains. The Banca does not publish or use in outlook analyses estimates for sub-population smaller than those for which reliable design-based estimates are available. In the survey data analysis when when estimates are needed also for small domains small methods can be helpful. This research is part of a project aimed at exploring a suitable small area methodology for the survey of Industrial and Service Firms.

The key idea behind small area estimation is that of complementing sample data with auxiliary information obtained from external sources. This can be done in several ways: we assume that auxiliary information is in the form of known totals or means for variables that are recorded also in the sample and that can be used to predict the variables of interest provided that a regression model express them as function of the auxiliary.

For a review of recent applications of small area estimation methods to business data see [5].

In this paper we are interested in a model-assisted approach, which means that, altough motivated by a regression model, the small area estimators make use of survey weights and are evaluated with respect to design properties. Specifically, we want the estimators to be design-consistent, a form of guarantee against possible model failures when the sample size is from moderate to large. Various small area estimators with good design properties have been proposed in the literature. The interested reader can find a review in [4].

A well known estimator in this class is the adaptation of the GREG to small area estimation (sae-GREG) reviewed in [7], (section 2.5). This estimator is design-unbiased but suffers from two limitations: it is very efficient and it is not robust to outliers that are quite often met in the analysis of business data.

A design-consistent estimator relying on the theory M-quantile regression [2] was proposed by [4]. This estimator is robust-projective in the sense of [3], is characterized by a small bias as the sae-GREG, but somewhat more efficient as it accounts for *area heterogeneity*. Its very small bias is nonetheless paid at the price of instability (relatively high mean square error) when large outliers are present. This problem has been studied by [3] who proposed robust-predictive estimators to improve efficiency of bias-corrected M-quantile estimators at the price of a moderate bias increase.

In this research we extend the [3] methodology, that was proposed in a model-based framework, to the model-assisted approach. This new methodology is applied to the analysis of a data set from the Banca d'Italia survey of Industrial and Service Firms. Although the data are real, the application is only for explorative purposes. We also present the results of a limited simulation exercise.

2 The Banca d'Italia survey of Industrial and Service firms

Banca d'Italia conducts the Survey of Industrial and service firms (INVIND, thereafter) since 1972 to collect yearly information on investments, gross sales, workforce, expectations and other economic variables relating to Italian industrial and service firms. From 2002 onwards, the sample is made up of about 4000 companies. The survey adopts a cut off stratified sample design. The sample excludes all units that have less than 20 employees. Strata are combinations of branch of activity, size class (in terms of number of employees) and region in which the firm's head office is located. Firms with more than 5,000 employees are self-representative units. Firms that have been interviewed in past waves and which are still in the target population are always eligible for a new interview. The survey was initially designed to produce accurate estimates at national level of total turnover and investments.

Recently, the Italian National Statistical Institute (Istat) has created a new statistical register for the annual production of profit-and-loss accounts of small and medium enterprises (SMEs). This register, called Frame SBS, makes use of firm-level administrative and fiscal data as primary sources of information to annually estimate the main variables of the economic accounts of Structural Business Statistics (SBS), while the traditional survey on SMEs is used as complementary source of information for estimating those variables which cannot be directly obtained from administrative sources. We use the Frame SBS register as a source of auxiliary information for our analysis.

3 Three model-assisted small area estimators

Suppose that, from a population U partitioned into m domains $U_1,\dots,U_i,\dots,U_m$ a sample s is drawn according to a possibly complex design; the design is characterized by first order inclusion probabilities π_{ij}, $(j = 1,\dots,n_{s_i}, i = 1,\dots,m)$ and sampling weights $w_{ij} = \pi_{ij}^{-1}$. Let's denote with $s_1,\dots,s_i,\dots,s_m$ the domain-specific samples. Although unplanned, we assume for simplicity that $|s_i| > 0$. Let y_{ij}, $\mathbf{x}_{ij}$

denote the sample values for the target and a p-vector of auxiliary variables for which we also assume that $\bar{\mathbf{x}}_i = N_i^{-1}\sum_{j\in U_i}\mathbf{x}_{it}^T$ are known from an external source. We assume that the target of our inference is $\bar{Y}_i = N_i^{-1}\sum_{j\in U_i} y_{ij}$.

Among the many model-assisted small area estimators proposed in the literature, we review those relevant to this paper. A first example in this class is the adaptation of the popular sae-GREG estimator discussed in [7] (section 2.5):

$$Y_i^{greg} = N_i^{-1}\sum_{j\in U_i}\mathbf{x}_{it}^T\hat{\beta}_w + N_i^{-1}\left\{\sum_{j\in s_i} w_{ij}(y_{ij}-\mathbf{x}_{ij}^T)\hat{\beta}_w\right\} \qquad (1)$$

where $\hat{\beta}_w = \left(\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\right)^{-1}\left(\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}y_{ij}^T\right)$. The first addend in (1) is a *synthetic* estimator that it is very likely to be have a small variance, but a potentially large bias as it does not consider any area heterogeneity beyond that explained by the auxiliary variables. Moreover the estimator β_w can be unduly influenced by outliers. The second addend corrects the ponential bias, but in doing so, it will produce an higher MSE. [7] notes that, for this reason the GREG is not very efficient and its variances is of order $O(n_{s_i}^{-1})$.

In a line of research stemming from [2], [4] propose an estimator based on M-quantile regression theory. An introduction to M-quantile regression and its application to small areas is beyond the scope of this short presentation. The interested reader is redirected to the last two quoted papers. We simply note how the recourse to M-estimation instead of plain quantile regression adds a degree of freedom in selecting the level of robustness we want for our procedure: it can be less or more than ordinary quantile regresssion according to the choice of an influence function $\psi(.)$. The estimator of [4] is as follows:

$$Y_i^{WMQ} = N_i^{-1}\sum_{j\in U_i}\mathbf{x}_{it}^T\hat{\beta}_{w\bar{\theta}_i} + N_i^{-1}\left\{\sum_{j\in s_i} w_{ij}(y_{ij}-\mathbf{x}_{ij}^T)\hat{\beta}_{w\bar{\theta}_i}\right\} \qquad (2)$$

The difference between (2) and (1) is in the slope coefficient. $\hat{\beta}_{w\bar{\theta}_i}$ is a design-consistent quantile-specific slope estimator (see [4], theorem 1) referred to the $\bar{\theta}_i$ quantile. This quantile is chosen assuming that, if there is area heterogeneity in the regression residuals, units within the same area will lie on close quantile regression planes. Thereby, observation specific quantiles are calculated for all sample observations from the same area and then averaged. The first addend in (2) is a *robust projective* synthetic estimator that projects sample non-outliers in out-of-sample prediction, while the second addend is a bias correction similar to that of (1). The estimator (2) is more efficient than (1) in most situations as the choice of an area-specific quantile regression plane incoroporates the idea of *borrowing strength* across areas. Nonetheless, the bias correction, although very effective is liable to the effect of outliers and reduces efficiency.

[3] introduce a new quantile based estimator that is *robust projective* in the synthetic component and *robust predictive* in the bias correction. The idea is to robustify the bias correction in order to curb down the impact of big outliers according to some influence function $\phi(.)$ with suitable properties. The methodological puropose

of this paper is to extend this estimator, that was developed in the context of a model based framework to the model-assisted approach:

$$Y_i^{WMQ-bc} = N_i^{-1} \sum_{j \in U_i} \mathbf{x}_{it}^T \hat{\beta}_{w\bar{\theta}_i} + N_i^{-1} \left\{ \sum_{j \in s_i} w_{ij} \hat{\sigma}_i \phi \left(\frac{y_{ij} - \mathbf{x}_{ij}^T \hat{\beta}_{w\bar{\theta}_i}}{\hat{\sigma}_i} \right) \right\} \qquad (3)$$

In the second addend of (3), $\phi(.)$ is a bounded monotone non-decreasing function over the real line with $\phi(0) = 0$ and $\hat{\sigma}_i = median|y_{ij} - \mathbf{x}_{ij}^T \hat{\beta}_{w\bar{\theta}_i}|/0.6745$. With respect to the influcence function implicitly adopted in the estimation of $\hat{\beta}_{w\bar{\theta}_i}$, in line with [3] we impose $|\phi(.)| \geq |\psi(.)|$.

With the same assumptions and regularity conditions assumed by [4] it can be proved that (2) is a design-consistent estimator of $\bar{Y}_i$.

As for the variance estimation, in view of the nearly unbiasedeness of (1) and (2) the simple weighted residuals estimator of [8] can be used provided second order inclusion probabilities are available. [4] prove that it works very well. Nonetheless in this research we use the bootstrap procedure illustrated in [4]. Also for (3) [3] introduced an analytical estimator that can be extended to the model assisted framework. Nonetheless we consider the same boostrap procedure as before.

4 A simulation exercise

In this Section we report results from a simulation that is based on the the Australian Agricultural and Grazing Industries Survey (AAGIS) data as in [4]. A sample of 1652 Australian broad-acre farms spread across twenty-nine regions of Australia is studied. A population of $N = 81982$ farms is generated by bootstrapping the original AAGIS sample. The synthetic AAGIS population consists of 15 variables for 81982 farms. The size of regions in terms of farms ranges from 79 to 10930. The variable of interest is the Total cash costs (`TCC`; Y); its distribution exhibits strong positive skewness. Auxiliary information available ($\mathbf{X}$) for each farm includes the total revenues received by the business during the financial year (`TTR`) and the total area of the farm in hectares (`FarmArea`); we also consider a group of six binary variables created cross-classifying the farms by climatic zone and size (`SizeZone`). We use two combinations of auxiliary variables to obtain two models with different values of R^2 calculated upon fitting an ordinary linear regression model: 1) low linear relationship with $R^2 = 0.40$ ($\mathbf{X}_1 = $ (`SizeZone,FarmArea`), `low` scenario), 2) high linear relationship between Y and $\mathbf{X}_2 = $ (`SizeZone,TTR`), $R^2 = 0.90$ (`high` scenario). Note that this strong correlation between Y and $\mathbf{X}_2$ is not due to the presence of outliers, as the inclusion of `TTR` reduces their size.

Samples are selected according to a fixed size, unequal probability without replacement sampling design using the maximum entropy method [9] implemented in the `Sampling` package [10] in `R`. We consider a sample size of $n = 578$ corresponding approximately to a 0.7% sampling rate. Two alternative sets of inclusion probabilities are defined to be proportional to two different size variables Z: *i*) livestock (beef, sheep and wool - `livestock`) and *ii*) a Uniform variable on the interval $(1,20)$. More specifically, in case *i*), we define $\pi_j = 0.2 \times z_j + 0.05$, $\forall j \in U$ to avoid that too many inclusion probabilities are equal to one. In our setting the first scenario is labelled non-ignorable design while the second one is called ignorable design.

The estimators of the small area means that we compare are the proposed bias corrected weighted M-quantile estimator (3), the weighted M-quantile estimator (2), the 'direct' estimator, the unweighted M-quantile estimator (MQ) and the sae-GREG estimator (1). The 'direct' estimator of the mean is a Hájek type estimator, i.e. $\hat{\bar{Y}}_i = \sum_{j \in s_i} \breve{w}_{ij} y_{ij}$. Note that for the M-quantile model the ψ function is set to be the Huber Proposal 2 with $c_\psi = 1.345$. If the reader would like to compare the performance of the estimator in the simulation experiment with other predictors that use sampling weights, as the pseudo-EBLUP estimator [12] and the GREG-LV estimator [6], see the results in Table 1 in [4].

The Monte-Carlo experiment consists of drawing $R = 5000$ samples from this population and calculating small area estimates of the mean of `TCC`. The performance of the different small area estimators is evaluated using the relative bias (RB) and the relative root mean squared error (RRMSE) of estimates of the small area parameters. The averages across the areas of the RB and RRMSE are reported for the considered estimators and the two scenarios in Tables 1. The comparisons include also an estimator labelled as MQ, which a the unweighted M-quantile regression based small area estimator that ignores the design. It is defined in [11].

Focusing on the proposed estimator we observe that, as expected, it shows a small bias. The WMQ is nearly design unbiased. We note that average relative biases are higher for the unweighted estimator (MQ) or for estimators that include unweighted components such as GREG-LV and pseudo-EBLUP, and this is particularly true when the design is non-ignorable. In terms of RRMSE, the proposed bias corrected predictor WMQ-bc is the most efficient estimator.

Table 1 Design-based simulation results using the AAGIS data. Results show the mean Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE) averaged across areas under the two scenarios.

Predictors	Average RB			
	non-ignorable design		ignorable design	
	`low`	`high`	`low`	`high`
MQ	10.13	7.86	-0.34	-0.83
WMQ	-0.41	0.84	-0.58	-0.59
WMQ-bc	-2.88	-0.79	-4.28	-2.63
sae-GREG	-0.21	1.54	-1.23	-0.22
Direct	1.01	1.01	-0.91	-0.91
	Average RRMSE			
MQ	27.53	20.23	18.43	14.13
WMQ	26.79	15.09	21.72	15.94
WMQ-bc	21.84	14.03	19.65	13.71
sae-GREG	35.59	16.73	25.40	16.86
Direct	40.35	40.35	35.43	35.43

5 Application to Banca d'Italia's data and discussion

In this section we compare the Horwitz-Thompson and the three model assisted design based estimators of Section 3 in two exercise data set with data from the Banca d'Italia survey of Industrial and Service firms. In both datasets we define domains to be a cross-classification of firms below 5000 employees by geography (19 regions), 2 sectors (industry, services) and size (below and above 100 employees). The target parameter is given in both cases by the average turnover.

In the first dataset the covariate is given by the lagged turnover (whose population total becomes available from the Frame SBS archive with delay with respect to survey data collection and analysis). This is a situation where a very powerful linear predictor is available ($R^2 = 0.98$, although influenced by points with high leverage). This scenario is called `high`.

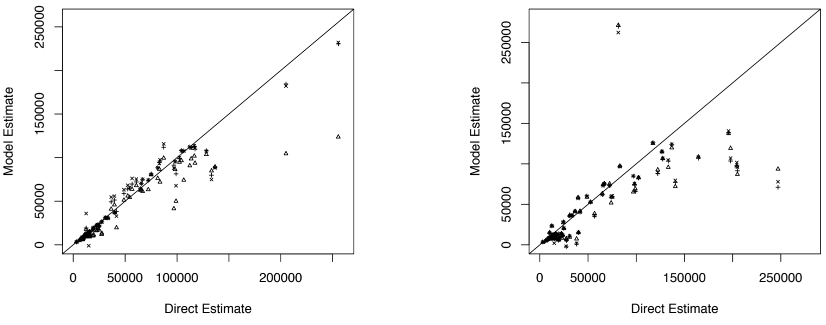
In the second dataset the covariate is given by the number of employees, a much weaker predictor of the turnover ($R^2 = 0.32$). This latter situation will allow us to explore what happens when, with a weaker predictivity, non-normality of the residuals' distribution and outliers are more likely. This scenario is called `low`.

In Table 2 we report the mean values across areas of the ratios of the estimated MSE of the WMQ, WMQ-bc and sae-GREG estimators to the estimated MSE of the direct estimator. The MSEs are estimated using the bootstrap procedure by [1]. In both scenarios the ordering of the MSEs is the same, with WMQ-bc being the most efficient. The scale of the efficiency improvement depends of course on the predictivity strength of the auxiliary information. The sae-GREG performs quite well relatively to the WMQ because in this application the intra-cluster correlation, i.e. area heterogeneity of the residuals is apparently small.

Table 2 Mean values across areas of the ratios of the estimated MSE of the weighted M-quantile (WMQ, WMQ-bc) and saeGREG estimators to the estimated MSE of the direct estimator.

Predictors	Scenario	
	`low`	`high`
WMQ	0.855	0.158
WMQ-bc	0.581	0.101
sae-GREG	0.918	0.182
Direct	1.000	1.000

Figure 1 compares the model-based estimates of the average turnover for each small area with the corresponding direct estimates. We note that the WMQ-bc estimates appear to be generally consistent with the direct estimates, with the correlation between the two sets of estimates being 0.92 in the low case and 0.95 in the high scenario. The corresponding correlation between the direct estimates and the WMQ estimates is 0.97 in low case and 0.94 in high case. Few remarkable differences can be seen on the right of both plots: they correspond to areas characterized by the presence of big outliers in the residuals. Curbing down these outliers can introduce some bias but protects estimates from instability that may translate in large year to year variations when these units enter or leave the sample.



(a) (b)

Fig. 1 Direct estimates versus corresponding WMQ (+), WMQ-bc (△) and sae-GREG (×) estimates in low (a) and high (b) scenarios.

References

1. Antal, E. and Tille, Y.: A direct bootstrap method for complex sampling designs from a finite population. Journal of the American Statistical Association **106**, 534-543 (2011)
2. Chambers, R., Tzavidis, N.: M-quantile models for small area estimation. Biometrika **93**, 255–268 (2006)
3. Chambers, R., Chandra, H., Salvati, N., Tzavidis. N.: Outlier robust small area estimation. J. R. Statist. Soc. B, **76**, 47–69 (2014)
4. Fabrizi, E. Salvati N., Pratesi M. Tzavidis, N.: Outlier robust model-assisted small area estimation. Biom. J. **169**, 157–175 (2014)
5. Fabrizi, E. Ferrante M.R., Trivisano C.: Bayesian small area estimation for skewed business survey variables. J. R. Statist. Soc. C (2017) doi: 10.1111/rssc.12254
6. Lehtonen R., Veijanen A. Domain estimation with logistic generalized regression and related estimators. IASS Satellite Conference on Small Area Estimation. Riga: Latvian Council of Science, 121–128 (1999)
7. Rao, J.N.K.: Small area estimation. Wiley, New York (2003)
8. Sárndal, C. E.: Implications of survey design for generalized regression estimation of linear functions. J. Stat. Plan. Inference **7**, 155–170 (1982)
9. Tillé Y.: Sampling algorithm. Springer Verlag, New York (2006)
10. Tillé Y., Matei A.: Package `Sampling` Functions for drawing and calibrating samples, downlodable at `http://cran.r-project.org/web/packages/sampling/`, (2009)
11. Tzavidis, N., Marchetti, S. and Chambers, R.: Robust estimation of small-area means and quantiles. Aust N Z J Stat, **52**, 167-186 (2010)
12. You Y., Rao J.N.K.: A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. Can J Stat, **30**, 431–439 (2002)

The Second Generation at School

Resilient students with migratory background

Studenti resilienti di origine immigrata

Anna Di Bartolomeo and Giuseppe Gabrielli

Abstract Numerous research has shown that there is direct relationship among high educational outcomes, employment, income and integration. Immigrants' children do show, on average, lower educational performances than natives' ones in addition of being more socio-economically disadvantaged. Countries' success in helping and integrating immigrants' children needs to find new educational policies and strategies. Using PISA survey data, we aim at providing further elements of discussion looking at "resilient" students. Given the strong impact of the students' social, cultural and economic family background, we aim at analysing how "disadvantaged" students can overcome their socio-economic and cultural obstacles and achieve good educational performances. Applying multilevel logit models, the paper analyses the factors linked to the probability of being resilient with a focus on parental and individual motivations and on school characteristics.

Abstract *Numerose sono le ricerche che hanno mostrato nel tempo la relazione diretta esistente tra rendimento scolastico, occupazione, reddito e integrazione. I figli di immigrati hanno, in media, un rendimento scolastico inferiore a quello dei figli di autoctoni e vivono in famiglie socio-economicamente svantaggiate. Sono necessarie nuove strategie e politiche in grado di raggiungere un miglior inserimento scolastico dei figli di immigrati. Utilizzando i dati dell'indagine PISA, lo scopo del presente lavoro è fornire ulteriori elementi di discussione analizzando gli studenti "resilienti". Nello specifico, l'intento è quello di definire come gli studenti, con un basso background socio-economico delle famiglie di origine, possano raggiungere risultati scolastici positivi. Applicando modelli logistici multilivello, sono analizzate i principali fattori legati all'essere studenti resilienti considerando nello specifico le motivazioni individuali e genitoriali e le caratteristiche scolastiche.*

Keywords: integration, second generation, school performances, resilience, multilevel

[1] Anna Di Bartolomeo, Università di Venezia Ca' Foscari, anna.dibartolomeo@unive.it
Giuseppe Gabrielli, Università di Napoli Federico II, giuseppe.gabrielli@unina.it

1 Introduction

Many studies in Europe have been devoted to the educational outcomes and social mobility pathways of immigrant descendants [among the others, 3, 7]. International research showed that positive educational paths and successful integration of immigrants' descendants promote social cohesion and development. There is direct relationship between high educational outcomes, employment, income and health.

However, immigrants' children have, on average, lower educational performances than natives' children in addition of being more socio-economically disadvantaged [8, 13]. When compared to their native peers, children of immigrants exhibit important signs that they are experiencing difficulties when integrating into the school system. These signs include more frequent school dropouts, less effective performances, and higher concentration of children of immigrants in vocational high schools where students enter the labour market immediately [3]. Scholars have tried to explain the worst immigrants' educational performances [among the others, 6, 11]. However, after controlling for variables at the individual level and the school level, differences between migrants and natives' children are still significant in different European countries [10, 11].

Countries' success in helping and integrating immigrants' children needs to find new educational policies and strategies to remove obstacles to student achievement and to help immigrant students fully develop their potential.

On average across OECD countries, students with migratory background are socio-economically disadvantaged compared to native students. Differences in socio-economic status explain over one-fifth of the gap between students with and without migratory background in the likelihood of obtaining academic skills [14]. It is debatable how disadvantaged students can overcome these obstacles and achieve good performance. Although most applied work identifies socio-economic disadvantage as a risk factor for poor academic performance [among the others, 15, 16], some disadvantaged (resilient) students beat the odds against them and achieve good academic.

Using PISA survey data, we aim at providing further elements of discussion looking at "resilient" students. Applying multilevel logistic models, we analyse the main determinants to be resilient students in EU countries considering both individual motivations and school resources by sex and migratory background. Resilient students are compared to other disadvantaged low achievers in order to individuate the main factors (such as environment, individual ability or motivation ability) to achieve positive school performances among "disadvantaged" students with immigrant background.

Which are the determinants associated with the probability of being resilient? What role is played by individual and parents' aspirations in having positive educational performance? Do schools' characteristics, activities and resources policies make a difference? To what extent do such patterns differ between children of natives and children of immigrants?

2 Data and methods

Data come from the Programme for International Students Assessment (PISA) dataset carried out in 2015 and developed by the Organisation for Economic Co-operation and Development (OECD). The Italian survey involves 11,583 students aged 15+ (15.8% with migratory background) attending 474 schools.

The PISA survey is conducted every three years with the aim of assessing the educational achievement of fifteen-year-old students in the most advanced economies. These cognitive skills are measured through standardized tests which aim not only at testing pupils' theoretical competencies but also – and especially – at assessing whether pupils are ready to enter their adulthood. These tests are later evaluated and synthesised into a continuous scale by a pool of international experts through the *Item Response Theory* (IRT), which gives five plausible values of pupils' performance for each student [12]. For the sake of country comparability, these scores are then standardized to an international mean of 500 and a standard deviation of 100.

Apart from test scores, data from students, parents and school characteristics were collected in separate questionnaires. A major limitation of PISA surveys relies, instead, on the fact that in Italy there are no detailed info on students' country of origin while it is well known as origin is likely to strongly affect immigrants' children trajectories in several fields.

Methodologically, the paper is based on multilevel regression analyses. This methodology is one of the most suitable methods for interpreting this interaction between levels [9] and frequently used in the analysis of student performances and behaviour patterns as it allows to understand the different levels in which students are grouped: classes, schools and educational systems [among the others, 7, 10].

In order to identify the – individual and school – determinants associated to the probability of being a resilient, a two-level logistic random intercept model is adopted. Specifically, we run the following model:

$$logit\{P[y_{ij} = 1|x_{ij},\zeta_j]\} = \beta_0 + \beta_1' x_{1ij} + \beta_2' x_{2j} + \zeta_{j,}$$

where y_{ij} is the response for student *i* in school *j* (y_{ij}=1 if i-th student in j-th school is resilient, y_{ij}=0 otherwise); x_{1ij} are the individual level covariates for student *i* in school *j*; x_{2j} are the school level covariates for school *j*.

Two separate models are run for children of immigrants and children of natives, respectively, for assessing – if any – immigrants' children specificities. Both models do only include students from a low socio-economic background (disadvantaged students), i.e. those with an ESCS index below the 40th percentile of whole distribution. The dependant variable equals 1 when students are resilient, i.e. when achieving performance higher or equal the 60th percentile and 0 when students are not resilient, i.e. do achieve performance below the 60th percentile. The performance subject of interest is science.

Independent variables at an individual level are related to socio-demographic aspects including sex and migrant generation (in the model of immigrants' children) and to the motivation dimension including students and parents' educational aspirations. At a school level, the following dimensions were considered: a) size of the school location; b) school resources and extra-curricular activities; c) school management; d) homework support. Annex Table 1 details the construction of variables.

3 Results

Overall, the share of resilient students is higher among children of immigrants (30.6%) than among children of natives (26.9%).

The first fundamental step when launching a multilevel model is verifying whether this approach is an added-value to the analysis, i.e. making sure that the uncorrelatedness assumption in the data is violated. By the null model, it is possible to decompose variance and to appreciate the quotes of variance to be attributed to each level of analysis, as measured by the Intraclass coefficient (ICC). The higher the ICC, the higher the importance of adopting multilevel tools as most of the differences observed across subjects on the response are actually derived from group differences. In our case, the quote of variance to be attributed to the school context is very high for both populations equalling 31% for children of natives and 42% for children of immigrants. This means that belonging to one school or another produces a strong selection and that multilevel techniques are extremely useful in the analysis (especially for immigrants' children).

Table 1 reports the results of the analysis including covariates at a student and school level. At an individual level, gender and generation' effects are in line with the literature. Being male denotes a higher probability of being resilient (for both groups) while being children of mixed couples is an advantage with respect to being second-generation and (even more) to first-generation students.

As far as parental aspirations are concerned, previous studies have shown as they have a positive effect on students' performance [2]. Our study adds that they are also positively associated with the probability of being resilient and thus help overcoming socio-economic gaps. This effect is strong and significant for both groups: the odds ratios are around 2. This clearly shows that even if parents are unable to assist their children with specific subjects or skill they can still play a vital role by encouraging students' feeling of competence and positive scholastic attitudes. Conversely, it is worth noticing that individual aspirations do play a significant role only for children of natives. In the case of children of migrants, once controlled for parental aspirations, the positive effect of students' motivations disappears. This may suggest that parental role tend to assume a pivotal role in affecting students' trajectories in this group by even prevailing over individual motivations.

Table 1: Determinants of resilience among Italian students and students with migratory background. Odd-ratios and p-values of two-level logistic regression analyses.

Independent variables	Italian students		Students with migratory background	
	OR	Sig.	OR	Sig.
Individual characteristics (level 1)				
Generation				
Transnational parents (rif.)				
Second generation			0.3	**
First generation			0.2	***
Sex				
Female (rif.)				
Male	1.7	***	2.4	**
PISA Index of students' motivations	1.3	**	1.4	n.s.
Parents' expectations about children scientific career				
"Negative" expectations (rif.)				
"Positive" expectations	2.0	***	2.1	**
School (level 2)				
School location size				
Small: less than 15,000 inhab. (rif.)				
Medium: from 15,000 to 100,000 inhab.	1.4	n.s.	1.2	n.s.
Large: more than 100,000 inhab.	1.3	n.s.	2.3	*
Extra-curricular non-science activities				
Less than two activities (rif.)				
Two or more activities	2.3	***	2.1	**
Extra-curricular science activities and resources				
Less than two activities (rif.)				
Two or more activities	2.4	**	1.9	*
School resources				
Less than two resources (rif.)				
Two or more resources	1.6	*	0.9	n.s.
Presence of school-entry selection criteria				
Less than two criteria (rif.)				
Two or more criteria	1.3	n.s.	2.4	**
PISA Index of school principal leadership	0.7	**	0.6	**
Support to homework				
No support (rif.)				
One or more supports	1.9	**	2.3	**
Constant	1.7	***	0.1	***
ICC (null model)	0.42		0.31	

Note: n.s. p>0.1; * p<0.1; ** p<0.05; ***p<0.01

School variables do clearly show that scholastic resources and characteristics may play the difference for developing virtuous paths. In other words, in line with other studies [1], we found that all students may profit from attending schools which develop certain kind of promoting activities and which may count on a number of resources. In particular, interestingly enough, students benefit from being inserted in schools where extra-curricular activities are promoted, regardless of whether they are related or not to science. In both groups, indeed, the probability of being resilient is positively and significantly associated with belonging to schools that develop two or more of these activities: the odd ratios of extra-curricular science activities equal 1.9 for children of migrants and 2.4 for children of natives. The same values for extra-curricular non-science activities stand at 2.1 and 2.3. Along the same line, a higher support to children to homework also provide strong benefits for being resilient. This suggest that supporting these activities – regardless of whether they are curricular oriented or not – may help students in being less reliant on their (scarce) home resources. School support thus enables to fill the gap of belonging to familiar disadvantaged conditions by enhancing positive educational trajectories.

Children outcomes are also similar with respect to the effect of selection school mechanisms: higher degree of entry-selection criteria is positively associated with the probability of being resilient especially in the case of children of migrants (odd ratio=2.4 vs. 1.3 of children of natives). The contrary occurs instead with more "authoritative" schools: the higher the principal leadership Index, the lower the quotes of resilient students.

Interesting enough are the results concerning school resources and location. As to the former, school resources have a positive effect on children of natives' resilience abilities while no effect is found for children of migrants. With respect to the latter, the only significant (positive) effect is found for children of migrants attending schools located in large cities (with more than 100thousand inhabitants).

4 Discussion

The impact of socio-economic disadvantages on educational trajectories is one of the most alarming form of inequality. Poor scholastic performance and the consequent difficulties in integrating into the labour market would indeed ensure that the unequal distribution of power resources between classes is transmitted through generations.

This is particularly worrying for immigrant families. Most migrants have low skills and lack established family businesses, accumulated wealth and long-standing local social networks. For them education represents a unique opportunity for social mobility with respect to the next generation. Success in the education system would allow their children to obtain higher-paying, higher-status jobs with a contemporaneous rise in the family's social standing [5]. In spite of this, some disadvantaged students succeed in overcoming such obstacles and achieve good performance results, i.e. the so-called resilient students.

Our main result supports the idea that, among both children of migrants and children of natives, parental motivations are a strong predictor of "resilience". Not surprisingly, parental expectations on children future careers represent a support for exiting their initial background disadvantage and well perform at school. As far as personal motivations are concerned, they result significantly associated with the probability of being resilient only for children of natives. At a school level, there seem to exist several strategies through which students can be supported in overcoming their structural socio-economic gap and develop virtuous mechanisms. Among them, extracurricular activities focusing on science topics and homework support tools seem are worth noticing. However, not only school-oriented but also extra-curricular activities do show a positive impact for both groups. The time spent in activities such as being involved in developing school yearbooks, newspapers or magazines, in participating to school musicals, art or science clubs or even sporting teams is useful in helping students to achieve good performance. This is possibly linked to the fact that taking part to these initiatives would stimulate students' cultural profiles by reducing the time they spent in less motivating familiar contexts. Interestingly enough, school material and cultural resources do represent an input only for children of natives. Conversely, children of migrants seemed to be advantaged when attending schools located in large cities. Possibly, in such big contexts, immigrants' children tend to be less stigmatized by both teachers and schoolmates by so enhancing their self-esteem and performance.

To conclude, this analysis show that motivations and schools can make the difference in overcoming social mobility inequalities within the Italian context. Accordingly, the necessity of investing on school autonomy tools for implementing more and diverse activities directed at targeted – more disadvantaged – groups seem favourable. On the other hand, it should be also guaranteed that school policies are not the privilege of few and selective schools, but to the whole scholastic population. School autonomy is positive but need to be monitored, as also suggested from the result concerning school leadership.

References

1. Agasisti T., Longobardi S. E.: Inequality in education: Can Italian disadvantaged students close the gap? Journal of Behavioral and Experimental Economics. 52, 8-20 (2014)
2. Alexander K.L., Entwisle D.R., Olson L.S.: Lasting Consequences of the Summer Learning Gap. American Sociological Review. 72, 167-180 (2007)
3. Buonomo, A., Strozza, S., Gabrielli, G.: Immigrant youths: Between early leaving and continue their studies. In: Merrill, B., Padilla Carmona, M.T., Monteagudo J.G. (eds.), Higher Education, Employability and Transitions to the Labour Market, pp. 131-147, EMPLOY Project & University of Seville (2018)
4. Chiswick, B.R, DebBurrnan, N.: Educational attainment: Analysis by immigrant generation. Econ. of Educ. Rev. 23(4), 361-379 (2004)
5. Di Bartolomeo A. Explaining the Gap in Educational Achievement between Second-Generation Immigrants and Natives: The Italian Case. Journal of Modern Italian Studies. 16(4), 437–449 (2011)
6. European Commission: Education and Training Monitor 2017. Publications Office of the European Union, Luxembourg (2017)
7. Fossati, F.: The Effect of Integration on Immigrants' School Performance: A Multilevel Estimate, Center for Comparative and International Studies. WP n. 57, University of Zurich (2010)
8. Heath, A., Rothon, C., Kilpi, E.: The Second Generation in Western Europe: Education, Unemployment, and Occupational Attainment. Annu. Rev. of Sociology. 34, 211-35 (2008)
9. Hox, J.J.; Applied Multilevel Analysis. TT– Publikaties, Amsterdam (1995)
10. Levels, M., Dronkers, J.: Educational performance of native and immigrant children from various countries of origin. Ethn. and Racial Stud. 31(8), 1404 (2008)
11. Mussino, E., Strozza, S.: The Delayed School Progress of the Children of Immigrants in Lower-Secondary Education in Italy. J. of Ethn. and Migr. Stud. 38(1), 41-57 (2012)
12. OECD: Learning beyond Fifteen: Ten Years after PISA. OECD Publishing, Paris (2012)
13. OECD: PISA 2015 Results - Volume I. Excellence and Equity in Education. OECD Publishing, Paris (2016)
14. OECD: The Resilience of Students with an Immigrant Background. Factors that shape well-being. OECD Publishing, Paris (2018)
15. Sirin, S.R.: Socioeconomic status and academic achievement: A meta-analytic review of research. Rev. of Educ. Res. 75(3), 417-453 (2005)
16. White, K.R.: The relation between socioeconomic status and academic achievement. Psychological Bull. 91(3), 461-481 (1982)

Annex Table 1: Description of variables.

Variable	Description

Migratory generation	It includes the following items: a) Students with one Italian parent; b) Students borne in Italy or arrived before secondary school; c) Students arrived during secondary school or later.
PISA Index of students' motivations	Students’ statements about themselves on a four-point Likert scale to the four following items: a) I want top grades in most or all of my courses; b) I want to be able to select from among the best opportunities available when I graduate; c) I want to be the best, whatever I do; d) I see myself as an ambitious person.
Parents' expectations about children scientific career	Do you expect your child will go into a science-related career? No, Yes.
Extra-curricular non-science activities	It includes the following activities: a) Band, orchestra or choir; b) School play or school musical; c) School yearbook, newspaper or magazine; d) Volunteering or service activities; e) Chess club; f) Club with a focus on computers/ Information and Communication Technology; g) Art club or art activities; h) Sporting team or sporting activities.
Extra-curricular science activities	It includes the following activities: a) Science club; b) Science competitions; c) A big share of extra-funding goes into improvement of our school science teaching; d) School science teachers are among our best educated staff members; e) Compared to similar schools, we have a well-equipped laboratory; f) The material for hands-on activities in school science is in good shape; g) We have enough laboratory material that all courses can regularly use it; h) We have extra laboratory staff that helps support school science teaching; i) Our school spends extra money on up-to-date school science equipment.
School resources	Students’ statements about themselves on a four-point Likert scale with respect to the following question: Is your school’s capacity to provide instruction hindered by any of the following issues? a) A lack of educational material (e.g. textbooks, IT equipment, library or laboratory material); b) Inadequate or poor quality educational material (e.g. textbooks, IT equipment, library or laboratory material); c) A lack of physical infrastructure (e.g. building, grounds, heating/cooling, lighting and acoustic systems); d) Inadequate or poor quality physical infrastructure (e.g. building, grounds, heating/cooling, lighting and acoustic systems).
Presence of school-entry selection criteria	It includes the following factors that are considered when students are admitted to the school: a) Student’s record of academic performance (including placement tests); b) Recommendation of feeder schools; c) Parents’ endorsement of the instructional or religious philosophy of the school; d) Whether the student requires or is interested in a special programme; e) Preference given to family members of current or former students; f) Residence in a particular area; g) Others.

Residential Proximity to Attended Schools among Immigrant-Origin Youths in Bologna

Vicinanza residenza-scuola fra i ragazzi di origine immigrata a Bologna

Federica Santangelo, Debora Mantovani and Giancarlo Gasperoni

Abstract This paper explores home-school proximity among students attending lower secondary schools and identifies student and school characteristics associated with proximity differences. After having implemented a geolocation procedure to a data-base supplied by the Italian National Institute for the Evaluation of the Educational System for Schooling and Training (INVALSI), distances between students' homes, their schools and other schools are examined for evidence of differences between native- and immigrant-origin students.

Abstract *Questo paper esplora la vicinanza casa-scuola tra gli studenti che frequentano le scuole secondarie di primo grado e individua le caratteristiche degli studenti e delle scuole associate alle differenze di distanza. Dopo aver implementato una procedura di geolocalizzazione su una base di dati fornita dall'Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione (INVALSI), viene esaminata la distanza tra la residenza dell'alunno, la scuola ad essa più vicina e la scuola effettivamente frequentata, per rilevare differenze tra studenti di origine nativa e immigrata.*

Keywords: Italy, home-school proximity, immigrant-origin students, socio-economic status

[1] Federica Santangelo, Dip. Scienze Politiche e Sociali, Univ. di Bologna: federica.santangelo@unibo.it;
Debora Mantovani, Dip. Scienze Politiche e Sociali, Univ. di Bologna: d.mantovani@unibo.it;
Giancarlo Gasperoni, Dip. Scienze Politiche e Sociali, Univ. di Bologna: giancarlo.gasperoni@unibo.it.

1 Home-school distance as an indicator of immigrant-origin student educational disadvantage?

This paper examines residential proximity to schools among students attending lower secondary schools in the city of Bologna, Italy, and in particular focuses on home-school distance differences between native- and immigrant-origin pupils. Several studies stress that school choice of students from lower social classes is mainly ascribable to area of residence and distance from school. More precisely, socially disadvantaged students are more likely to attend the nearest school to their home, since school selection depends more on convenience rather than on an evaluation of multiple schools' pros and cons in a medium/long-term perspective [4]. The decision may be shaped by several aspects, such as: the degree to which values, norms and beliefs are shared by parents and teachers; parents' level of understanding of the school system; economic, cultural and social resources. Since immigrant-origin students are overrepresented in working class families, we might expect that: i) they are more likely to attend the nearest school to their home; ii) immigrant-origin status is irrelevant after controlling for social origins.

Therefore, differential home-school distance is a potential indicator of school segregation mechanisms, possibly reflecting both residential separation among native and immigrant populations and more specific school-based mechanisms that reinforce (or weaken) the residential component of segregation. To this end, the analysis will distinguish between (aggregations of) neighbourhoods in the Bologna area. The main data, provided by the Italian National Institute for the Evaluation of the Educational System for Schooling and Training (INVALSI), involves a sample of about 2,000 students living and attending schools in Bologna in the 2014/15 school year. Other data is drawn from the 2011 Italian Population Census conducted by the Italian National Institute for Statistics (ISTAT).

Residence-school proximity has often been studied as a factor influencing adolescents' levels of physical activity and health [1,2], but here we reflect on its capacity to convey differences between native and immigrant-origin families' strategies as regards their children lower secondary school choices and the potential for ensuing school segregation, which has received little attention in Italy [5]. We remind readers that in Italy lower secondary education is mandatory and comprehensive in nature (vertical tracking begins with upper secondary education).

More specifically, this paper will address the following questions with a primarily exploratory approach. Do immigrant-origin children attend the closest school to the same degree as children of Italian nationality? Do they travel the same distance? Do differences vary according to area of residence within the city, and/or are they associated with other student, family or area characteristics?

The preliminary findings reported here are part of a research project titled "Social Exclusion and Selection in Lower Secondary Education: School Segregation Dynamics and Criteria Concerning Immigrant-Origin Students". The project is financed by the University of Bologna, through its Alma Idea grant programme supporting basic research. The overall aim of the project is to explore the segregation of native- and immigrant-origin students in lower secondary schools in larger cities in Central and Northern Italy, via an account of foreigners' presence in these territorial contexts, a description of how non-Italian students are distributed among schools and classes, the identification of mechanisms underlying such differential allocation, and the explanation of these mechanisms' effects on learning levels. This paper is a preliminary effort developed within this project.

2 Data and Variables

The following subsections describe the source of the data-base (and its limits) used for our analyses, as well as the dependent variables, the main explanatory variables and other covariates, and the basic data analysis strategies implemented.

2.1 Sources of data

The data primarily used in our analyses were provided by INVALSI and refer to students enrolled in the final year of lower secondary schools, in the 2014/15 school year, across 36 Italian provinces (chosen in such a way as to include major cities with a significant immigrant-origin population). The data included, most importantly, students' home addresses as well as the addresses of the schools they attended at the time, but also comprises other information, including students' native/immigrant-origin status (and, if applicable, their belonging to so-called first or second generations), gender, and other socio-demographic characteristics concerning the pupils themselves, their families, their schools and their classes. The current analysis focuses only on students living and attending schools in the city of Bologna. A major limit of the INVALSI data-base is an underrepresentation of immigrant-origin pupils as compared to the MIUR (Italian Ministry of Education) data-base, comprising the entire student population.

The data-base was submitted to an extensive cleaning and augmentation process, including the georeferentiation of home and school addresses, which required an initial reformatting of addresses (in order to adapt them to the requirements of the geolocation software) and the use of the QGIS platform and its MMQGIS geocoding plug-in. The data-base was also enriched with census-area information (2011 ISTAT census data) concerning selected features of the resident population, including its composition broken down by citizenship.

2.2 Dependent variables

Comparatively short distances from residence to school may reflect families' choosing to enrol their children in the most readily available learning institution and thus adhering to school catchment boundaries, whereas selecting a comparatively distant school may correspond to a more discerning behaviour based on information-gathering and evaluation strategies. It could be argued that the distinction roughly mirrors a passive versus active choice dynamic. On the other hand, greater distances also imply, other things being equal, greater travel times and efforts. The selection of a school distant from one's residence might also depend on other convenience criteria, such as ease of access due to a school's proximity to a parent's workplace or grandparents' home. Available data do not permit exploration of these aspects.

The georeferentiation of home and school addresses allows for the estimation of a (Euclidean) distance for each student's home-school dyad. Indeed, the association of each address to its geographical coordinates makes it possible to calculate the distance between a pupil's home and *any* school and thus to ascertain whether the school he/she actually attends is the closest to home. Also, it is possible to compare actual proximity with the greatest theoretical proximity (difference between actual home-school distance and the shortest possible distance from the pupil's home to a lower secondary school).

This information puts at our disposal a set of potential dependent variables:

- D1. Attendance of school closest to home: no = 0, yes = 1 (binary variable);
- D2. Actual home-school distance (cardinal variable, with no 0 values);
- D3. Supplementary home-school distance, i.e., the difference between actual home-school distance and the distance between home and nearest school (cardinal variable, equal to 0 when the nearest school and the actual school coincide).

The sample's students are enrolled in 28 schools (an additional 2 schools are located in the city of Bologna, but they are not attended by any students in the INVALSI data-set).

2.3 Explanatory variables and covariates

We focus on two main explanatory variables. The first is *immigrant/native status*. Our sample, once cases with missing values have been eliminated, comprises 1,917 units, of which 1,615 defined as native Italians, 150 members of the so-called second generation (G2) and 152 first-generation immigrants (G1). G1s are expected to display a weaker degree of acculturation in the destination country, which may lead to less informed school choices.

[2] Each dependent variable features two major limits. The first one is computational: home-school road distance may considerably differ from the Euclidian one, in terms of both kilometres travelled and travel times. The second is methodological: the closest school to home is not necessarily the school that the student "should" select considering his/her residence, since our analyses do not take into account formal catchment areas. Future efforts will address these issues.

The second variable is *area of residence*. Figure 1 highlights Bologna's administrative zones, in use from 1966 to 1985 and still widely employed in everyday experience. These 18 *quartieri* were subsequently grouped into 9 larger areas and, more recently, into just six, but we've used the older boundaries as the basis for a different 8-area aggregation, that takes into account social similarities as well as the number of immigrant-origin students in our data-base:

- Bologna West: Borgo Panigale, Barca, Santa Viola, Saffi;
- Historic Centre: Marconi, Irnerio, Malpighi, Galvani;
- Costa-Saragozza / Colli / Murri;
- San Ruffillo / Mazzini;
- San Vitale;
- San Donato;
- Corticella / Lame;
- Bolognina.

Analyses at the local level are crucial for examining possible school segregation effects. In fact, analyses conducted at the national or regional level might suppress forms of school segregation in play. Only a detailed examination of local areas' characteristics (socio-economic and ethnic features, distribution of schools) may reveal operative segregation [3,6].

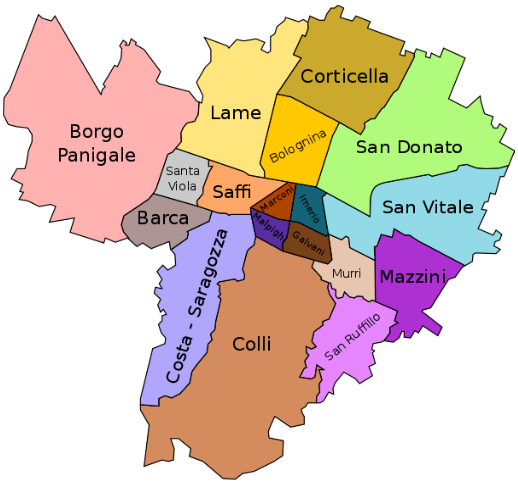


Figure 1: Bologna neighbourhoods

Table 1 displays some basic information about the 8 areas used in our analysis. The more socially disadvantaged areas are San Donato and Bolognina, as reflected by the relatively lower incidence of residents with a university degree, the higher unemployment rates and the higher concentration of foreign-origin residents. The INVALSI data-base shows that San Donato and Bolognina have the highest residential concentration of immigrant-origin students. Some differences exist between these two ex-*quartieri*: in San Donato, the incidence of first and second generations is relatively similar, even if G1s exceed G2s (respectively 16.5 and 12.4%), whereas in Bolognina G2s are more numerous than G1s.

The multivariate analyses will involve the following covariates: student gender and parents' highest education level.

Table 1: Some socio-economic characteristics of population residing in Bologna's areas (2011 Census data) and students' distribution in lower secondary schools per areas and migratory status (INVALSI data)

	Population (2011 Census)				INVALSI data-base				
	Pop. (000s)	% Non-Italians	Unempl. rate	% univ. graduates	No. pupils	No. schools	% Non-Italians	% G1	% G2
Bologna West	75,482	12.2	5.4	15.0	399	5	9.8	5.3	4.5
Historic Centre	50,720	10.0	5.6	39.3	187	6	11.2	4.3	6.9
Costa-Saragozza / Colli / Murri	59,786	9.1	4.8	32.0	345	4	7.3	3.8	3.5
S. Ruffillo / Mazzini	57,925	10.1	5.2	19.2	206	4	15.5	6.8	8.7
S. Vitale	32,574	11.7	5.6	22.8	137	1	16.8	8.8	8.0
S. Donato	30,557	14.2	7.6	14.7	194	2	28.9	16.5	12.4
Corticella / Lame	31,366	12.1	5.8	12.7	236	2	17.8	11.9	5.9
Bolognina	32,764	20.0	6.2	17.4	213	4	30.1	11.3	18.8
Total	371,174	11.9	5.6	22.4	1,917	28	15.7	7.9	7.8

3 Preliminary findings

As a whole, Italian and immigrant-origin students show a high propensity towards school mobility: over two students out of five attend the closest school to home (Table 2). Aggregate data suggest that immigrant-origin students are slightly more likely than natives to choose the school closest to home.

Nonetheless, analyses broken down by area reveal noteworthy differences: Italian students residing in Bologna West and the Historic Centre are more likely to be less mobile, whereas the same behaviour is more common among immigrant-origin students living in San Donato, one of the city's most socio-economically disadvantaged outskirts. A lower propensity towards mobility is also evident in both groups of students residing in Corticella / Lame.

Table 2 also highlights differences between G2s and G1s. In particular, G1s have a higher propensity to opt for the nearest school to home (D1). This result could be explained by the fact that G1s (and their families) have a lower knowledge of the Italian school system due to their more recent (as compared to G2s's families) arrival in Italy. However, G1s' likelihood of attending the school closest to home is evident only in some areas, such as San Donato and Corticella / Lame.

Table 2: Students attending the closest lower secondary school to home (D1), overall and supplementary home-school distances (in km) travelled by students (D2 and D3) by area and migratory status

	D1 (% closest school)				D2 (mean distance)				D3 (mean suppl. distance)[*]			
	Ital.	Imm.	G1	G2	Ital.	Imm.	G1	G2	Ital.	Imm.	G1	G2
Bologna West	58.6	46.1	42.9	50.0	1.39	1.78	1.86	1.69	1.36	2.06	2.23	1.84
Historic Centre	51.8	47.6	37.5	52.9	0.73	0.84	0.49	1.05	0.75	0.98	0.15	1.66
C.-Sar. / Colli / M.	25.3	24.0	23.1	25.0	1.20	1.35	1.44	1.24	0.78	1.02	1.12	0.92
S. Ruffillo / Mazz.	37.9	34.4	14.3	50.0	1.39	1.06	1.19	0.96	1.31	0.74	0.55	0.99
S. Vitale	28.1	34.8	33.3	33.4	1.71	1.51	1.39	1.64	1.21	1.12	1.00	1.26
S. Donato	42.0	60.7	78.1	37.5	1.54	1.21	1.01	1.48	1.44	1.42	1.55	1.35
Corticella / Lame	52.6	66.7	71.4	57.1	1.79	1.48	1.49	1.47	2.17	2.25	2.62	1.75
Bolognina	33.6	40.6	33.3	45.0	1.11	1.04	1.10	1.01	0.97	1.00	0.93	1.05
Total	42.5	46.7	48.7	44.7	1.34	1.28	1.28	1.27	1.19	1.28	1.28	1.28

[*] D3 calculated only for students not attending school closest to home.

As far as the home-school distance is concerned (D2), on average an Italian student travels 1.34 km, whereas an immigrant-origin student 1.28 km. Differences between G1s and G2s are negligible. Analysis by *quartiere* reveals some distinct features: immigrant-origin students, especially G1s, are more likely to travel longer distances than Italians if they reside in Bologna West. In this case, higher mobility (D1) is associated with greater travelled distance. On the contrary, although G1s residing in the Historic Centre show a similar propensity to school mobility (37.5%), they travel very short distances to reach their schools (0.49 km). This finding might reflect a process of school segregation: in the more socio-economically prestigious area of the Historic Centre, students are relatively mobile and may travel short distances to reach the school that fits their needs. As a result, Italian students might be concentrated in some schools and immigrant-origin students in other schools, which are very close to each other and located in the same area.

In order to explore the different propensity to mobility in selecting schools among Italian and immigrant-origin students, two multivariate analyses have been developed: a logistic regression on the probability to choose the closest school (D1) and a linear regression on the additional distance (D3).

The former suggests if – all other things being equal (student gender, family's level of education and area of residence) – migratory background is associated with selection of the closest school to home. The latter focuses on a cardinal dependent variable, with values determined as follows: the difference between the distance from home to the selected school and the distance from home to the closest school. The linear regression was estimated only on the subsample of students who do not attend the nearest school in order to investigate if specific characteristics are detectable among students travelling an additional distance to go to school.

[3] Analyses control for all interactions between parents' highest level of education, immigrant status and area of residence, but average marginal effects have been calculated as if several regressions had been

Preliminary results show that G2s are more similar to Italians than to G1s with respect to both the choice to attend a school on a distance basis and the additional distance that students travel in order to attend a school that does not coincide with the closest to home. Furthermore, G1s exhibit a lower probability than Italians to select a school far from home when they reside in San Donato. On the contrary, they are more inclined to travel farther than Italians when they live in Mazzini-San Ruffillo. As far as the additional distance is concerned, G1s tend not to traverse a longer distance than Italians.

Irrespective of immigrant status, two mechanisms appear to be in play: with the only exception of Bolognina inhabitants, families living in the most disadvantaged neighbourhoods are less likely to adopt school choice strategies other than convenience. Families living in wealthier neighbourhoods, like the Historic Centre, have the same behaviour; in this case, however, they live in areas which are the preferred destination for families who decide to place their children in non-near schools. Despite the fact that families residing in Corticella / Lame are comparatively less willing to enrol their children in distant schools, more educated parents are also the most willing to have their children travel longer distances.

To tentatively conclude, data does not reveal any operational process of school segregation on an ethnic basis, and school selection seems to be driven more by family's socio-cultural and economic traits. As a consequence, immigrant-origin students' (especially G1s) school choice patterns are more associated with their over-representation in working class families than with their migratory status.

References

1. Duncan, S. et alii: Active Transport, Physical Activity, and Distance Between Home and School in Children and Adolescents. J. Physical Activity & Health (2016): https://doi.org/10.1123/jpah.2015-0054
2. Easton, S., Ferrari, E.: Children's Travel to School—The Interaction of Individual, Neighbourhood and School Factors. Transport Policy (2015) https://doi.org/10.1016/j.tranpol.2015.05.023
3. Gibson, A., Asthana, S.: Local Markets and the Polarization of Public-Sector Schools in England and Wales. Transactions of the Institute of British Geographers (2000) doi: 10.1111/j.0020-2754.2000.00303
4. Oberti, M.: Social and School Differentiation in Urban Space: Inequalities and Local Configurations. Environment and Planning (2007) doi: 10.1068/a39159
5. Pacchi, C., Ranci, C. (eds): White flight a Milano. La segregazione sociale ed etnica nelle scuole dell'obbligo, Milan, Franco Angeli
6. Raveaud, M., van Zanten, A.: Choosing the Local School: Middle Class Parents' Values and Social and Ethnic Mix in London and Paris (2007) doi: 10.1080/026809360601065817

developed for each area of residence. Level of education, migrant status and the area of residence could possibly be correlated to the socio-economic status, causing an endogeneity problem.

From school to ... future: strategies, paths and perspectives of immigrant immediate descendants in Naples

Dalla scuola al …. futuro: strategie, percorsi e prospettive di giovani immigrati di seconda generazione a Napoli

Giustina Orientale Caputo and Giuseppe Gargiulo

Abstract: The study subject presented here is the school inclusion process of immigrant immediate descendants enrolled in some middle schools of Naples. The research's general objective was to study the immigrant immediate descendants integration and, in particular, to identify the determinants of the choice, once achieved the compulsory school title, to continue studies or enter the labour market by using a qualitative approach. In fact, it seemed interesting to study this transition phase, which is crucial for adolescents, because they are in constant evolution in terms of aspirations, constraints and opportunities. First results have shown how it is important to consider the simultaneous interconnections between several factors such as: the territorial context; the individual and family characteristics; the family migratory history; schools institution and educational policies on theme about intercultural education and pedagogy. We have seen how the interweaving between the weaknesses of each of the characteristics we mentioned above, could often define paths different from those imagined - by themself or by their families. These weaknesses are also cause of biographical shocks in the delicate transition from school to the future due to not understood, not supported, disincentive and obstructed aspirations.

Abstract*: Oggetto dello studio di cui qui presentiamo i primi risultati è il processo di inclusione scolastica dei giovani immigrati iscritti in alcune scuole secondarie di primo grado a Napoli. Obiettivo generale della ricerca era studiare l'integrazione degli immigrati di seconde generazioni ed in particolare, attraverso un approccio qualitativo, cogliere le determinati relative alla scelta, successiva al conseguimento*

[1] Giustina Orientale Caputo, Università degli Studi di Napoli Federico II; email: oriental@unina.it

Giuseppe Gargiulo, Università degli Studi di Napoli Federico II; mail: giuseppe.gargiulo2unina.it

del titolo di studio dell'obbligo, di proseguire gli studi o di entrare nel mercato del lavoro. Ci è sembrato infatti interessante studiare una fase di passaggio che in generale per tutti gli adolescenti è cruciale poiché pone i ragazzi in un continuo confronto tra desideri, vincoli e opportunità disponibili. Dai primi risultati emerge come sia stato significativo considerare le simultanee interconnessioni tra più fattori quali: il contesto territoriale; le caratteristiche individuali e familiari; la storia migratoria del nucleo familiare; l'istituzione scuola e le politiche educative in tema di educazione e pedagogia interculturale. Nello specifico, si è visto come siano spesso proprio i punti di debolezza di ciascuno di questi elementi che nel loro intrecciarsi in modi diversi possono definire percorsi diversi da quelli immaginati – dal soggetto o dalle loro famiglie – e anche provocare shock biografici nel delicato passaggio dalla scuola al futuro dovuti ad aspirazioni non comprese, non sostenute, disincentivate o, al limite, ostacolate.

Key words: immigrant immediate descendants, educational insertion, secondary school, Naples.

1 Introduction

Over time Migrant phenomenon has taken different structures and traits, shaped by economic, social and political variables typical of the different phases(stages) of European and World History. Comply with Ambrosini's Periodization (1999), the transition from temporary migration and for job reasons to long-term settlement proposal, changes the own nature of the migrant phenomenon. Subsequently is the permanent and growing presence of the migrant families, reunited or newly created, and children in the society of residence (Ambrosini 2005) with a deep change in the human and social territorial geography and social spaces involved (Basteneir, Dasetto 1990). The steady presence, increase in number the relations and the new subjects demand to become part of the social context surrounded (bringing out the demand of the new subjects in the territory to become part of the social context surrounded); and also stimulate the institutions and civil society to provide responses to the migrant demands. Parents and children are faced with an unknown and complicated situation, going through different ways (Bindi 2005).

Migration involves the questioning of minors born in Italy from foreign parents, in that case it shows a dis-identification from the parents' s culture and an identity transfer more closed (projected) to the birth country. Children, whether they are young with family reunited or born in Italy, are in a different phase compared to their parents, in which the socialization process and identity definition is still flowing (or has just begun) and it happens in a different system of expectations shaped from another culture that either does not be part (belong to) of them or is not very immediate, but in which every day they do experiences.

The local school system is an essential socialization context with the surrounding world and, at the same time, a unique opportunity to glean the necessary training for social inclusion (Strozza, de Filippo, Buonomo 2014). The educational insertion of migrant descendants is the primary aim of the modern, multinational and multicultural European societies because the accumulation of human resource is a basically requirement for the social inclusion. This has inflamed the debate of the Italian schools context on topics as: multiculturalism, organization in the school, educational content, pedagogy and teaching methods (Strozza 2015). The multicultural transformation of Italian school groups imposes to the school system a lot of civic and educational dares if it would like to keep his aims of equity, inclusion and democratic citizen education. The research we are going to discuss in this paper, intends implementing an perspective shown relationships between the practice education, often taken for granted - or not problematized - in the political-academic and professional debate and the assumption on which such practices are founded (Serpieri, Grimaldi 2014). Our cognitive object, in the first time, will be to analyze the integration / inclusion level of the second-generation migrant children in the Italian schools; the second question, it will be to understand this youngsters future orientations – continue to study or have a job - and it also concerns educational practice. Therefore, one of our aims it will be to bring out the existence of an equity problem in the educational system; linked to inequalities of opportunity and result that systematically disadvantage not-Italian pupils; disadvantage exacerbating in the higher school levels. In fact, considering the evolutions of the school system, a marked shift of school selection emerges from the secondary school of first degree towards the first years of secondary school so that for this order of school we can speak of deferred selection (Besozzi, Colombo 2014). For that reason our attention was shaped to the transition from secondary school to grade II.

The sociological approach through a qualitative methodology intends to achieve primarily the goal of reading the new needs they bring immigrant children and consequently those of the families where they belong, in the belief that "Children of immigrants (...) could (or should) be the main actors in the construction of the future society "(Strozza, Serpieri, de Filippo, Grimaldi (eds. 2014, p. 34). And if these companies must be "multicultural and low-low level of conflict"(Zincone 2000) then it must be built primarily from the integration of children in schooling and training. Just these paths must therefore carefully studied because they provide to young subjects - immigrants or not - on the one hand the tools for better construction and understanding of themselves and their own abilities and the other keys to access opportunities - that the context inside living can offer them - in terms of further training and / or integration into the labour market.

2 Methodology, technique and research phases

Our sociology-driven qualitative research perform the task of surveying, integrating and putting into context the statistical data and information collected during the first part of the research. Through field analysis we actually try to describe the conditions and experiences of minor immigrants and their families in migrants reception contexts, in order to identify the variables which have greater influence – whether negative or positive – on school, and later on social, integration of minor immigrants to prevent situations of marginalization, social exclusion and on their future choices. Immigrant immediate descendants represent, indeed, a growing trend both from the numerical point of view and from that of social importance The analysis take into account the different nationality groups to which the minors belong, their wider family background, the length of their stay in Italy, the links their families have with the labour market, so as to better understand minors' school progression and their successes and failures.

The project aim is to fill gaps in research on the subject of immigration in the Italian context through the providing understanding of the school inclusion process of children and youths of not Italian origins. The contribution which this part of the research provide looks rather remarkable to us, given the lack of surveys on the extent of the phenomena we are going to investigate, and of studies on the conditions, strategies, difficulties, changes and adaption to our schooling system foreign minors have to face while they are here. In other words, we intend analyse the determinant of the school drop-out and failures, the risk of perceived school discrimination, the school insertion, the training and employment expectations. On the other hand, this step is all the more necessary since the current cohort of second-generation is also the first, especially in Naples where they were defined as "frontrunners social category" (Spanò 2011). The analysis is also be oriented at understanding the present state of educational and social policies implemented by the individual schools for these foreign pupils, who can be defined somehow non standard, since they do not always possess the pre-requisites and base knowledge held by native minors. In our opinion, all that acquires further value in a context such as the city of Naples with a structurally weak educational offer, a common condition in Southern Italy (Landri, Milione, Vecchione 2014).

Even though data show a lower presence of minor immigrants in Naples schools compared to the other big Italian cities, we are witnessing here as well, ongoing changes in the presence of immigrants, and as a consequence we are facing rapid changes in the composition of the pupils mix. The research group is obviously aware that this rapidity does not always correspond with a rapidity in the transformation of the social policies applied locally and, most of all, with the institutions' capability of identifying rapidly the most effective actions to implement for the school inclusion of foreign minors. However, we are going to examine and illustrate these actions.

For a deeper understanding of these aspects, we have chosen the Neapolitan schools with a strong presence of immigrant minors, in the city centre, to be able to represent the weight of local contexts, at the analysis stage.

Therefore our object was the process of school inclusion of second-generation migrants attending the secondary schools of first degree in the historical center of Naples. Our aim was to understand the elements that determined processes of their scholastic and social inclusion and identifying if there are any problematic aspects of the process not already solved. The specific objectives of the qualitative research are the identification of the strong points and weak points of the school inclusion process and academic insertion. The aim will be to produce recommendation for any future interventions of integration policies developed.

The research question was: what are the determinants of the school inclusion process of the second-generation children attend the three-year period of secondary schools first degree? What are their training needs in the educational path? What are the variables that guide their choices of continuation or interruption their studies at the end of the course?

First of all, we are going to study the theoretical contribute of the statistical, pedagogical, sociological approaches and obviously the analysis of public policies. The data analysis from the available databases was preparatory for the field survey, made through qualitative research tools including: participant observation, interviews with privileged observers, autobiographical texts, focus group. About focus groups, actually we just made meeting separately children and their parents.

The survey has been performed as following:

- mapping secondary schools of first degree (level); consultation, contact and identification of those available to participate for the research;
- partake planning and/or interviews with school chief in order to identify specific detection areas, shared construction of survey tools and the identification of the class of each school.

Focus group made with:

- class group, with the aim to show the intergenerational questions in a peer to peer contest ; identify the role of peer relationships, the trans-generational relationship, the socialization contexts; observe the training needs and expectations of this age group;
- parents of second-generation migrant pupils with the aim to understand which was the impact and the influences of the family migration project on the children's educational career; representations, causal attributions in relation to the choice of scholastic paths and to the educational success; parental styles; models, beliefs and educational aims; requests and expectations with respect to training institutions.

At the first time, the phenomenon we want to understand was superior to the individual, therefore it was circumscribed to a group, even if not representative can be considered significant in order to understand our interested dynamics. The investigation method as through the focus group technique was the approach used in this phase and we are going to show this in the paper as first results. This investigation technique does not allow to reach the phenomenon quantification and not to their generalization but it is considered incisive for obtaining a lot of information on little-known or delicate subjects.

In the type of group interview, the stimulus is invariant and the reactions are recorded individually and in the same way. Moreover, in addition to the content of the reporting, the interactions between the members of the group are observed. The recent diffusion of this interview technique is contemporary with the rediscovery of the community, as a reaction to globalization and depersonalization and indicates a growing interest in implicit and explicit negotiation processes in the everyday sphere (Besozzi, Colombo, 2014) It is good to remember that, especially in socio-educational contexts, the interviewed subjects require particular caution. With children and adolescents some elements must be considered, such as: mastering a more or less restricted vocabulary; lower ability to use abstract concepts; asymmetrical relationship with the adult interviewer; limited time of attention to purely intellectual activities; lack of understanding of the situation; respond mechanically or in a complacent manner. For these reasons, until the nineties, children and young people were mostly excluded from large-scale investigations. However, in recent years, also in Italy, research has taken place with the children active protagonists in the consideration that minors are social subjects endowed with their own agency capacity and therefore able not only to report themselves with a language and a peculiar reflexivity but to affirm cultural needs and contents in the public sphere.

Specifically, the topics discussed during the meetings were:

- Family / cultural educational models
- School inclusion and discrimination
- Linguistic aspects
- Free time, relationships with the peer group, as represented in the speech of peers.

Field research started in September 2017.

Starting from city schools mapping, those institutes were contacted, due to the numerous and heterogeneity of enrolled student stalls of not-Italian origin, could better serve to understand the themes of the research. Not without difficulty, the schools that showed their availability were 6 and represent precisely those from the most complex and articulate student groups. Moreover, these institutions all insist on the vast territory of the historic center of Naples.

Received the availability of school leaders, the subsequent contacts were made with representatives of the institutes, with whom they shared the subsequent stages of identification of specific areas of research detection and identification of classes participating in the focus groups. At this point, the subsequent contacts were with the teacher coordinators of the classes identified in the six schools with which, through preliminary meetings, the knowledge of the composition of the group of students was deepened.

At present, meetings have been held with teachers and managers, with whom it is important to return to speak at the end of the meetings with pupils and parents in order to be able to make an overall assessment of the issues raised; all the focus has been completed with the boys and the first three focuses with the parents were conducted.

3 Preliminary research results

Although the research is still going, the material collected so far and the observations made may be the object of some consideration.

School, family, peer group and future, in fact, are the spaces - sometimes physical, sometimes even mental - with which the participants expressed their ability to "interpretive reproduction" (Corsaro, 2003).

The first possible observations are related to the social context in which the schools are located and the children live. The quarters of the Neapolitan historical centre in which the schools analysed by us fall are a common and varied space crossed by similar social phenomenon but quite peculiar to each neighbourhood. San Lorenzo, Vicaria, Zona Industriale and Montecalvario are neighbourhoods in which the presence of students of non-Italian origin, and therefore of foreign families, helps to make the typing of the Neapolitan historical centre. These areas are generally characterized by the presence of a particularly difficult social fabric, in which poverty, unemployment, the presence of multi-problem families in great difficulty appears high. This affects the growth of children and constitutes an element that defines a large part of their lives, as emerged from our survey. The urban and architectural structure of the neighbourhoods, the almost total absence of public green and equipped for children, the high population density, the school structures often inadequate and poorly equipped are just some of the problems that the area shows at first glance. And minors immediately reported their weight on their daily lives. Not less than their families did.

Present for several years in our country, mostly born in Italy, foreign children we heard all express themselves in Italian quite correctly, often speak very well the dialect and have a frequency of bilingualism in everyday use or otherwise express themselves and they understand both languages - the family (which often keep talking at home) and the Italian one. The most significant relationships that the boys have established in the classroom are in most cases with other boys of their same origin and in some cases even with Italians these relationships also continue outside the classroom and are an important part of their free time.

Refers to the aspects of daily interaction and the issue of discrimination, it can be said that the children felt cannot be touched by the problem: a problem of discrimination exists - where it is stronger where it is milder - and this strongly recalls the body's responsibility teacher and the choices that they often make in everyday life: diversifying the expectations among Italian and foreign students, also physically distancing the good guys from the less, not favouring and in some cases hindering the interactions between the groups. The success or failure of the study and the different disciplines is somehow predictive of the orientation towards the future that the boys show. And so the boys who have a better education (Italian or foreign) have a stronger orientation to continue their studies than others. In general all indicate that they want to continue their studies, even with the approval of their parents and no one has declared that they will interrupt them (the obligation of the other party will lead them to continue at least two more years after the conclusion of the cycle they currently attend), as well as in general foreign children more than the Italians put higher expectations in the study compared to professional achievement. In general, children of foreign origins have declared desires for higher professions and professional achievements of their Italian peers. If young Italians want to play all the players (the males) and the dancers or the hairdressers (the females), the young foreigners want to be a veterinarian, an anatomical pathologist, an engineer, a tour guide. These statements seem to indicate not only a different horizon between Italian parents (of classes strongly deprived as mentioned) and foreign parents, but also a different degree of confidence in the future and in the possibility of growth and improvement of the condition of departure of the children through the study.

References

1. Ambrosini, M.: Utili invasori: l'inserimento degli immigrati nel mercato del lavoro italiano. Franco Angeli, Milano (1999)
2. Ambrosini, M.: Sociologia delle migrazioni. il Mulino, Bologna (2005)
3. Bastenier, A., Dasetto, F.: Nodi conflittuali conseguenti all'insediamento definitivo delle popolazioni immigrate nei paesi europei. Fondazione Giovanni Agnelli, Torino (1990)
4. Besozzi, E., Colombo, M.: Metodologia della ricerca sociale nei contesti socio-educativi. Guerini Scientifica, Milano (2014)
5. Bindi, L.: Uscire dall'invisibilità. Bambini e adolescenti di origine straniera in Italia. UNICEF, Roma (2005)
6. Corsaro, W.: Le culture dei bambini. il Mulino, Bologna (2003)
7. Landri, P., Milione, A., Vecchione, A.G.: Allievi non standard? Strategie e tattiche di inclusione degli allievi con cittadinanza non italiana nelle scuole di Napoli. In: Strozza, S., Serpieri, R., de Filippo, E., Grimaldi, E. (eds.): Una scuola che include. Formazione, mediazione e networking. L'esperienza della scuola napoletana. Franco Angeli, Milano, 83--105 (2014)
8. Serpieri, R., Grimaldi, E.: Inclusione e scuola multiculturale. Pratiche e contesti a Napoli. Franco Angeli, Milano (2014)
9. Spanò, A. (eds.): Esistere, coesistere, resistere. Progetti di vita e processi di identificazione dei giovani di origine straniera a Napoli. Franco Angeli, Milano (2011)
10. Strozza, S.: L'inserimento scolastico dei figli degli immigrati: una questione aperta. Rivista delle politiche sociali, 2-3, 127--146 (2015)
11. Strozza, S., Serpieri, R., de Filippo, E., Grimaldi, E. (eds.): Una scuola che include. Formazione, mediazione e networking. L'esperienza della scuola napoletana. Franco Angeli, Milano, (2014)
12. Strozza, S., de Filippo E., Buonomo, A.: Immigrati, figli di immigrati e loro inserimento scolastico: Italia, Campania e Napoli. In: Strozza, S., Serpieri, R., de Filippo, E., Grimaldi, E. (eds.): Una scuola che include. Formazione, mediazione e networking. L'esperienza della scuola napoletana. Franco Angeli, Milano, 33--68 (2014)
13. Zincone, G.: Primo rapporto sull'integrazione degli immigrati in Italia. il Mulino, Bologna (2000)

Tourism Destinations, Household, Firms

The Pricing Behaviour of Firms in the On-line Accommodation Market: Evidence from a Metropolitan City

Andrea Guizzardi and Flavio Maria Emanuele Pons

Abstract The widespread diffusion of on-line travel agencies has opened the possibility, for hoteliers, to update continously quality and prices offered along the advance booking. We study firms' pricing behaviour in a business-oriented environment considering time series of daily best available rates for 107 hotels in Milan, over a period of 9 months, from 0 to 28 days of advance booking. Throught a panel-VAR approach we assess if the typical planning of the price trajectory, including dummies for holidays and fairs as covariates. Results suggest that strategies put into effect by firms reflect some of the basic principles of the online revenue management. Price trajectories are planned considering both firms expectations on the prices they hope to charge last-minute, and their need to guarantee price stability along the advance booking. Fairs and holidays show a different impact on price dynamics. While the response caused by an ”holiday shock” tends to be flat, room rates during fairs raise in the immediate future, then accommodate to the equilibrium in about three days.

Key words: RevPOR, RevPAR, Dynamic Pricing, panel-VAR

1 Introduction

In recent years, rapid technological progress and the proliferation of on-line booking platforms have deeply modified the behaviour of touristic firms, in particular concerning pricing and the management of the occupation rate. The opportunity to

AndreaGuizzardi
Department of Statistics, University of Bologna, Via delle Belle Arti 41, Bologna, e-mail: andrea.guizzardi@unibo.it

Flavio Maria Emanuele Pons
Department of Statistics, University of Bologna, Via delle Belle Arti 41, Bologna, e-mail: flaviomaria.pons2@unibo.it

update quality and prices of the room offered on-line in real time has boosted the development of new methods and algorithms to perform an effective revenue management. Moreover, the great availability of free data deriving from on-line travel agencies increases the transparency of the market and the possibility to study firm competition.

Considering the industry characteristics (perishable inventory, short-term constrained capacity, high fixed costs respect to variable costs) managers are motivated to adjust prices in real time in order to sell all the rooms out by the target day emphasizing maximization of daily revenues ([Wang and Brennan, 2014]). However, hotels operate a segmentation of the room market, considering that there exists a negative correlation between the ability to purchase and the advance between booking and arrival date. In fact, clients with higher spending possibility, in general business travellers, tend to make a reservation on the target day or with a short notice of one or two days ([Guo et al., 2013]).

This situation induces a trade-off between the strong incentive to sell all the rooms by the target day (minimizing strategically - unsold capacity) and the higher profitability of the rooms sold with a short advance booking (maximizing tactically - average Revenue Per Occupied Room (RevPOR)). In the on-line market, the advance booking enables firms to mix strategic and tactic pricing, in order to maximize Revenue Per Available Room (RevPAR). This operation is not straigtforward, particularly in a location, such as Milan, where the low elasticity to price of business customers concurs in differentiating the economic effect of tactic and strategy in revenue management.

Previous researches (e.g. [Abrate and Viglia, 2016]) have exploited the full potential of a panel dataset to explain both cross-section and time variability of prices considering also contextual variables like: room and hotel quality, offered services, restriction placed on prices or the spatial density of competitors or destination occupancy. Price is taken as response variable in a random-effect regression model and advance booking is an explicative variable. However, [Guizzardi et al., 2017] found that price trajectories can be seen as a stationary AR(1) process, suggesting that dynamics is present not only between successive arrival dates, but also along the advance booking, the role of which cannot then be properly assessed simply considering the booking lag as a covariate.

In this article, we overcome this issue by considering a panel-VAR model. The VAR setting enables us to consider the price trajectory as a multivariate endogenous variable, modelling also the interdependencies between different advance bookings, other than the serial correlation in each time series. The panel generalization of VAR models lets us build consistent GMM estimators considering the cross-sectional nature of the dataset. Moreover, the computation of the impulse response functions (IRF) enables us to interpret the effect of exogenous shocks (in our case holidays and fairs) in terms of forecasted price response. Thanks to this approach, we assess: if and how hotel managers account for demand patterns and exogenous shocks in their price competition on the on-line market; if and when they change from strategy to tattic along the advance booking.

The rest of the paper is organized as follows: in Section 2 we describe the dataset and the modellistic framework, introducing the general formulation of panel-VAR models, and discussing the techniques adopted for model selection and estimation. In Section 3 we present the main results, which include significant model coefficients and the relevant impulse response functions; in Section 4 we discuss briefly our findings, linking them to the existing literature about dynamic pricing, and present our concluding remarks and an outline of future extensions of this work.

2 Data and Methods

We consider a dataset consisting of best available rates (BARs) recorded every day at 00:00 AM for a panel of 107 hotels in Milan, from January, 1st to September, 30th, 2016. The data source was `Expedia.com`. For each arrival date, the room price has been recorded from 28 to 0 days of advance booking.

Let $i = 1,2,\dots,N = 107$ index the hotel, $t = 1,2,\dots,T = 274$ the arrival date and $k = 0,1,\dots,K = 28$ the number of days of advance booking: let us denote the natural logarithm of the BAR for hotel i, arrival date t and advance booking k as $y_{it}^{(k)}$. We call *price trajectory* for the arrival date t at hotel i the vector of values $y_{it} = \{y_{it}^{(k)}\}_{k=0}^{28}$. In order to reduce dimensionality, we consider a sub-sample of the price trajectory, including a limited number of values of the advance booking to sample short, medium and long term with respect to the arrival date. In particular, we only consider lags = 0, 1, 7, 14, 21, 28, so that we reduce the dimension of the dependent variable from $K = 29$ to $K = 6$.

Given the panel nature of our dataset, and that for each hotel y_{it} is a vector time series, it is natural to consider the extension to panel-VAR models, which permits to take simultaneously into account the longitudinal and multivariate time series nature of the data. From our preliminary analysis, we establish that none of the hotel-specific VARs contains cointegration, so that no differencing is needed to achieve stationarity.

The panel-VAR model of order p reads ([Abrigo et al., 2016])

$$y_{it} = \sum_{j=1}^{p} y_{it-j}A_j + x_{it}B + u_i + \varepsilon_{it} \tag{1}$$

where y_{it} is a $(1 \times k)$ dependent vector, x_{it} is a $(1 \times l)$ vector of exogenous variables, u_i is a $(1 \times K)$ vector of hotel-specific fixed effects and ε_i is a $(1 \times K)$ vector of idiosyncratic errors, such that $E(\varepsilon_{it}) = 0$, $E(\varepsilon_{it}'\varepsilon_{it}) = \Sigma$, $E(\varepsilon_{it}'\varepsilon_{is}) = 0$ for all $t > s$. A_j and B are, respectively, $(K \times K)$ and $(l \times K)$ matrices of parameters to be estimated.

The covariates x_{it} consist of four dummies indicating weekend nights (from Friday to Sunday), national holidays, fairs, and the month of August, during which business activity in Italy is considerably reduced. From a preliminary analysis of their marginal effects, we found that seasonal variability is not very evident, except for the months of August and September, during which prices are, respectively, lower and higher than the annual average. However, we can only link to tourism seasonality the lower price levels during August, while the higher fares in September are expectedly due to particular events, such as an high number of fairs during the month. Moreover, weekend nights appear to be associated to prices lower that the weekly average, including Fridays. Regarding short events displacing demand from its equilibrium state, fairs are associated to significantly higher room rates, while holidays do not seem to marginally affect room prices in Milan.

Estimation

We use the package `pvar` in `Stata`, described in [Abrigo et al., 2016], to specify the panel-VAR model in Eq. 1. Model estimation is conducted in a GMM framework, while model selection is based on model and moment selection criteria (MMSC).

Concerning the estimation of the coefficient matrices A and B in Eq. 1, the package follows a GMM approach which generalizes the bivariate procedure introduced by [Holtz-Eakin et al., 1988]. As suggested by [Arellano and Bover, 1995], we remove fixed effects through forward orthogonal deviation (FOD), which consists of subtracting the mean of only the future observations at any time and for any unit. This way, the variables transformed to remove fixed effects and the lagged dependent variables remain orthogonal, allowing to use the latter as instruments for estimation. The FOD procedure results in the following transformed variables, which are subsequently used for GMM estimation: $y^*_{it} = \frac{y_{it} - \overline{y_{it}}^+}{\sqrt{T_{it}/(T_{it}+1)}}$, where $\overline{y_{it}}^+$ is the mean of the future observations at time t, T_{it} is the number of remaining future observations at time t, and y^*_{it} is a $(1 \times K)$ vector $y^*_{it} = [y^{*1}_{it}\, y^{*2}_{it} \cdots y^{*K}_{it}]$.

Considering the model as a system, rather than equation-by-equation, makes it possible to achieve efficiency. In practice, this is obtained by writing the model equation in reduced form, $y^*_{it} = \tilde{y}^*_{it} A + \varepsilon^*_{it}$. The resulting GMM estimator for A is obtained stacking observations over panel and then over time, $\hat{A} = (\tilde{y}'^* z \hat{W} z' \tilde{y}^*)^{-1} (\tilde{y}'^* z \hat{W} z' \tilde{y}^*)$ where z are the available instruments, and $\hat{W}$ is a $(L \times L)$ weighing matrix, with $L = Kp + l$, assumed to be nonsingular, symmetric, positive semidefinite, and can be chosen to maximize efficiency according to [Hansen, 1982]. The GMM estimator is consistent if the orthogonality condition between instruments and idiosyncratic errors $E(z'\varepsilon) = 0$ holds and $rank E(\tilde{y}'^*_{it} z) = Kp + l$

Model selection

Panel-VAR model selection is based on the choice of a maximum lag for the set of instrumental variables and for the autoregressive model. For our dataset, we choose $P = 8$ as maximum autoregressive order; this choice is made considering that we may observe weekly dependencies which could be better caught by an AR(7) model, so that we want $p = 7$ to be among the eligible values. The selection of the maximum lag of the endogenous variables to be used as instruments is less arbitrary. With a view to testing the validity of the restrictions imposed by the postulated model, [Hansen, 1982] proposed an extension of the specification test introduced by [Sargan, 1959]. Such test requires over-identification, i.e. the number of available instruments must be larger than $kP+l$: since the exogenous variables provide the l valid instruments, we need to choose a number $q > p$ of lagged endogenous K-dimensional variables. Moreover, to avoid consistency loss due to endogeneity, instrumental variables are valid if incorrelated to the estimation variables, so that the smallest lag available as a valid instrument for y_{it-1} is y_{it-2}, and the smallest lag available for y_{it-2} is y_{it-3}. In practice, the minimum set of instruments for a panel-VAR(p) model is $[y_{it-2} \cdots y_{it-p-2}]$, which also implies that, while conducting the model selection, for every set of q instruments, we can only achieve over-identification up to $p = q-1$.

The order selection is based on model and moment selection criteria (MMSC) introduced by [Andrews and Lu, 2001] to include the number of parameters p, the number of instruments q and the test statistics $J(p,q,K)$ for over-identification defined by [Hansen, 1982].

3 Results

The procedure described in the previous section leads to the final selection of a panel VAR(2) with $[y_{it-2}, \cdots, y_{it-6}]$ as instruments to guarantee overidentification. The modulus of all the eigenvalues of the companion matrix lies inside the unit circle, so that the model is stable. Sargan's test does not reject the null hypothesis of correctly specified model, with a p-value = 0.080. Estimation results are shown in Table 3, only for the exogenous variables and significant coefficients of endogenous variables.

The only endogenous variables that are significant in an explanatory sense are prices at advance booking 0, 1 and 21. Prices at advance booking 0 and 1 are significant "predictors" of prices at larger advance bookings for arrival dates in the immediate future (which have been fixed several days or weeks earlier). This indicates that hotel managers not only practice revenue management in a dynamic pricing framework, but also that along the considered advance booking they make an implicit prediction of the last-minute occupation rate and of competitors' prices.

Concerning the exogenous variables, we can notice that periodic effects, such as month and day of week, have significant effects at all the considered lags of advance booking. Also fairs significantly affect price levels at all advance bookings along the trajectory, always with positive coefficients. On the other hand, while holidays do not show significant marginal impact on prices, a small positive holiday effect can be recognized for advance bookings in the close proximity to the arrival date. The weakness of this holiday effect and its limited extension along the trajectory, opposed to the stronger and always significant impact of fairs reflect the vocation of Milan as a business destination. We can also notice that the negative effect due to weekends is stronger for last-minute deals, while discounts associated to Summer holidays are stronger for early purchasers. Finally, the positive effect of fairs seems to be higher for larger advance booking. This highlights, for the examined period, a tendency of managers to be too optimistic about the increase in occupation rates stimulated by fairs, and then on feasible prices. This reflects a recent decline in the fair market in Milan, as noted by [Guizzardi, 2016].

Thanks to the possibility to compute impulse response functions (IRF), we can interpret the role of these effects also in a forecasting setting. Dynamic multipliers, i.e. the IRFs for exogenous variables, are shown in Fig. 1 and 2: while the IRF of the holiday dummy is flat on zero at all future lags, fairs clearly result in higher prices in the immediate future (i.e. the duration of the event), with a return to equilibrium around three days later. The estimated IRFs are shown with 95% Monte Carlo (500 replicates) confidence intervals.

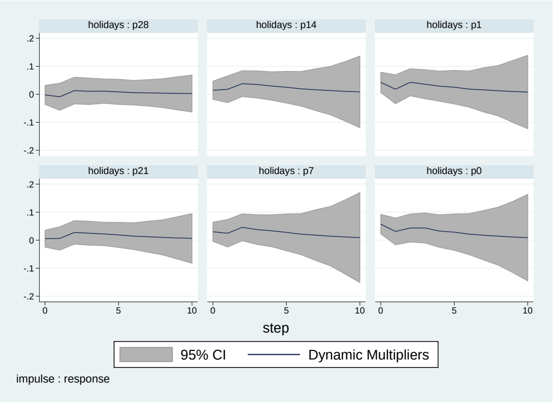


Fig. 1 Impulse response function on a 10 days horizon for holiday shocks. The gray shaded area marks the 95% confidence intervals obtained from 500 Monte Carlo replications.

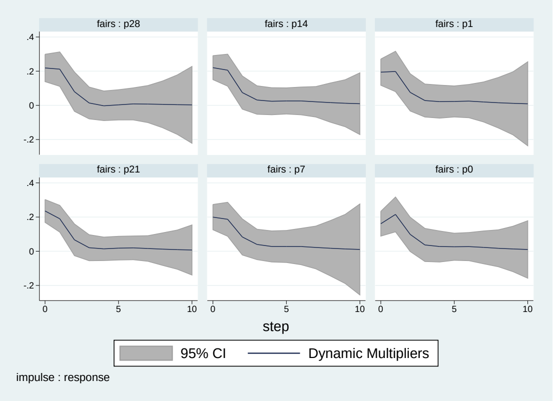


Fig. 2 Impulse response function on a 10 days horizon for fair shocks. The gray shaded area marks the 95% confidence intervals obtained from 500 Monte Carlo replications.

Table 1 Panel-VAR(2) significant coefficients.

y^0	Coef.	Std. Err.	p-value	y^1	Coef.	Std. Err.	p-value	y^7	Coef.	Std. Err.	p-value
y^0 lag=1	-	-	-	y^0 lag=1	-	-	-	y^0 lag=1	1.450	0.553	0.009
lag=2	0.310	0.114	0.007	lag=2	0.336	0.119	0.005	lag=2	0.234	0.107	0.029
y^1 lag=1	-1.514	0.702	0.031	y^1 lag=1	-1.830	0.710	0.01	y^1 lag=1	-2.126	0.673	0.002
lag=2	-	-	-	lag=2	-	-	-	lag=2	-	-	-
y^{21} lag=1	-	-	-	y^{21} lag=1	-	-	-	y^{21} lag=1	1.456	0.616	0.018
lag=2	-0.268	0.356	0.451	lag=2	-0.606	0.253	0.017	lag=2	-0.666	0.243	0.006
weekend	-0.199	0.015	0.000	weekend	-0.196	0.016	0.000	weekend	-0.144	0.016	0.000
august	-0.097	0.040	0.017	august	-0.121	0.042	0.004	august	-0.117	0.040	0.003
holidays	0.057	0.018	0.001	holidays	0.043	0.019	0.024	holidays	-	-	-
fairs	0.161	0.038	0.000	fairs	0.194	0.038	0.000	fairs	0.200	0.036	0.000
y^{14}	Coef.	Std. Err.	p-value	y^{21}	Coef.	Std. Err.	p-value	y^{28}	Coef.	Std. Err.	p-value
y^0 lag=1	1.056	0.519	0.041	y^0 lag=1	0.960	0.476	0.044	y^0 lag=1	1.172	0.572	0.04
lag=2	0.226	0.102	0.026	lag=2	0.208	0.092	0.024	lag=2	-	-	-
y^1 lag=1	-1.660	0.634	0.009	y^1 lag=1	-1.49	0.58	0.01	y^1 lag=1	-1.592	0.671	0.018
lag=2	-	-	-	lag=2	-	-	-	lag=2	-	-	-
y^{21} lag=1	1.880	0.531	0.001	y^{21} lag=1	1.950	0.645	0.044	y^{21} lag=1	2.078	0.591	0.000
lag=2	-0.763	0.239	0.001	lag=2	-0.634	0.361	0.024	lag=2	-0.854	0.233	0.000
weekends	-0.125	0.142	0.000	weekend	-0.123	0.013	0.000	weekend	-0.111	0.015	0.000
august	-0.130	0.038	0.001	august	0.141	0.035	0.000	august	-0.181	0.041	0.000
holidays	-	-	-	holidays	-	-	-	holidays	-	-	-
fairs	0.221	0.034	0.000	fairs	0.236	0.032	0.000	fairs	0.219	0.039	0.000

4 Discussion

In this work, we have shown that hotels of the high rating segment in Milan determine on-line BARs on the base of both their intentions about planning last-minute prices and the need to guarantee a certain stability of the price trajectory along the advance booking. In particular, the causality structure resulting from the VAR estimation suggests the significance of both last-minute and mid-term (three weeks in advance) room rates to explain the price dynamics. Prices in the mid-range of the trajectory (7-14 days) are not significant in an explanatory sense. The balance between tactic and strategic pricing holds up until short advance booking, where the tactic gains importance. Rates at advance booking 0 and 1, become negatively correlated with pricing strategies pursued at 21 days of advance booking. This negative correlation reflects tactic price adjustments in response to under/overbooking levels, caused by improper pricing choices in the mid-term, or an incorrect assessment of competitors on-line last-minute pricing. This behaviour is amplified by the prevalence of business travellers, who tend to make reservation with short advance booking while having low elasticity towards price.

We included four exogenous variables, all controlling for periods or events that modify the destinations business activity, so that we expect them to change the level of aggregate demand for the destination, occupation rates and, consequently, the pressure on firms to discount or increase prices. We find that fairs have a positive effect on prices, slightly higher for larger advance bookings, while the negative coefficients corresponding to weekends and the month of August call for a discount strategy. It is worth to notice that the negative effect due to Summer holidays is stronger for early purchasers, while discounts associated to weekends are stronger for last-minute deals. This suggest that managers have high confidence in planning pricing during August, while they give less weight to strategic pricing regarding weekends, given the unpredictability of weekend occupation rates, and lacking early information about: weather, events for leisure tourists and/or the concomitance with other significant events (also in the adjacent days) that interest the business segment. Also the effect of national holidays is significant only in short-term, while positive, for the reasons just mentioned regarding weekend pricing and for the fact that holidays are irregularly placed during the year. Overall, it appears that offer is not saturated until the advance booking is very small, so that pricing strategies defined three weeks in advance are only adjusted very close to the target date.

Concerning shock propagation, the impulse response functions for fairs and holidays show a different impact on price dynamics. While the response caused by a holiday tends to be zero at all future lags, fairs clearly result in higher prices in the immediate future (i.e. the average duration of the event), with a return to equilibrium around three days later.

In synthesis, we show that the actual strategies put into effect by the hotel managers reflect some of the basic principles of the online revenue management. Further development of this work will assess the question if, how, and to what extent, while practicing this mixture of tactical and strategic dynamic pricing, firms base their on-line pricing behaviour on the observation of prices published online by their competitors.

References

[Abrate and Viglia, 2016] Abrate, G. and Viglia, G. (2016). Strategic and tactical price decisions in hotel revenue management. *Tourism Management*, 55:123–132.

[Abrigo et al., 2016] Abrigo, M. R., Love, I., et al. (2016). Estimation of panel vector autoregression in stata. *Stata Journal*, 16(3):778–804.

[Andrews and Lu, 2001] Andrews, D. W. and Lu, B. (2001). Consistent model and moment selection procedures for gmm estimation with application to dynamic panel data models. *Journal of Econometrics*, 101(1):123–164.

[Arellano and Bover, 1995] Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of econometrics*, 68(1):29–51.

[Guizzardi, 2016] Guizzardi, A. (2016). Viaggi, fatturato e soddisfazione dei clienti del turismo daffari italiano nel 2015. *Turismo dAffari.*

[Guizzardi et al., 2017] Guizzardi, A., Pons, F. M. E., and Ranieri, E. (2017). Advance booking and hotel price variability online: Any opportunity for business customers? *International Journal of Hospitality Management*, 64:85–93.

[Guo et al., 2013] Guo, X., Ling, L., Yang, C., Li, Z., and Liang, L. (2013). Optimal pricing strategy based on market segmentation for service products using online reservation systems: An application to hotel rooms. *International Journal of Hospitality Management*, 35:274–281.

[Hansen, 1982] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

[Holtz-Eakin et al., 1988] Holtz-Eakin, D., Newey, W., and Rosen, H. S. (1988). Estimating vector autoregressions with panel data. *Econometrica: Journal of the Econometric Society*, pages 1371–1395.

[Sargan, 1959] Sargan, J. D. (1959). The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 91–105.

[Wang and Brennan, 2014] Wang, X. L. and Brennan, R. (2014). A framework for key account management and revenue management integration. *Industrial Marketing Management*, 43(7):1172–1181.

The Migration-Led-Tourism Hypothesis for Italy: A Survey

Rassegna sull'evidenza empirica a sostegno della relazione tra turismo e fenomeno migratorio in Italia

Carla Massidda, Romano Piras and Ivan Etzo[1]

Abstract Over recent decades, the literature on trade and factor mobility has shown an increasing interest in the relationship between migration and tourism. This literature has favored the demand side perspective, leaving the supply dimension almost entirely overlooked. Among the countries interested to investigate the tourism-migration nexus, Italy represents an attractive case to study. The present contribution provides a brief survey of the main empirical findings regarding this Country.

Abstract Nel Corso degli ultimi decenni, la letteratura sul commercio internazionale ha mostrato un interesse crescente verso la relazione tra turismo e fenomeno migratorio. In seno a tali studi, il maggior interesse si è concentrato sulla prospettiva della domanda, mentre la dimensione dell'offerta turistica è stata quasi completamente trascurata. Tra i paesi particolarmente interessanti da analizzare, per la sua storia passata e recente, vi è certamente l'Italia. Il presente contributo ha lo scopo di offrire una rassegna dei principali risultati finora raggiunti per questo paese.

Key words: tourism, migration, migration-led-tourism

1. Introduction

Starting from the works of Jackson (1990) and Williams and Hall (2002), the literature on trade and factor mobility has shown an increasing interest in the relationship between migration and tourism. Williams and Hall (2002) discuss three mechanisms that can link tourism demand to migration: a causal relationship running from migration to tourism trips (migration-led-tourism, MLT) motivated by the pleasure to visit friends and relatives (VFR); a causation from tourism to migration (tourism-led-migration, TLM); a bi-directional causal link between these

[1] Carla Massidda, University of Cagliari; massidda@unica.it
Romano Piras, University of Cagliari, pirasr@unica.it
Ivan Etzo, University of Cagliari, ietzo@unica.it

two phenomena. Although empirical findings effectively places migration among the major determinants of VFR flows, more recent studies favour an extensive interpretation of the MLT hypothesis (cf., inter al., Etzo et al., 2014). Accordingly, immigrants can pull arrivals and push departures, whether or not VFR is the main motivation. In the footsteps of this initial work, empirical studies on this issue have grown considerably and, almost unanimously, they conclude in favour of the MLT (Cf. *inter al*., Boyne et al. (2002), Prescott *et al*. (2005), Seetaram and Dwyer (2009), Gheasi *et al*. (2011), Seetaram (2012a,b), Tadesse and White (2012), Leitão and Shahbaz (2012), Genç (2013), Law et al. (2013), Massidda *et al*. (2015), Etzo et al. (2014), Massidda and Piras (2015), Balli *et al*. (2016), Etzo (2017), Massidda *et al*. (2017).

Surprisingly, this literature has almost entirely overlooked the supply-side perspective. This point is discussed in Massidda et al. (2017) on the base of the recent literature regarding the general impacts of immigration on host economies (cf. *inter al*., Ottaviano and Peri, 2012; Olney, 2013). The main issue is that, taking fully into account the skill complementarities between workers, immigrants can be an opportunity for receiving countries. On the one hand, immigrants, pulling tourist inflows, stimulate the demand for goods and services and, therefore, the production activities aimed at its fulfilment. On the other hand, foreign population represents workforce, mostly low-skilled, which often allows firms to pay lower wages and increase the scale of production. At the same time, immigrants can be entrepreneurs that, again, contribute to raise the number of firms and their employees.

Among the countries that are potentially interested to investigate the tourism-migration nexus, Italy represents an interesting case to study. This is because tourism and migration have strongly characterised the history of this Country, influencing its current social and economic shape. Italy has been an important source of international migration and, during the last decades, it has also become a country of immigration. At the same time, today Italy ranks among the top tourism destinations and top spenders worldwide. In spite of these features, there is a lack of empirical research on the determinants of Italian tourism flows (outbound, inbound and domestic) and, above all, a lack of investigations on the mechanisms linking tourism to migration. Only recently, has the literature started to fill this gap with the works of Massidda *et al*. (2015), Etzo *et al*. (2014), Massidda and Piras (2015) and Massidda *et al*. (2017). The aim of the present paper is to provide a survey of the main empirical findings provided until now.

2. Tourism and migration in Italy: an overview

The Italian tourism sector has increased steadily over time becoming one of the largest worldwide. In 2016, domestic tourism accounts for about 51% of the total industry for both arrivals and nights. With respect to arrivals (ISTAT, 2017), the Northern macro-area registers the highest share (55%), followed by Centre (23%) and South (22%). These shares are only slightly different in term of nights (52%, 23%, 25%, respectively). Regional market shares highlight Emilia Romagna, Lombardia, Veneto and Toscana amongst the most popular destinations for arrivals, whereas Emilia Romagna, Veneto, Toscana and Trentino are amongst the regions with highest number of nights. Holiday purpose is the main factor driving Italian domestic tourism, followed by VFR. As far as inbound tourism (UNWTO, 2017), among the top world destination countries, in 2016 Italy ranks 5th for arrivals and 6th for receipts. Germany is the main sending country, followed by France and United Kingdom (Bank of Italy, 2017). In terms of percentage variation, the eastern European countries show the highest growth rates, in particular Russia and Poland. Expenditures are mainly motivated by holiday purposes (68%), followed by business visits (14%) and VFR (10%). Turning the attention to outbound tourism, Italy ranks 8th among the main source markets in terms of expenditure (UNWTO, 2017). In 2016 the number of total outbound travellers has grown by approximately 1.6% and the European countries are the preferred destinations (Bank of Italy, 2017). Expenditure are higher in France, USA and Spain. With regard to the main purpose of visit, expenditures are mainly motivated by holiday purposes (40%), followed by business visits (34%) and VFR (9%).

Turning the attention to immigration flows, as it is well known, during the last decades of the 20th century Italy experienced a transition from being one of the most important sending countries to become also one of the principal destinations. According to the Registry of Italian citizens Residing Abroad, the stock of Italians abroad has reached 4.9 million in 2016, while the stock of immigrants residing in Italy is slightly above 5 million (Fondazione Leone Moressa, 2017). Romania is the first sending country, followed by Albania and Marocco. As far as internal migration is concerned, it started to boom after WWII: as it is well known in about twenty years (1951-1975), more than three million of individuals moved from the South to the Central and Northern areas of the country. During the seventies up to the mid-eighties, the internal migration flows started to decrease, to re-gain momentum from the mid-nineties onward (Piras, 2005). In particular, estimates show that the stock of internal migrants climbed to over 13 million in 2010, approximately one fifth of the Italian population. Lombardia, Piemonte and Lazio are the regions with the highest stocks of immigrants. Conversely, the lowest stocks are reported by Valle d'Aosta, Molise and Basilicata.

3. Migration and inbound tourism flows in Italy

The present section provides evidence of the channels through which migration stocks, defined at both origin and destination, affect Italian inbound tourism. The main source of this analysis is the contribution of Massidda *et al.* (2015). The model of tourism demand is specified in terms of total arrivals and arrivals disaggregated by purpose of visits, i.e. Visiting Friends and Relatives (VFR), Holiday and Business. The other covariates have been chosen following the extant literature. The investigation is performed over the period 2005-2011 and considers a panel of 65 countries. Data are taken from Bank of Italy, ISTAT, AIRE (Archivio degli italiani residenti all'estero) and World Bank.

To perform the analysis, with lower-case letters that indicate log-transformed variables, the following dynamic econometric model is specified:

(1) $y_{i,t,m} = \beta_0 + \beta_1 y_{i,t\text{-}1,m} + \beta_2 m\_ita_{i,t} + \beta_3 m\_for_{i,t} + \beta_4 gdp_{i,t} + \beta_5 rer_{i,t} + \beta_6 dist_i + \beta_7 cont_i + \gamma_t + \mu_i + \varepsilon_{i,t}$

where the subscript i=1, 2, … 65 denotes the countries of origin, the subscript t=1, 2, … 7 refers to the time period and m stands for the purpose of the visit. The empirical estimation is carried out by means of the one step system GMM estimator and the variables are defined as follows:[2]

- $y_{i,t,m}$ indicates arrivals;
- $m\_ita_{i,t}$ is the stock of Italian citizens residing in country i;
- $m\_for_{i,t}$ is the stock of country i's foreign citizens residing in Italy;
- $gdp_{i,t}$ is the real per capita GDP in the source country;
- $rer_{i,t}$ is the real exchange rate.
- $dist_i$ is the distance between Rome and the Capital of country i;
- $cont_i$ is a dummy variable that controls for common border effects;
- μ_i and γ_t are country-specific and time-specific fixed effects, respectively;
- $\varepsilon_{i,t}$ is the stochastic error term.

The main empirical findings provided by the cited study is that the stocks of Italian citizens residing abroad have a positive and statistically significant impact on total arrivals, on VFR (with the highest estimated coefficient) and on Holiday tourism. As for $m\_for_{i,t}$, the estimated coefficient is positive and statistically significant for all the trip categories. The highest impact is detected for Holiday tourism. A positive impact is detected also for $gdp_{i,t}$. Holidays demand, with an above unity estimated coefficient, seems to behave as a luxury good, whereas VFR and Business trips report the characteristic of a normal good. Among the other covariates, distance, when significant, report a negative influence on tourism flows, while the real exchange rate seems to not having any influence on international arrivals. Further comments on results are avoided for the sake of space.

4. Migration and outbound tourism flows from Italy

This section focuses on the relationship between migration and Italian outbound tourism. The main source for this analysis is the investigation proposed by Etzo *et al.* (2014). The empirical setting is closely related to the study proposed in Section 3, whereas now the dependent variable is specified in terms of departures. Data are

[2] For a more detailed description of the variables, data source and empirical methodology, see the original paper.

taken from Bank of Italy, ISTAT, AIRE (Archivio degli italiani residenti all'estero) and World Bank. The dynamic econometric model is specified as follows, with lower-case letters that indicate log-transformed variables:

$$(2)\quad y_{i,t,m} = \beta_0 + \beta_1\, y_{i,t\text{-}1,m} + \beta_2\, m\_ita_{i,t} + \beta_3\, m\_for_{i,t} + \beta_4\, p_{i,t} + \beta_5\, gdp_t + \beta_6\, dist_i + \beta_7\, cont_i + \beta_8\, crt_i + \beta_9\, (p_{i,t} \times crt_i) + \gamma_t + \mu_i + \varepsilon_{i,t}$$

where the subscript i=1, 2, … 65 denotes the destination country, the subscript t=1, 2, …7 refers to the time period and m stands for the purpose of visit. Again, the empirical estimation is carried out by means of the one step system GMM estimator and the variables are defined as follows:[3]

- $y_{i,t,m}$ represents the number of Italian outbound tourism trips;
- $m\_ita_{i,t}$ is the stock of Italian citizens residing in country i;
- $m\_for_{i,t}$ is the stock of country i's foreign citizens residing in Italy;
- $p_{i,t}$ is the log of the price competitiveness index $P_{i,t}$ (destination i is cheaper than Italy with a value of $P_{i,t}$ lower than one);
- gdp_t is the annual Italian real *GDP per capita*;
- $dist_i$ is the distance between Rome and the Capital of country i;
- $cont_i$ is a dummy variable that controls for the presence of common border effects
- crt_i is a dummy variable which takes on the value of one if real GDP per capita at destination is equal to or higher than Italian's real GDP per capita;
- $(p_{i,t} \times crt_i)$ is the interaction that allows to differentiate the effect of price depending on the development level at destination.

The main results is the positive impact of the stocks of Italian citizens residing abroad on all tourist groups. The highest effect is registered for Holiday. As for the effect of $m\_for_{i,t}$, the impact is statistically significant for all groups of tourists, but for Holiday. VFR trips seem to show the highest estimated impact. Turning the attention to the other covariates, it is interesting to observe the coefficient reported by the price competitiveness index. In line with expectations, when significant, it reports a positive coefficient for countries less developed than Italy and negative for countries as developed as Italy. The remaining covariates report the expected estimated impacts. Further comments are avoided for the sake of space.

5. Migration and domestic tourism in Italy

This section is devoted to the discussion of the relationship between internal migration and domestic tourism in Italy. The main source for this analysis is the contribution of Massidda and Piras (2015) where it is proposed a heterogeneous dynamic panel investigation of the role of interregional migration on domestic

[3] For more details, see the original paper.

tourism for the twenty Italian regions over the period 1987-2010. For the scope, the Pool Mean Group (PMG) panel estimation procedure is applied. The main sources of the data are ISTAT (various sources and years) and Svimez (2011).

The analysis is performed within a tourism demand theoretical framework. The number of per capita domestic bed nights in region *i* and time *t* ($NP_{i,t}$) is assumed to depend on the ratio of the stock of internal migrants to resident population in the region of destination ($Imp_{i,t}$). This variable, in absence of official data, has been constructed following the approach of White (2007). Other determinants are[4]:

- $Pr_{i,t}$ is the relative prices index;
- $Yp_{i,t}$ is real per capita gross domestic product at destination;
- $Hp_{i,t}$ is the number of per capita hotels;
- $Dens_{i,t}$ captures the population density calculated as the ratio of population per square kilometre.

Since the variables are proved to be $I(1)$ and cointegrated, the following error-correction model can be considered (lower-case letters indicate log-transformed variables):

$$(3) \quad \Delta np_{i,t} = \phi_{\mathrm{i}} \, (np_{i,t\text{-}1} - \beta_{0,i} - \beta_{1,i}\, pr_{i,t} - \beta_{2,i}\, yp_{i,t} - \beta_{3,i}\, imp_{i,t} - \beta_{4,i}\, hp_{i,t} - \beta_{5,i}\, dens_{i,t}) + \gamma_{11,i}\, \Delta pr_{i,t} + \gamma_{21,i}\, \Delta yp_{i,t} + \gamma_{31,i}\, \Delta imp_{i,t} + \gamma_{41,i}\, \Delta hp_{i,t} + \gamma_{51,i}\, \Delta dens_{i,t} + \varepsilon_{i,t}$$

where all coefficients are obtained as re-parameterization of an ARDL (1, 1, 1, 1, 1, 1). In Eq (3), ϕ_i represents the error correction coefficient, whereas $\varepsilon_{i,t}$ indicates the remainder disturbance term which is assumed to be independently distributed across *i* and *t*, with mean 0 and variance $\sigma_i^2 > 0$.

The main outcome of this study is the positive long-run nexus between internal migration and domestic tourism nights. Moving to the other covariates, Italian residents seem highly discouraged by price fluctuations (*pr*), whereas they are attracted by the general wellbeing of a region (*yp*), by tourism infrastructure (*hp*). A positive impact is also estimated for the variable *dens*.

6. Migration and tourism firms in Italy

The relationship between immigration and the Italian tourism supply has been the issue of the recent contribution of Massidda *et al.* (2017). This work provides an investigation of the relationship between the share of working age foreign-born population and the number of local units and their employees in the Hotel and Restaurant sector. The investigation is carried out at nation-wide level and, separately, for Centre-Northern and Southern macro-areas. Data refer to a panel of 103 Italian provinces during the 2004-2010 time period and are taken from the Statistical Archive of Active Enterprise (ASIA-ISTAT) for local units, from ISTAT

[4] For details on variables definition, data source and econometric issues, see the original paper.

archives for immigrants and from ISTAT data warehouse for unemployment and population density.

The empirical model is specified as follows (lower-case letters indicate log-transformed variables):

$$(4) \quad y_{i,t} = \beta_0 + \beta_1 sh\_Imm_{i,t} + \beta_j x_{i,t} + \gamma_t + \mu_i + \varepsilon_{i,t}$$

where the dependent variable $y_{i,t}$ is, alternatively, the log of the number of establishments and of the number of employees, in province *i* at year *t*. The variable of interest ($sh\_Imm_{i,t}$) is the log of the share of working age foreign born population resident in province *i* at year *t*, whereas $x_{i,t}$ is a set of three control variables including population density, unemployment rate and the growth rate of value added per capita. Fixed effects μ_i and temporal dummies γ_t are also included in the model. Finally, $\varepsilon_{i,t}$ is the error term uncorrelated with the covariates. To avoid inconsistency due to the potential endogeneity of the migration variable, estimates are obtained through the two stages least square (2SLS) estimator.

The main outcome of this analysis is that the share of immigrants positively affects the number of establishments, at both national and macro-area level. At macro-area level, this relationship appears to be stronger in the South. As for the effect of the other covariates, population density always reports a positive and statistically significant coefficient, whereas the unemployment rate seems to negatively affect only the number of local units in Southern provinces. Results on employees again highlight the positive influence of the provincial share of immigrants. Here again, Southern provinces are more responsive than Centre-Northern ones. As for the other covariates, population density keeps its positive role in the H&R sector and the unemployment rate negatively affects employment in the South.

Conclusions

Recent empirical studies demonstrate that in Italy the two phenomenon of tourism and migration show important connections. From the demand side perspective, it seems that the foreign-born communities residing in Italy and Italian emigrants leaving abroad stimulate both inbound and outbound tourism flows. A strong relationship also exists between interregional migration and the domestic tourism demand. On the supply side perspective, it seems that, besides opening new establishments and/or re-locating existing ones, Italian tourism firms respond to the relatively abundancy of foreign workers by increasing the labor demand.

References

1. Balli F., Balli HO & Jean Louis R. (2016). The impacts of immigrants and institutions on bilateral tourism flows. Tourism management, 52: 221-229.
2. Bank of Italy (2017), Indagine sul turismo internazionale, Bank of Italy website.

3. Boyne S., Carswell F. & Hall D. (2002), Reconceptualising VFR Tourism: Friends, Relatives and Migration in a Domestic Context, in Hall C. M. & Williams A. M. (eds), Tourism and Migration: New Relationship between Production and Consumption, Kluwer Academic Publishers, Dordrecht.
4. Etzo I., Massidda C. and Piras R. (2014). Migration and outbound tourism: evidence from Italy. Annals of Tourism Research, 48: 235-249.
5. Etzo I. (2016). The impact of migration on tourism demand: evidence from Japan. Review of Asian and Pacific Studies, No. 41, pp. 79-98.
6. Fondazione Leone Moressa, (2017). Rapporto annuale sull'economia dell'immigrazione. Bologna, Il Mulino.
7. Genç, M. (2013), Migration and tourism flows to New Zealand, in Matias Á., Nijkamp P., & Sarmento M. (eds), Quantitative Methods in Tourism Economics. Springer-Verlag, Heidelberg.
8. Gheasi, M., Nijkamp, P., & Rietveld, P. (2011). Migration and Tourism flows, in Matias, Á., Nijkamp, P., & Sarmento, M. (eds.), Tourism Economics, Physica-Verlag, Heidelberg.
9. Jackson R. (1990). VFR tourism: is it underestimated? Journal of Tourism Studies, 1(2): 10-17.
10. ISTAT (2017), Annuario statistico italiano, ISTAT web site.
11. Law D., Genç, M., & Bryant, J. (2013). Trade, diaspora and migration to New Zealand. The World Economy, 36, 582-606.
12. Leitão N. C. & Shahbaz M. (2012), Migration and tourism demand, Theoretical and Applied Economics, XIX, 39-48.
13. Massidda C., Etzo I. and Piras R. (2015). Migration and inbound tourism: An Italian perspective. Current Issues in Tourism, 18(12): 1152-1171.
14. Massidda C. and Piras R. (2015). Does internal migration affect Italian domestic tourism? A panel data analysis. Tourism Economics, 21(3): 581-600.
15. Massidda C., Etzo I. and Piras R. (2017). The Relationship between immigration and tourism firms. Tourism Economics, 23(8): 1537-1552.
16. Olney W. (2013). Immigration and firms expansion. Journal of Regional Science, 53(1): 142–157.
17. Ottaviano GIP and Peri G (2012). Rethinking the effect of immigration on wages. Journal of the European Economic Association, 10(1): 152–197.
18. Prescott, D., Wilton, D., Dadayli, C., & Dickson, A. (2005), Travel to Canada: the role of Canada's immigrants populations. Applied Economics, 37, 651-663.
19. Seetaram, N., & Dwyer, L. (2009), Immigration and tourism demand in Australia: A panel data analysis. Anatolia: An international Journal of Tourism and Hospitality Research, 20, 212-222.
20. Seetaram, N. (2012a), Immigration and international inbound tourism: Empirical evidence from Australia. Tourism Management, 33, 1535-1543.
21. Seetaram (2012b), Estimating demand elasticities for Australia's international outbound tourism. Tourism Economics, 18, 999-1017.
22. SVIMEZ (2011), 150 anni di statistiche italiane: Nord e Sud, 1861-2011. Il Mulino, Bologna.
23. Tadesse, B., & White, R. (2012), Do immigrants enhance international trade services? The case of US tourism services export. International Journal of Tourism Research, 14, 567-585.
24. Williams A. M., and Hall C. M. (2002). Tourism, migration, circulation and mobility. The contingencies of time and place, in Hall C. M. and Williams A. M. (eds.) Tourism and Migration. New Relationship between Production and Consumption. Dordrecht: Kluwer Academic Publishers.
25. UNWTO (2017) Tourism highlights, World Tourism Organization web site.

Tourism Statistics: development and potential uses

Le Statistiche sul turismo: sviluppi e possibili utilizzi

Fabrizio Antolini

Abstract The paper aims at highlighting the role and importance of tourism statistics in describing tourism sector but also the business cycle. So far, despite WTO recommendations, and the EU regulations on the European statistics on tourism (Reg. 692/2011), there is no standardized Tourism Statistics System (TSS) across countries.

This study systematically and critically examines metadata of tourism statistics for a significant group of countries, aiming at highlighting those methodological issues that are afflicting tourism statistics

Abstract *Il presente lavoro intende evidenziare il ruolo e l'importanza delle statistiche sul turismo nel descrivere il settore del turismo, ma anche l'andamento del ciclo economico. Finora, nonostante le raccomandazioni del WTO e gli appositi regolamenti UE (Reg. 692/2011), non esiste un sistema di statistiche sul turismo (STS) standardizzato per i vari paesi. Questo studio analizza i metadati delle statistiche sul turismo di alcuni Paesi, con l'obiettivo di evidenziare le principali differenze metodologiche esistenti.*

Key words: Tourism statistics; metadata; quality of statistics.

1. "Tourism statistics" and "statistical information about tourism": the NSI and the ONT organizations.

Tourism is connected to nearly all areas of human social activity although, for its contribution to economic growth, it is considered an economic activity itself,

[1] Fabrizio Antolini, Università degli Studi di Teramo; fantolini@unite.it

prevalently concentrated in the services sector. Tourist behaviour depends on several reasons, and tourism is originated from different purposes that are not only the motivation of the people who originate the decision to travel, but also the specific characteristics of the place or country visited (Cunha, 2014). Actually, for a better comprehension of tourism sector and tourist behaviour, researchers needs and subjective preferences suggest to have a detailed and integrated statistical information. It follows the opportunity to build a System of Tourism Statistics (STS) at national and sub national level (OECD, 2011). The construction of a STS should be a part of the National Statistical System (NSS) "providing reliable, consistent and appropriate statistical information on the socio-economic aspects related to tourism, integrated within all the economic and social statistics related to other fields, at different territorial levels (national – or federal, where appropriate - infra national and international)" (UNWTO, 2011).

In the perspective of building a comprehensive STS, it is important to distinguish the statistical information on tourism from tourism statistics, the latter being only a part of the statistical information useful to evaluate tourism sector. While each National Statistical Institute (NSI) has the primary goal to transmit to Eurostat, data on the variables provided in Reg.692/2011 (Annex I and II), differently ONTs should organize a dashboard to evaluate the competitiveness of tourist destinations. In fact: "the National Tourism Observatories (ONT) could be considered as a one stop-stop shop for analysing economic and statistical data provided by different stakeholders so that to create a tourism intelligence centre for increased statistical co-operation with the local authority and the tourist operators" (OECD, 2011, p. 17). At present, in some countries, there are independent ONTs, while in others we find ONTs inside NSIs. Among the 27 EU Countries, 16 have an ONT, but only 10 have an independent ONTs (EC2010).

Moreover, ONTs can be organized in a very flexible way, depending on the needs of national or local stakeholders, thus covering topics that are not assigned to NSI. For instance, tourist operators and policy makers are interested in forecasting tourism figures in the long and short term (nowcasting), while such needs are not generally met by NSI as, according to Reg.692/2011, they do not fall within NSI duties. The availability of a STS creates a benefit also in the interpretation of other phenomena, not strictly linked to the tourism field. For instance, considering the attributes that a short-term indicator must have (Eurostat, 2006), the number of holiday trips can be used as a proxy of the household confidence about their future expectations, as well as the change in the "average length of stay" has proved to be an indirect measures of economic crisis in Italy.

2 The framework of tourism statistics: the past, the present and the future

The shift from the directive EC 95/57 to the EU Regulation 692/2011 cannot be considered as occasional. It represents the answer to the changing nature of tourism behavior with: the growing importance of short trips and same-day visits, the increasing importance of non-rented accommodations or accommodation in smaller establishments, as well as the growing impact of the internet on the tourism industry. Then the production of tourism statistics should be adapted (EC, 2011). For instance, in the new definition, the day visitors are inside the tourism sector, although they are not tourists. The nomenclature of economic activity is more detailed in its application, in particular for the code 55.3 of NACE REV.2 (Camping grounds recreational vehicle parks and trailer parks) . Furthermore, a specific classification of territories is provided (densely, intermediate and thinly areas), as well as it is also possible to classify the local administrative units in coastal and non-coastal. Tourism statistics have, as their primary goal, "to recognized the role of tourism as a tool of development and socioeconomic integration" (EC95/57), so that "monthly data is needed in order to measure the seasonal influences of demand on tourist accommodation capacity and thereby help public authorities and economic operators develop more suitable strategies and policies for improving the seasonal spread of holidays and tourism activities or to enable assessment of the macroeconomic importance of tourism in the economies of the Member States by Satellite Accounts" (EC 692/2011). With the "Agenda for a sustainable and competitive European tourism" (EC, 2007) the right balance between the welfare of tourists, the needs of the natural and cultural environment, linked competitiveness with sustainability were connected. Competitiveness and sustainability, during the years, have become the main concepts that guide the policies on tourism and, indirectly, the production of tourism statistics. While a set of indicators has been found to measures sustainability (EC, 2016), providing a sounded system of tourism information by ETIS indicators, for competitiveness the goal seems to be more complex to achieve. In fact, although some key elements of competitiveness are shared by countries, measuring competitiveness can be more difficult because of a lack of data. In fact, in many countries, the variables needed to build the indicators are not produced. For instance, researchers or politicians do not know the level of price of many services (and goods). Originally the volume of arrivals or of nights spent have been used as a proxy to measure competitiveness but now, with a significant change in mobility (Wang 2000; White and White, 2007) those aggregates cannot be used in the same way . Finally, it is important to point out the attractiveness of a territory in order to highlight its capacity to become a gravitational area, and that is why we need an integrated tourism statistical system.

3 Tourism statistics: operational definitions and different methodology across countries

Tourism statistics cover the internal tourism, in terms of capacity and occupancy of tourism accommodation establishment and national tourism, in terms of tourism demand, that concerns the characteristics of tourism trips and visitors and same day visits. The statistical unit is the visitor. In fact 'tourism' means the activity of visitors taking a trip to a main destination outside their usual environment, for less than one year, for any main purpose, including business, leisure or other personal purpose, other than to be employed by a resident entity in the place visited" (EC, 2011, p. 19). Thus (Table 1), a visitor is a traveller out his/her usual environment, which" identifies the geographical area, though not necessarily a contiguous one, within which an individual conducts his regular life routines" (EC, 2011, p. 19). The history of the definition of "usual environment" and "visitor" starts in the 1963 United Nations Conference (IUOTO, 1963). The conceptual meaning of usual environment (or, conversely, unusual environment) is "as residential place or as a part of physical space that can be defined in relation to people's individual experience and its routine activities" (Govers, Van Hecke & Cabus, 2008). Thus, the detection of "visitor" (and, consequently, "usual environment") is obtained excluding those people that every day or every week go from their home to their workplace or study-place (Cunha, 2014). The relationship between usual environment, space and place (Tuan, 1974; Cresswell, 2004) is crucial for demand side tourism statistics, considering the usual environment not as a spatial continuum, but as a collection of places with different surface areas.

Table 1: Action for travellers and visitors

Action	*Traveller*	*Visitor*
Movements	Between geographical locations	Destination outside his/her usual environment
Purposes	Any	Any purpose other than to be employed by a resident entity in the placed visited
Duration	Any	Less than year

A number of issues emerge from UN (UN, 2016, p. 19):

- the usual environment concept as a respondent category: "it introduces subjectivity, confusion and unsystematic variation in reported travel activity":
- the travel distance criterion: "introduces a false appearance of objectivity by masking subjective differences in respondents' abilities to recall and accurately measure travel distance thereby contributing to increased uncontrolled variance and volatility in subsequent data";
- crossing an administrative boundary: "could potentially provide an arbitrary gross standard as a 'minimum basis of comparison' for the purposes of international reporting, cumulative statistics and analysis".

Table 2 provides the operational definition of unusual environment for a representative set of EU countries. That table is partially different from the one (dated 2011) included in the publication UN (2016, p. 22) as it comprises also Italy and other EU countries. Data for the compilation of Table 2 are derived from Eurostat metadata and the examination of the questionnaires available at the Eurostat web pages.

It is clear that any combination of the criteria illustrated shifts the classification of visitors/non-visitors. For instance, traveller A who makes a unusual trip (suppose less than one a week) that exceeds 30 Km (hypothetical limit), with a duration of more than 2 hours (hypothetical limits), but does not pass any administrative border could not be considered a visitor. The second critical element regards how to recognize the "tourism trip" that requires "the stay in the place visited" but does not necessarily involve any "overnight", and that is why tourism visit is conceptually different from tourist. The notion of stay, implies a stop, in fact entering into a geographical area without stopping there, does not qualify, as a visit to that area, but a common limit to identify the minimum duration has not been found yet.

Looking at internal tourism data provided by the census on capacity and occupancy of tourist accommodation establishments (supply side source), the process design is different across countries. According to the Eurostat metadata, is possible (Table 3) to cluster the countries as those with a:

- dedicated and autonomous census survey (Italy, Portugal Malta);
- census survey from business register (Germany; Croatia; Finland);
- stratified sample survey (France and Spain).

The distinction between "autonomous" census survey and census survey from business register can have important effect on the quality of the data (Santoro and Petrei, 2015), and on the measure proposed. In Italy, for the year 2015, the Business Register ASIA reports 27330 Local Units of hotels (ATECO code 55.1, overall 53106), whilst the tourism administrative archives count 33199 establishments (overall 167718). Also comparing the number of overnight stays from supply and demand side sources, there are deep inconsistencies as shown for example in Guizzardi and Bernini (2012).

Table 2*: Operational definitions of unusual environment*

Country	*Distance threshold (km)*	*Administrative borders*	*Respondents' self-evaluation*	*Frequency of visit*	*Duration threshold (hours)*
Austria		X	X	Less than twice per month (*)	
Finland	30			Less than once a week	3 (+)
France	100 (*)		X		
Germany		X	X		

Italy	X	X	Less than once a week	
Malta	X (country)			
Portugal	X (*)	X		3 (*)
Spain	X	X		

(*) one-day visit; (+) trips abroad; Source: Eurostat metadata

Table 3: *Capacity of accommodation establishments*

Country	*Regional Coverage*	*NACE Rev. 2 Code*	*Source of data*	*Frequency of data collection*	*Time coverage*
Italy	LAU2 (municipality)	55.1 (Hotels and similar accommodation)	Census survey (via local authorities or tourism bodies)	Annually	1956-2015
		55.2 (Holiday and other short-stay accommodation)	Census survey (via local authorities or tourism bodies)	Annually	1963-2015
		55.3 (Camping grounds, etc.)	Census survey (via local authorities or tourism bodies)	Annually	1963-2015
Germany	NUTS 0 (federal territory)	55.1 (Hotels and similar accommodation)	Census survey (data are collected by business register)	Monthly	1992 >>
	NUTS 1 (federal states)	55.2 (Holiday and other short-stay accommodation)	Census survey (data are collected by business register)	Monthly	1992 >>
	NUTS 2 (districts)				
	NUTS 3 (rural and urban districts)	55.3 (Camping grounds, etc.)	Census survey (data are collected by business register)	Monthly	1992 >>
	LAU2 (municipality)				
France	NUTS 3	55.1 (Hotels and similar accommodation)	Sample survey (from business registers)	Monthly	2000 >>
		55.2 (Holiday and other short-stay accommodation)	Sample survey (from business registers)	Monthly	2011 >>
		55.3 (Camping grounds, etc.)	Sample survey (from business registers)	Monthly	2003 >>

Spain	NUTS 2 (17 Autonomous Communities)	55.1 (Hotels and similar accommodation)	Sample survey (stratified population; from administrative registers)	Monthly	1964 >>
	2 Autonomous Cities (just for NACE 55.1)				
	NUTS 3 (52 provinces)	55.2 (Holiday and other short-stay accommodation)	Sample survey (stratified population; from administrative registers)	Monthly	Holiday Dwellings: 2000 >>
					Rural Tourism Accommodations: 2001 >>
	LAU 2 (tourist areas and tourist sites)				Hostels: 2014 >>
		55.3 (Camping grounds, etc.)	Census survey	Monthly	1964 >>
Portugal	LAU2 (municipality)	55.1 (Hotels and similar accommodation)	Census survey	Monthly	2000 >> (with a break in 2008)
		55.2 (Holiday and other short-stay accommodation)	Census survey	Monthly	2000 >> (with a break in 2008)
		55.3 (Camping grounds, etc.)	Census survey	Monthly	2000 >> (with a break in 2008)
		55.3 (Camping grounds, etc.)	Not avalaible	Not avalaible	Not avalaible

Source: Eurostat Metadata

4. Conclusions

The present analysis reveals the implementation of different process designs, different data collection modes and different operational definitions for two types of tourism data: the ones collected from the supply side ("capacity and occupancy data") and the ones from the demand side (households surveys). As illustrated in the article, the population frames of supply side data can be administrative, statistical or mixed (Business Register). Differently, for demand side data, the main operational problem is the definition of "usual environment". The definition of "usual environment" is crucial even in the integration of official statistics with new data sources like mobile phone data. In fact, the use of mobile phone data should respect the basic definitions of usual environment given by the international regulations (Raun & Ahas, 2016).

References

1. Cresswell, T. (2004) Place, A Short Introduction. Malden, USA: Blackwell
2. Cunha, L. (2014) The Definition and Scope of Tourism: a Necessary Inquiry. Cogitur, Journal of Tourism Studies, [S.l.], n. 5, may 2014. http://revistas.ulusofona.pt/index.php/jts/article/view/4426:
3. EC95/57 On the collection of statistical information in the field of tourism
4. EC (2007). Agenda for a sustainable and competitive European tourism, http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52007DC0621 (accessed 2018.04.16)
5. EC (2010) European Tourism Stakeholders' Conference, 14-15 April, Madrid
6. EC (2011). Regulation (EU) No 692/2011 Of the European Parliament and of the Council of 6 July 2011. Official Journal of the European Union, L192/17.
7. Govers, R.,Van Hecke E., & Cabus, P. (2008). Delineating Tourism Defining the Usual Environment. Annals of Tourism Research, 35(4), 1055-1073
8. Guizardi, A., Bernini (2012). Measuring underreporting in accommodation statistics: evidence from Italy. Currentt Issues in Tourism, 15(6), 597-602
9. IUOTO (1963). La Conférence des Nations Unies sur le Tourisme et les Vojages Internationaux. Aout-Septembre. Résolutions et recommandotions. Paris: Publication de L'Union Internationale des Organismes Officiels de Tourisme.
10. Lam, C., McKercher B. (2013). "The tourism data gap: The utility of official touirism information for the hospitality and tourism industry", Tourism management perspectives, 6 (2013), 82-94.
11. Masiero, L. (2016). International tourism statistics: UNWTO benchmark and cross-country comparison of definitions and sampling issues. Statistics and TSA: Issue Paper Series.
12. OECD (2011). OECD Studies on Tourism: Italy: Review of Issues and Policies, OECD Publishing. http://dx.doi.org/10.1787/9789264114258-en
13. Petrei , F. & Santoro M.R (2015), " Tourism Statistics and the business Register in Italy: A Comparative analysis and future outlooks Integrations, in Journal of Statistical Application, Vol.3, No. 1-15.
14. Raun, J. & Ahas, R. (2016). Defining usual environment with mobile tracking data. 14th Global Forum on Tourism Statistics, 23-25 November 2016, Venice, Italy.
15. Tuan, Y.-F. (1974) Space and Place: Humanistic Perspective. Progress in Geography, 6:211–225.
16. Wang, N. (2000). Tourism and modernity: A sociological analysis. Oxford, UK: Pergamon Press.
17. White N. R. & White, P. B. (2007) Home and away: Tourists in a Connected World. Annals of Tourism Research, Volume 34, Issue 1, January 2007, 88-104.
18. UNWTO (2011) The System of tourism statistics as the foundation for a tourism information system, Madrid, http://statistics.unwto.org/content/sts_nss

Tourism attractiveness in Italy. Some empirical evidence comparing origin-destination domestic tourism flows.

L'attrattività turistica in Italia. Alcune evidenze empiriche confrontando metodi per flussi origine-destinazione.

Francesca Giambona, Emanuela Dreassi, and Alessandro Magrini

Abstract This paper aims to model tourism attractiveness for the twenty Italian regions by using origin-destination flows. To this purpose we consider the Italian domestic tourism flows and a wide range of determinants within the theoretical framework referring to the destination competitiveness theories. Using the same set of covariates (selected from Istat and Enac), we propose a comparison between the Gravity model (commonly used in tourism research) and the Bradley-Terry modelling approach (to date not yet used for tourism). Using different model specifications different empirical findings are obtained. Strengths and weaknesses of both modelling approaches will be analysed and explained.

Abstract *Questo lavoro intende modellare l'attrattiva turistica per le venti regioni italiane utilizzando i flussi turistici interregionali origine-destinazione. A tal fine sono stati considerati i flussi turistici nazionali italiani e una vasta gamma di determinanti all'interno del quadro teorico che si riferisce alle teorie della competitivitá di destinazione. Utilizzando lo stesso set di covariate (selezionato da Istat ed Enac), proponiamo un confronto tra il modello Gravity (comunemente usato in ambito turistico) e il modello Bradley-Terry (fino ad oggi non ancora utilizzato per il l'analisi dell'attrattivit turistica). Usando diverse specificazioni teoriche si otterranno risultati empirici diversi. I punti di forza e di debolezza di entrambi gli approcci di modellazione verranno analizzati e spiegati.*

Key words: Tourism attractiveness, Domestic tourism flows, Bradley-Terry model, Gravity model.

Francesca Giambona
Dip. di Statistica, Informatica, Applicazioni (DiSIA), e-mail: emanuela.dreassi@unifi.it

Emanuela Dreassi
Dip. di Statistica, Informatica, Applicazioni (DiSIA), e-mail: francesca.giambona@unifi.it

Alessandro Magrini
Dip. di Statistica, Informatica, Applicazioni (DiSIA), e-mail: alessandro.magrini@unifi.it

1 Introduction

Domestic tourism represents about 80% of internal (inbound and domestic) tourism consumption in the OECD area [1]. The determinants of both domestic and international tourism flows have been extensively studied and results are documented by empirical literature [2]. As far as local development is concerned, domestic tourism is a key driver also in a mature country like Italy. The recent strategic plan for tourism of the Italian government [3] explicitly includes domestic tourism among the targets and interventions related to the reinforcement of the Italian brand's positioning and attractiveness. A deepen understanding of domestic tourism at subnational level is a condition for a more affective national policies planning.
Tourism attractiveness is an attribute of tourism competitiveness as competitiveness for a destination is ”about the ability of the place to optimize its attractiveness for residents and non-residents, to deliver quality, innovative, and attractive (e.g. providing good value for money) tourism services to consumers and to gain market shares on the domestic and global market places, while ensuring that the available resources supporting tourism are used efficiently and in a sustainable way” [4] ; consequently, attractiveness has a key role in competitiveness as the increase of tourist flows (and more general tourism demand) is one target of competitiveness. Tourism attractiveness, in this paper, is analysed by using a model-comparison approach, by comparing the widely used gravity model (GM) respect to the Bradley-Terry model (BTM) not yet used in tourism issue. Both, BTM and GM are used to modelling origin-destination flows, but with different model specification and results interpretation. Gravity models have often been used to analyse tourism flows [5, 6, 7] with extensions to panel data [8, 9, 10] and spatial modelling [11]. Gravity models consider that bilateral flows between two countries are directly proportional to the countries economic masses and inversely proportional to the distance between them. However, the availability of origin-destination flows also allows the use of alternative models. The best-known model in the pairwise framework is the BTM [12]. It has been widely used in empirical applications when the structure of the data and the research questions have to be analysed using a pairwise comparison model [13]. In this paper we compare these modelling approaches in order to analyse the determinants of regional tourism attractiveness and to assess differences between GM and BTM.

2 Data and Models

Following the OECD definition domestic tourism is the tourism of resident visitors within the economic territory of the country of reference. In order to make a comparison between models used for O-D flows we use as dependent variable the Italian domestic flows between regions (NUTS2) available for the year 2016. The choice of covariates that could affect tourism attractiveness refer mainly on the push-pull factors theory. Push factors are related to individual motivations and even perceptions

of the destinations quality of life and image [14, 15, 16], so they are generally represented by individual level data. Pull factors are characteristics of the destination that arouse the desire for travel in the potential tourist and attract tourists to specific destinations. (generally include destination variables like natural attractions, cultural resources, recreational activities, and so on [17, 18]). Covariates are chosen on the basis of the theoretical framework by [19] who identified the main dimensions of competitiveness. In particular, we defined the following five dimensions within the supply side model: environment and scenery, heritage and culture, general infrastructures, tourism activities and situational conditions. For each dimensions several variable have been taken into account, but the final subset of explanatory variable is selected through a stepwise procedure (forward selection), considering the reduction of residual standard deviation.
Gravity model and the Bradley-Terry model are statistical tools to analyse flows when origin-destination matrix is available although they deal with data in a very different way. The basic differences between the two models are mainly due to the different belonging to two groups of different models, as GM belongs to the models with complementarity, whilst BTM refers to the models with competition [20]. In the following a more detailed methodological explanation will be provided.

2.1 Bradley-Terry model

The standard Bradley-Terry model [12, 22] considers the regions as players (e.g., i and j) with different abilities. If the ability of i (for $i = 1,\dots,M$) is higher than the ability of j (for all j), the number of times that i beats j is expected to be higher than the number of times j beats i, that is the number of tourists who prefer the region i coming from the area j. The model specifies the probability that in a pairwise comparison between i and j (for j that range from 1 to $M-1$) tourists prefer the region i to j, as follows:

$$P(i \text{ beats } j) = \pi_{ij} = \frac{\alpha_i}{\alpha_i + \alpha_j} \tag{1}$$

where α_i and α_j represent the ability parameters that measure the intensity of an unobservable (latent) trait in the two players. In the analysis of tourism the ability parameters are the attractiveness parameter of the competing regions. By expressing the model in the logit form, equation 1 becomes

$$\text{logit}(\pi_{ij}) = \lambda_i - \lambda_j \tag{2}$$

where $\lambda_i = \log \alpha_i$ and $\lambda_j = \log \alpha_j$ may be fixed or random parameter.

The basic model allows to make generalisations in several directions (Turner and Firth, 2012), for example, to specify ability as a function of covariates. If player covariates ($r = 1,\dots,p$) are used to explain differences in players' abilities, the parameters λ_i and λ_j are related to the covariates by a linear predictor

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir} + U_i \tag{3}$$

where U_i (and U_j) are normally distributed random terms. Following, equation 2 becomes

$$\text{logit}(\pi_i) = \sum_{r=1}^{p} \beta_r (x_{ir} - x_{jr}) + U_i - U_j \tag{4}$$

In the framework of the Bradley-Terry models, differences in attractiveness parameters (as measured by a fixed or random parameter shared by all pairs in which the same region is involved) are the factors that lead tourists to prefer one region over another. We call them ”ability” of the region to attract tourists.

2.2 Gravity model

Let Y_{ij} the flow from the i-th origin region to the j-th destination region. We consider a Poisson Generalized Linear Mixed Model (GLMM). We consider a set of fixed effects and a set of random intercepts and slopes (one for each destination region). The slopes relate to the logarithm of the distance between the centroid of the origin and destination regions. This, represents a gravity model [21]. The set of random slopes (exponentialized) describes the multiplicative effect of the logarithm of the distance for each destination region on the logarithm of the flow. Fixed effects are considered accordingly with covariates included o the Bradley-Terry model. However, a weakly linear relation is suggested for these and the logarithm of the flows. So, we consider $Y_{ij} \sim Poisson\,(\mu_{ij})$, then the linear predictor

$$\log \mu_{ij} = V_j + W_j \log(\text{distance}_{ij}) + \beta_1 x_i + \sum_{r=2}^{p} \beta_r x_{jr} \tag{5}$$

where V_j and W_j represent, respectively, sets of random intercepts and slopes (each referred to a destination j). A set of $p-1$ covariates, i.e. x_{jr}, are referred to the destination j, another, i.e. x_i, to the origin i. We estimate the model considering or not the internal regional flows. The classical Gravity model include these latter, but to compare results with the BTM we decided to estimate also the GM without the internal regional flows. The distance for internal regional flows are settled considering the radios of the circle with the same area of the region.

3 Results

The final subset of explanatory variable is selected through a stepwise procedure (forward selection), considering the reduction of residual standard deviation for the BTM. The analysis considers the twenty regions (level-2 units) and compares the

BTM and GM models. Random effects for the BTM have a standard deviation equal to 0.23788. For GM the estimate of standard deviation is 10.5737 for the intercepts and 0.8085 for the slopes. When we consider also the within flows we obtain a standard deviation of 12.9196 for intercepts and 0.9715 for slopes. Covariates are representative of local resources (*VILLAGES*= landscapes with historical villages (counts/1 000 km2) and *CULTENDOW*= cultural endowment index (counts/100 km2)), the tourism supply (*BEDS*= number of bed places over number of inhabitants) and transport services (*LOWCOST*=percentage of passengers of low-cost flights). Furthermore, we account for regional population size (*POP*=resident population (thousands) adjusted for outgoing flows). The signs of the covariates are as expected except for the *VILLAGES* variable. Its negative sign can be interpreted as follows: what matters is the presence of attractive historical villages although they are sparse over the territory. However, this result can be also the consequences of a heterogeneity across regions that is not taken into account trough a fixed effect. Table 1 reports the fixed effects estimates from both models. For the gravity model we have considered also the possibility of including internal regional flows. In GM the linear relations between the included covariates and the logarithm of the mean of the flows, are very weak, while in BTM are stronger. Figure 1 displays GM fitted flows (without internal regional flows) random intercept and slopes respect to the distance. Friuli and Trentino are more affected to distance than (on the contrary) Sicily or Molise.

Figure 2 describes the intercept random effects for the two models; both models do not consider internal regional flows. In is interesting the different positioning of Italian regions referring to the fitted values of GM and BTM: i) regions with concordant values as Molise and Lazio (low, low) or Trentino (high, high) or Toscana (medium, medium), ii) regions with different values as Piemonte or Friuli (low, high) or Marche (medium-low, high). Considering the standard GM (i.e. with internal flows) the discrepancy between the fitted flows using GM and BTM is more marked (Figure 3).

4 Conclusions

Although domestic tourism is only a part of total tourism flows, it remains a key driver of competitiveness in Italy. Recent economic crises have revealed a weakness of domestic tourism, which has undergone a period of stagnation and decline, recovering only since 2013, and only in some regions. From a destination point of view, if tourists are likely to find attractive and unattractive a specific destination is a key topic, as this is the key to improving destination performance and assisting, in this case, the Italian tourism industry to regain its attractiveness (competitiveness). In this contribution, a measurement of domestic attractiveness for the twenty Italian regions has been proposed based on an analysis of regional origin-destination tourism flows, by comparing two modelling approaches: the well-known Gravity model and the less-known (especially in tourism) Bradley-Terry model. The

Fig. 1 Random intercepts and slopes from the Gravity model without internal regional flows (the Poisson linear mixed model)

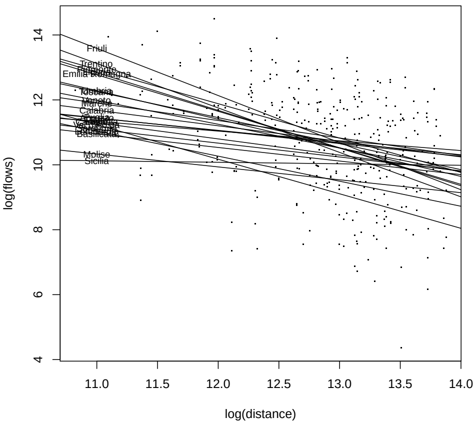
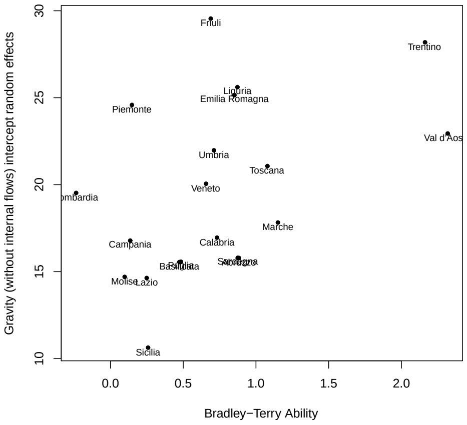


Fig. 2 Random intercepts from Gravity model *versus* the ability parameters from the Bradley-Terry model



Bradley-Terry model found new application on tourism system evaluation, and compared to the usual Gravity model, the Bradley-Terry model changes drastically point of view as it specify attractiveness in a competition point of view. There are many differences between the two models. Mainly, the gravity model, compared to the Bradley-Terry model: *i*) models a flow and not a probability, as it belongs to the category of models with complementarity (nor competition), *ii*) takes into account the distance between origin and destination, *iii*) by including slope random effects for the distance between origin and destination, evaluates for each destination the decay effect of attraction of tourism respect to the distance, *iv*) it can consider also within flows (flows of tourists that move inside each region), *v*) includes covariates,

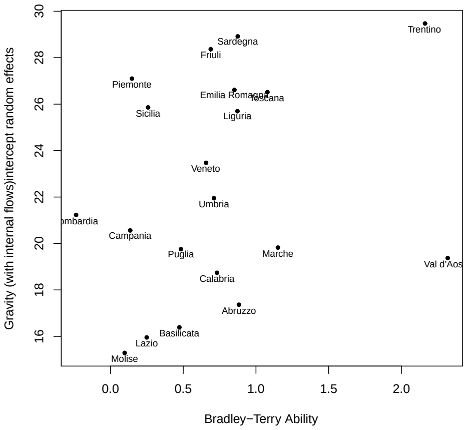


Fig. 3 Random intercepts from Gravity model considering regional within flows *versus* the ability parameters from the Bradley-Terry model

Table 1 Fixed effects from Bradley-Terry and Gravity models

covariates	estimate	s.e.	p-value
Bradley-Terry Model			
population	-6.921e-05	2.840e-05	0.01481
beds	5.709	0.6812	< 2e-16
villages	-0.05768	0.02011	0.00413
cultendow	2.299e-03	9.858e-04	0.01968
lowcost	7.762e-03	4.787e-03	0.10494
Gravity Model without internal regional flows			
internal regional flows			
intercept	10.31	0.2641	< 2e-16
population	3.792e-04	4.226e-05	< 2e-16
beds	7.769e-01	1.889e-03	< 2e-16
villages	-1.798e-01	7.218e-05	< 2e-16
cultendow	2.668e-03	2.724e-06	< 2e-16
lowcost	2.242e-03	1.313e-05	< 2e-16
Gravity Model with internal regional flows			
internal regional flows			
intercept	1.031e+01	2.707e-01	<2e-16
population	3.815e-04	3.941e-05	<2e-16
beds	-1.499e-01	1.686e-03	<2e-16
villages	-1.585e-01	6.448e-05	<2e-16
cultendow	1.783e-03	2.480e-06	<2e-16
lowcost	2.008e-05	1.158e-05	0.083

referring them to the origin or destination, accordingly to their meaning. The use of GM or BTM is not a trivial issue, indeed empirical findings highlight that regions are more or less attractive on the basis of model specification, even if for some regions, i.e. the better and the worst in attractiveness terms, results are very similar.

The availability of a more detailed tourist flows matrix (at the provincial level, for example) could be useful to investigate more deeply the determinants of tourism attractiveness. However, at present, only provincial information is recorded only for the destination, i.e. we would have only region-province tourism flows.

References

1. OECD: OECD Studies on Tourism, Italy, Review of Issues and Policies, OECD Publishing, (2016)
2. Song, H., Dwyer, L., Li, G., Cao, Z.: Tourism economics research: A review and assessment, Annals of Tourism Research, **39**, 1653–1682 (2012)
3. MiBACT: PST 2017-2022. Strategic plan for tourism, Ministero dei beni e delle attivit culturali e del turismo, Roma (2016)
4. Dupeyras, A., MacCallum, N.: Indicators for Measuring Competitiveness in Tourism: A Guidance Document, OECD Tourism Papers (2013)
5. Durden, G. C., Silberman, J.: The determinants of Florida tourist flows: A gravity model approach. Review of Regional Studies, **5**, 31–41 (1975)
6. Vietze, C.: Cultural effects on inbound tourism into the USA: A gravity approach, Tourism Economics. Tourism Economics, **18**, 121–138 (2012)
7. Morley, C.A., Rossell, J., Santana-Gallego, M..: Gravity models for tourism demand: theory and use. Annals of Tourism Research, **48**, 1–10 (2014)
8. Eilat, Y., Einav, L., Santana-Gallego, M.: The determinants of international tourism: A three-dimensional panel data analysis. Applied Economics, **36**, 1351–1328 (2004)
9. Keum, K.: Tourism flows and trade theory: A panel data analysis with the gravity model. The Annals of Regional Science, **44**, 541–557 (2010)
10. Massidda, C., Etzo, I.: The determinants of Italian domestic tourism: a panel data analysis. Tourism management, **33**, 603–610 (2012)
11. Marrocu, E., Paci, R.: Different tourists to different destinations. Evidence from spatial interaction models. Tourism management, **39**, 71–83 (2013)
12. Bradley, R.A., Terry, M.E.: The rank analysis of incmplete block designs. The method of paired comparisons. Biometrika, **39**, 324–345 (1952)
13. Varin, C., Cattelan, M., Firth, D.: Statistical modelling of citation exchange between statistics journals (with discussion). Journal of the Royal Statistical Society A, **179**, 1–63 (2016)
14. Crompton, J.: Motivations of pleasure vacations. Annals of Tourism Research, **6**, 408–424 (1979)
15. Kim, S., Lee, C.: Push and pull relationships. Annals of Tourism Research, **29**, 257–260 (2002)
16. Kwang-Hoon, L.: The conceptualisation of country attractiveness; a review of research. International review of administrative sciences, **82(4)**, 807–826 (2016)
17. Baloglu, S., Muzaffer, U.: Market segments of push and pull motivations: a canonical correlation approach. International Journal of Contemporary Hospitality Management, **8(3)**, 32–38 (1997)
18. Kim, S., Lee, C., Klenosky, D.: The influence of push and pull factors at Korean national parks. Tourism Management, **24(2)**, 169–180 (2004)
19. Dwyer, L., Chulwon, K.: Destination competitiveness: Determinants and Indicators. Current Issues in Tourism, **6**, 369–413.
20. Jockmans, K.: Modified-likelihood estimation of the S-model, Sciences Po Economics Discussion Papers 2016-01, Sciences Po Departement of Economics, **39**, 1–19 (2016)
21. Bailey,T.C., Gatrell, A.C.: Interactive spatial data analysis, Prentice Hall, London (1995)
22. Agresti, A.: Categorical data analysis, Wiley, New York (2002)
23. Turner, H., Firth, D.: BradleyTerrymodels in R: The BradleyTerry2package, Journal of StatisticalSoftware, **48**, 1–21 (2012)

What’s Happening in Africa

Environmental shocks and internal migration in Tanzania

Shock ambientali e migrazioni interne in Tanzania

Maria Francesca Marino, Alessandra Petrucci, and Elena Pirani

Abstract There is substantial evidence that climate has been changed in recent decades, and this trend will intensify in the near future. Climate changes are expected to exert additional pressure on areas which mainly rely on agricultural activities and, as result, act as push factors for population movements. By implementing a gravity-type model, in this study we aim at examining migration flows across Tanzanian districts. We analyze the potential impact of climate changes on inter-district migration in Tanzania, while controlling for socio-economic and geo-physical features for both the district of origin and destination.

Abstract Numerosi studi dimostrano come il clima sia cambiato negli ultimi decenni e come questa tendenza si intensificherà nel prossimo futuro. Si prevede che tali cambiamenti avranno ripercussioni soprattutto su aree la cui attività economica è prevalentemente basata sull'agricoltura e, in tali aree, agiranno come fattore di spinta per le migrazioni. Implementando un modello gravitazionale, in questo studio si propone un'analisi delle migrazioni interne in Tanzania. In particolare, si analizza l'effetto di shock ambientali sui flussi migratori inter-distrettuali in Tanzania, tenendo conto delle diverse caratteristiche socio-economiche e geofisiche delle aree interessate dal processo migratorio, ovvero quella di partenza e quella di arrivo.

Key words: Climate change, gravity-type models, population flows, socio-economic conditions.

Maria Francesca Marino
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, e-mail: mariafrancesca.marino@unifi.it

Alessandra Petrucci
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, e-mail: alessandra.petrucci@unifi.it

Elena Pirani
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, e-mail: elena.pirani@unifi.it

1 Introduction

Recent studies have clearly shown that climate has been changed in recent decades and that this trend will not only persist but will also intensify in the near future. Climate changes are expected to have an impact especially on those areas whose economy mainly relies on agricultural activities; here, they are expected to exert additional pressure on the economy and, as a result, act as push factors for population movements. Despite the growing acknowledgement that the environment is becoming increasingly important in explaining large-scale movements of migrants, some of whom have been described as environmental refugees [5, 6, 8], understanding the association between migrations flows and changes in climate conditions remains a challenging task, especially in developing countries. Most of the previous studies are based on a micro approach [10, 7]. That is, they focus on the individual migrating unit (person, group or household) and study those factors that influence the decision of the potential migrant to remain in the current location or to move. In contrast, when considering a macro perspective, the interest is in analyzing aggregated migration flows, and understanding the relation between migration and objectively-determined macro variables, such as population sizes, economic descriptors of the context, and/or environmental conditions.

In this study, we follow the macro approach and focus on migration flows across Tanzania districts. This country represents an interesting case study due to its diversity in terms of agro-climatic conditions and ecological zones, and for the richness of available data on climate conditions. In this respect, our aim is that of analyzing the potential impact of climate changes on inter-districts migration in Tanzania via a gravity-type model. This also allows us to properly control for socio-economic and geo-physical features of both the district of origin and destination.

2 Tanzania National Panel Survey

Tanzania National Panel Survey (TNZPS) is a longitudinal survey carried out in Tanzania between October 2008 and January 2016. The survey collects information on a wide range of topics, including agricultural production, non-farm income generating activities, consumption expenditures, and other socio-economic features, with the aim of monitoring poverty dynamics in the country. The survey is made up by three different questionnaires: (*i*) the household questionnaire, which collects information on households economic conditions (expenditures, loans and credits, crimes and justice, etc...) and on education, labor, and health conditions of household members; (*ii*) the agriculture questionnaire, which collects information on household's agricultural activities in terms of production and sales; (*iii*) the community questionnaire, which provides information on physical and economic infrastructure and events in surveyed communities. The first wave of the survey (October 2008 - September 2009) provides information on $3,265$ households; the second (October 2010 - November 2011) and the third (October 2012 - November 2013) provide information on $3,924$ and $5,010$ households, respectively. At the last wave (October 2014 - January 2016), $3,352$ households were involved in the study. In the following, we will focus on data from the household questionnaire coming from the third wave of the survey, that is, that providing the highest sample size.

Combining information on the household's district of residence with the answers provided by the household's head to items "For how many years have you lived in this community?" and "From which district did you move?", we derive information on: (*i*) the migration status of the household (the household migrated if the household's head moved from one district to another in the 5 years prior to the interview); (*ii*) the district of origin and of destination of the household. By aggregating these information at the district level, we are able to build a migration variable, M_{ij}, counting the number of households migrating between district i and j during the five years prior to the interview, with $i = 1,\dots,129$, $j = 1,\dots,129$, and $j \neq i$.

3 A gravity-type model for migration flows across Tanzanian districts

The use of gravity models represents a well-established practice in the economic literature. Recently, thanks to the enhanced availability of migration data, this class of models has attracted researcher's interest also for the analysis of migration flows. The traditional gravity model (e.g. [11] , so-called by analogy with Newtons gravity law) is defined as

$$M_{ij} = \frac{P_i P_j}{d_{ij}},$$

where, as before, M_{ij} denotes the number of recorded migrants between area i and j, P_i and P_j the corresponding population size, and d_{ij} denotes the distance between the centroids of i and j. This equation has the nice property that, taking the logarithm of both sides, we end up with a multiple linear regression model, which is simply to fit and to interpret. However, the restrictive assumptions upon which the log-normal model is built have led researchers to consider more elaborated gravity-type models. Two main extensions have been explored: (*i*) migration flows are assumed to follow a Poisson distribution, rather than a log-normal one; (*ii*) together with the origin and destination populations and the corresponding distance, additional environmental, demographic, and socio-economic factors are included in the model with the aim of providing a better description of the phenomenon under investigation. See e.g. [4, 2, 3]. In this framework, migration flows are modeled according to the following regression model:

$$\log M_{ij} = \beta_0 + \beta_1 P_i + \beta_2 P_j + \beta_3 d_{ij} + x_i'\gamma_1 + x_j\gamma_2 + w_i'\phi_1 + w_j'\phi_2,$$

where x_i and x_j denote the vector of environmental covariates for the area of origin and destination, respectively, and γ_1 and γ_2 the corresponding vectors of parameters. Here we focus on three environmental shock variables: for each district, we consider the percentage of households in the survey indicating to have been affected by (*i*) drought or floods, (*ii*) crop disease or pests, (*iii*) severe water shortage, on the total number of households. These events have been cited as negative important events for a relevant part of the Tanzanian population in the period 2007/08 – 2012/13, with a substantial variation across districts (Table 1).

On the other hand, w_i and w_j denote the vector of socio-economic covariates, with ϕ_1 and ϕ_2 summarizing the corresponding effects on the (transformed) response M_{ij}. Given the scarcity of information on contextual socio-economic characteristics at district level from external databases, we decided to derive such information directly from the TNZSP survey. Based on previous theoretical and empirical literature [9], we built some indicators describing the demographic structure, the level of social and economic development and urbanization, and some geo-physical features of each district. In preliminary analyses, we tested a large set of indicators. At end, the variables included in the model are those listed and described in Table 1, together with their distribution across Tanzanian districts.

Table 1: Gravity-type model for migration flows: variables description

Variable	**Description**	**Mean**	**Std dev.**	**Min**	**Max**
Environmental variables					
Drought or floods	n. of household indicating to have been affected by drought or floods in the last 5 years	22.5	16.5	0	78.1
Crop disease or pests	n. of household indicating to have been affected by crop disease or crop pests in the last 5 years	13.0	14.4	0	87.9
Severe water shortage	n. of household indicating to have been affected by severe water shortage in the last 5 years	20.7	11.7	0	63.6
Demographic and socio-economic variables					
Cultivated area	% under agriculture in the area	27.4	15.9	0.8	84
Mean age	Mean age of the population	23.3	3.1	17	33.5
Urbanization	% of persons living in urbanized areas	22.5	28.4	0	100
Education	% of people with secondary of higher education	16.6	15.9	0	74.9
Tenure status	% of households of property	74.5	17.2	30.6	100

4 Model results

To evaluate the role of environmental shocks in internal migration, we estimated a set of models. First, we considered the basic gravity model (Model 1) including only the population size of the district of origin and of destination and the distance between them. Second, we added to Model 1 the variables associated to perceived environmental shocks in the district of origin and of destination (Model 2). Finally, we included also those variables describing the socio-economic context of the districts involved in the migration (Model 3). Model parameter estimates are reported in Table 2. In the last lines of the table, we also report the value of the log-likelihood function, the deviance, and the AIC index [1] of all three models under investigation to be used for model comparison.

Table 2: Estimates, standard errors, and p-values of model parameters under different model specifications

	Mod 1			Mod 2			Mod 3		
	Est.	Std Er.	p-value	Est.	Std Er.	p-value	Est.	Std Er.	p-value
Intercept	0.554	0.046	0.000	0.673	0.061	0.000	0.642	0.394	0.103
$\text{Pop}_i/1000$	0.002	0.000	0.000	-0.001	0.001	0.046	-0.002	0.001	0.025
$\text{Pop}_j/1000$	0.002	0.000	0.000	0.003	0.001	0.000	0.002	0.001	0.006
d_{ij}	-0.244	0.016	0.000	-0.238	0.016	0.000	-0.235	0.016	0.000
Crop disease or pests$_i$				-0.017	0.005	0.001	0.020	0.005	0.000
Crop disease or pests$_j$				-0.013	0.005	0.009	0.015	0.006	0.006
Drought or floods$_i$				0.020	0.005	0.000	-0.010	0.006	0.092
Drought or floods$_j$				0.003	0.005	0.605	-0.010	0.006	0.095
Water shortage$_i$				0.024	0.004	0.000	0.020	0.005	0.000
Water shortage$_j$				0.005	0.004	0.222	0.004	0.005	0.478
Urbanization$_i$							0.002	0.001	0.097
Urbanization$_j$							0.006	0.001	0.000
Cultivated area$_i$							-0.005	0.002	0.002
Cultivated area$_j$							0.002	0.002	0.266
Education$_i$							0.000	0.002	0.908
Education$_j$							0.007	0.002	0.001
Tenure status$_i$							0.000	0.002	0.845
Tenure status$_j$							0.009	0.002	0.000
Mean Age$_i$							-0.007	0.010	0.484
Mean Age$_j$							-0.036	0.012	0.003
Deviance	975.774			931.291			856.831		
AIC	3049.133			3016.651			2962.190		

By looking at the last three lines of Table 2, we may observe that the optimal specification includes in the linear predictor both environmental and socio-economic factors. In particular, by comparing the deviance of Model 2 with that of Model 1, we obtain a Likelihood Ratio Test (LRT) equal to 44.48 with 6 degrees of freedom; the low p-value (< 0.001) associated to such a statistic leads us to prefer the more complex model specification (i.e. Model 2). Similarly, when comparing Model 3 with Model 2, we obtain a LRT equal to 74.46 with 10 degrees of freedom. Once again, the low p-value (< 0.001) of such a test statistic leads us to retain the more complex model specification which includes in the linear predictor the population size of the district of origin and destination, the distance between then, the environmental variables, as well as the socio-economic ones. Identical conclusions can be drawn when looking at the AIC values reported in the last line of Table 2. Based on these findings, we will discuss in the following only results obtained under Model 3.

By looking at the estimates derived under the optimal model specification, we may firstly observe that the largest the population of destination, the biggest the migration flow, whereas, the population of origin seems to be negatively linked to migratory movements (significance level at 5%). In agreement with the standard theory behind gravity-type models, longer distances discourages movements (coefficient equal to -0.23); that is, when Tanzanian inhabitants migrate, they generally tend to prefer districts which are not that far from the district of origin.

As for environmental factors, we found that being affected by crop diseases or crop pest, or by a severe water shortage are important factors in pushing internal migrations (both coefficients equal to 0.020): as expected, the higher is the number of households experiencing one of such environmental shocks, the higher is the likelihood for them to move. On the other hand, experiencing droughts or floods does not seem to have a significant effect on migration flows. Also crop-related shocks associated to the district of destination are positively associated to migration flows, even if the magnitude of this effect is lower (coefficient equal to 0.015). This result may be possibly related to the negative sign of the parameter associated to the distance between districts. Tanzanian households tend to prefer closer destinations, which are more likely to be affected by similar problems in term of crop disease or pests with respect to the area of origin. Overall, estimated parameters suggest that environmental variables mainly act as as pushing factors for migration. Similarly, we found that the percentage of cultivated area of the district of origin – a characteristic which is related to the environment, although not representing a shock – is negatively related to inter-districts migrations: the average number of migrants is lower for those districts characterized by a low percentage of cultivated areas.

As for the variables describing the demographic, the social and the economic characteristics of the district, we found that migrants prefer destinations where the population is younger (probably the more dynamic areas). The variables referring to the percentage of population living in urban areas and the percentage of people who are owner of their home can be interpreted as proxies of the richness and development of an area. In this sense, the higher these indicators are for the districts of destination, the higher is the incoming migration flow they attract. Finally, in line with the literature, the attractiveness of an area increases when the corresponding educational level is higher. Generally, when considering demographic and socio-economic factors, it seems that they are relevant especially for the district of destination; in other words, such factors appear to be important as pull factors rather than as pushing ones.

We tested the effect of other objective climatic variables, coming from external databases, such as temperature and rainfall, but none of them were statistically significant. In the analysis, we also considered other demographic and socio-economic variables (such as the percentage of household headed by a woman; the illiteracy rate; the proportion of male population), but once again they were not significant.

5 Conclusions

The results of this analysis highlight that, as expected, demographic and socio-economic characteristics of Tanzanian districts are correlated to migration patterns, and that environmental variables are also significant. In particular, the contribution of environmental variables is mainly important in explaining the departure from a given district, whereas demographic and socio-economic characteristics are especially relevant in explaining the destination of the migration process. It is also worthwhile noticing that the distance remains the most important factor in determining the district of destination (the coefficient, equals to -0.23, has the biggest magnitude). In conclusion, Tanzanian inter-districts migrations are especially pushed by crop-related shocks and severe water shortage. At the same time, migrants generally prefer closer destinations, where they can expect to improve their social and economic situation (i.e., more developed and higher educated areas).

In the next steps of the analysis, we aim at adding a contiguity variable in the model specification, which is expected to better characterize migration flows (previous literature stated that areas that share a boundary tend to have more migration between them because this migration will include a proportion of short-distance moves from one side of the boundary to the other). Moreover, we aim at improving the description of the origin and destination districts, testing the effect of other environmental and socio-economic variables.

References

1. AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (1973), B. N. Petrov and F. Csaki, Eds., Akadémiai Kiado, pp. 267–281.
2. FLOWERDEW, R. Poisson regression modelling of migration. *Migration models: Macro and micro approaches* (1991), 92–112.
3. FLOWERDEW, R. Modelling migration with poisson regression. *Technologies for migration and commuting analysis: Spatial interaction data applications* (2010), 261–279.
4. FLOWERDEW, R., AND AITKIN, M. A method of fitting the gravity model based on the poisson distribution. *Journal of regional science 22* (1982), 191–202.
5. MYERS, N. Environmental refugees. *Population and environment 19* (1997), 167–182.
6. MYERS, N., KENT, J., ET AL. *Environmental exodus: an emergent crisis in the global arena.* Climate Institute, 1995.
7. OCELLO, C., PETRUCCI, A., TESTA, M. R., AND VIGNOLI, D. Environmental aspects of internal migration in tanzania. *Population and Environment 37* (2015), 99–108.
8. RAMLOGAN, R. Environmental refugees: a review. *Environmental conservation 23* (1996), 81–88.
9. STILLWELL, J. Inter-regional migration modelling-a review and assessment. Tech. rep., European Regional Science Association, 2005.
10. WAJDI, N., MULDER, C. H., AND ADIOETOMO, S. M. Inter-regional migration in indonesia: a micro approach. *Journal of Population Research 34* (2017), 253–277.
11. ZIPF, G. K. The P_1P_2/D hypothesis: on the intercity movement of persons. *American sociological review 11* (1946), 677–686.

Determinants and geographical disparities of BMI in African Countries: a measurement error small area approach.

Determinanti e disparità geografiche del BMI nelle regioni africane: un modello di piccola area con errore di misurazione.

Serena Arima and Silvia Polettini

Abstract Food insecurity remains one of the greatest challenges in many African countries, hindering their economic development. Among related indicators, women's body mass index (BMI), measuring women's nutritional status, is a key indicator of the socio-economic development of a country. Despite recent intervention programmes, geographic and socio-economic disparities remain in the BMI distribution. Therefore, it would be important to rely on accurate estimates of women's mean BMI levels across domains. We consider a small area model with area-specific random effects that capture the regional differences in BMI levels. We propose a Bayesian model to investigate the role on BMI of a number of socio-economic characteristics such as age, wealth, parity, education, while accounting for regional variation. Since it is reasonable to assume that some of these variables are measured with error, we develop a suitable methodology and investigate the effect of neglecting measurement error in covariates on the assessment of the regression effects and on the prediction of area-specific BMI mean levels. We apply the proposed model to DHS data to explore the geographical variability of the BMI in two different regions, namely Ethiopia and Nigeria, and compare the determinants of women's nutritional status in these countries.

Abstract *L'insicurezza alimentare rimane una delle maggiori sfide per molti paesi dell'Africa. L'indice di massa corporea (BMI) femminile non fornisce solo una indicazione sullo stato di salute delle donne, ma è uno degli indicatori più utilizzati per valutare lo sviluppo culturale ed economico dei paesi. Nonostante le numerose azioni intraprese dai governi, ancora oggi permangono grandi disparità geografiche e socio-economiche a livello nutrizionale. Per la valutazione e la pianificazione delle politiche occorrono stime accurate dei livelli medi del BMI delle donne a livello regionale; a tale scopo, il lavoro propone un modello bayesiano di*

Serena Arima
Dip. di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza University of Rome, via del Castro Laurenziano 9, 00161 Roma, e-mail: serena.arima@uniroma1.it

Silvia Polettini
Name, Address of Institute e-mail: silvia.polettini@uniroma1.it

piccola area con effetti casuali atti a cogliere la variabilità intra-regionale della risposta. Il modello lega la variabile di interesse a una serie di caratteristiche socio-economiche individuali quali età, indice di benessere economico, livello di istruzione, numero di figli, tenendo conto della variabilità geografica del BMI. Tuttavia per alcune di tali variabili è ragionevole ipotizzare che siano state misurate con errore. Il modello proposto è stato quindi esteso includendo nella sua formulazione anche la presenza di errore di misurazione nelle covariate. Utilizzando i dati indiviuali risultanti dalle indagini DHS, il modello proposto viene applicato per stimare e confrontare i livelli medi del BMI delle donne di due Paesi, Etiopia e Nigeria, che presentano caratteristiche molto diverse. Le stime dei parametri del modello consentono inoltre di valutare e confrontare le determinanti dello stato nutrizionale delle donne nei due Paesi.

Key words: BMI, Food insecurity, measurement error, small area models

1 Introduction

Although the proportion and absolute number of chronically undernourished people has declined worldwide, progress has been uneven among developing countries. Women are clearly the most critical target group from a nutrition standpoint, therefore data on women's nutritional status is essential in monitoring the socioeconomic development of a country. Indeed it has a direct impact on women's health status, and several indirect effects through the multiple roles of women in generating income inside and outside the household, bearing children and being responsible for their families' nutrition and care. Not surprisingly, for many countries this aspect has been the object of prioritized interventions in the achievement of the Millennium Development Goals.

In this paper we study the Body Mass Index (BMI) in two African countries: Ethiopia and Nigeria. These two countries are very different from several points of view. First of all their geographical position: Nigeria is located in West Africa while Ethiopia is located in the Horn of Africa. Nigeria is often referred to as the "Giant of Africa", owing to its large population and economy. With 186 million inhabitants, Nigeria is the most populous country in Africa and the seventh most populous country in the world. With over 102 million inhabitants, Ethiopia is the most populous landlocked country in the world and the second most populous nation on the African continent.

Despite progress toward eliminating extreme poverty, Ethiopia remains one of the poorest countries in the world, due both to rapid population growth and a low starting base. More than 70% of Ethiopia's population is still employed in the agricultural sector, but services have surpassed agriculture as the principal source of GDP. According to the 2016 World Bank figures, life expectancy in Ethiopia is about 65 years while in Nigeria it is about 53. However, the human development index has been estimated to be equal to 0.448 in Ethiopia and 0.527 in Nigeria, where the

GDP per capita is more than three times as high as Ethiopia (5861$ vs 1734$, data in constant 2011 international dollars). Also, the literacy rate among the population aged 15 years and older is much lower in Ethiopia (39% in 2007) than in Nigeria (55% in the same year).

The figures published for Nigeria by the National Population Commission & ICF International in 2014 based on the 2013 DHS survey data indicate that the mean BMI among women aged 15-49 is 23.0 kg/m2. Instead, the 2011 Ethiopia DHS data give a mean BMI of 20.2 kg/m2. While rural/urban disparities exist in both countries, small geographical differences emerge from the figures published for Nigeria, although the North West has the lowest mean BMI (21.9 kg/m2). Invariably, mean BMI increases with increasing education level and shows a steady increase with increasing wealth. In Nigeria, 11% of women of reproductive age are thin or undernourished (BMI less than 18.5 kg/m2), as opposed to Ethiopia, where, according to the 2011 DHS figures, the same percentage amounts to 27%. Besides this, in Nigeria obesity is a public health problem, with a 17% of women being overweight (BMI of 25-29 kg/m2), and 8% obese (BMI of 30 kg/m2 or above). Notice that in Ethiopia only 6% of women are overweight or obese. In both countries the prevalence of overweight and obesity among women of reproductive age increases with age and is reportedly higher in urban areas than in rural areas. In addition, the wealth index seems to be strongly associated with being overweight or obese.

In this work we study the BMI of women aged 15-49: we consider data from the 2011 Ethiopia DHS and 2013 Nigeria DHS. Data are available at `www.measuredhs.com`. Both surveys were designed to provide population and health indicators at the national (urban and rural) and regional levels (for Ethiopia, the following 11 regions: Tigray, Affar, Amhara, Oromiya, Somali, Benishangul-Gumuz, SNNP, Gambela, Harari, and two city administrations, Addis Ababa and Dire-Dawa; for Nigeria, the 36 states plus the Federal Capital Territory are planned domains, but we consider for comparison the following 6 geo-political zones: South East, South South, South West, North Central, North West and North East). For both countries, the number of observations sampled in each region of interest is not particularly small. However, especially for Ethiopia, high geographical variability and the large population size make the problem of estimating the mean BMI at the domain level worth to be framed in a small area context. This also allows us to investigate the individual determinants of BMI while accounting for geographical variability.

We develop a small area model for studying the effect on BMI level of several potential explanatory variables: for each women, we have considered the number of sons, the education level, if they live in urban or rural centers, the age and the wealth index. The wealth index is built from from sample information on asset ownership, housing characteristics and water and sanitation facilities; it is obtained via a three-step procedure, based on principal components analysis, designed to take better account of urban-rural differences in wealth indicators. Being the result of a complex procedure, we treat the wealth index as a categorical covariate subject to misclassification. We also consider age, being self reported, as a continuous variable observed with error.

2 Small area models

In recent years, small area estimation has emerged as an important area of statistics as a tool for extracting the maximum information from sample survey data. Sample surveys are generally designed to provide estimates of totals and means of variables of interest for large subpopulations or domains. However, governments and policy makers are more and more interested in obtaining statistical summaries for smaller domains such as states or provinces. These domains are called small areas and are usually unplanned so that a small number of units is allocated in each of these areas. *Indirect estimators* are often employed in order to increase the effective domain sample size by borrowing strength from the related areas using linking models, census, administrative data and other auxiliary variables associated with the small areas. Depending on the type of data available, small area models are classified into two types: area-level and unit-level. Area level models relate the small area means to area-specific auxiliary variables. Such models are essential if unit level data are not available. Unit level models relate the unit values of the study variable to unit-specific auxiliary variables with known area means. In this paper we focus on unit-level models within a Bayesian framework. See [9] for an up-to-date review.
In this paper we focus on unit-level small area models given the availability of record-level data, described in Section 1.
Suppose there are m areas and let N_i be the known population size of area i. We denote by Y_{ij} the response of the $j-$th unit in the $i-$th area ($i = 1,...,m$; $j = 1,...,N_i$). A random sample of size n_i is drawn from the $i-$th area. The goal is to predict the small area means

$$\theta_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij} \tag{1}$$

based on the observed sample. To develop reliable estimates, auxiliary information, often in forms of covariates, measured at the unit or at the area level, may be exploited.Adopting a superpopulation approach to finite population sampling, a unit-level small area model is defined as

$$Y_{ij} = \alpha + \beta x_i + u_i + \varepsilon_{ij} \qquad i = 1,...,m; \quad j = 1,...,N_i \tag{2}$$

where x_i is an auxiliary variable observed for each area. ε_{ij} and u_i are assumed independent, $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0,\sigma_e^2)$ and $u_i \overset{\text{iid}}{\sim} N(0,\sigma_u^2)$. A random sample of size n_i is selected from the $i-$th small area ($i = 1,...,m$).
The model in (2) may be estimated based on maximum likelihood [2, 8], Empirical Bayes [5] and hierarchical Bayes approaches [3].
As stressed in [6, 7] auxiliary variables may be measured with error: It is well recognized that the presence of measurement error in covariates causes biases in estimated model parameters and leads to loss of power for detecting interesting relationships among variables. Several solutions exist, also in the small area literature: indeed corrections of the unit-level and area-level estimators have been proposed both in a frequentist and Bayesian context [10, 4, 1]. Relying on the model proposed in

[6], we extend it in order to account for measurement error in both continuous and discrete covariates and explore the impact of our procedure in the assessment of covariates' effect in a small area model designed to estimate regional mean BMI level in Nigeria and Ethiopia.

3 Measurement error small area models

Consider a finite population, whose units are divided into m small areas. As in the previous section, let the population size of the i-th area be N_i, $i = 1, \ldots, m$. Let Y_{ij} be the value of the variable of interest associated with the j-th unit ($j = 1, \ldots, N_i$) in the i-th area ($i = 1, \ldots, m$). A random sample of size $n_i \geq 1$ is drawn from the $i-$th area population and the sample data are denoted by y_{ij} ($i = 1, \ldots, m$; $j = 1, \ldots, n_i$). For each area, we consider the following covariates: t_{ij} – the vector of p continuous or discrete covariates measured without error, w_{ij} and x_{ij} – respectively, a vector of q continuous covariates and h discrete variables (with a total of K categories), both measured with error. Denote by s_{ij} and z_{ij} the observed values of the latent w_{ij} and x_{ij}, respectively. We assume that the perturbation only depends on the unobserved category of the latent variable, so if $h > 1$ we assume independent misclassification. Without loss of generality, in what follows we assume $h = 1$.
Following the notation in [6], the proposed measurement error model can be written in the usual multi-stage way: for $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ and for $k, k' = 1, \ldots, K$

$$\text{Stage 1. } y_{ij} = \theta_{ij} + e_{ij} \qquad e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$$

$$\text{Stage 2. } \theta_{ij} = t_{ij}^{'}\delta + w_{ij}^{'}\gamma + \sum_{k=1}^{K} I(x_{ij} = k)\beta_k + u_i \qquad u_i \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$$

$$\text{Stage 3. } S_{ij}|w_{ij} \overset{\text{iid}}{\sim} N(w_{ij}, \sigma_s^2), \qquad w_{ij} \overset{\text{iid}}{\sim} N(\mu_W, \Sigma_w)$$

$$\mu_W \sim N(0, \sigma_\mu^2 I)$$

$$Pr(Z_{ij} = k|X_{ij} = k') = p_{k'k} \qquad p_{k'.} \sim \text{Dirichlet}(\alpha_{k',1}, \ldots, \alpha_{k',K})$$

$$Pr(X_{ij} = k') = \frac{1}{K}$$

We also assume that $\beta, \delta, \gamma, \sigma_e^2, \sigma_u^2, \sigma_s^2$ are, loosely speaking, a-priori mutually independent; in particular, $\beta \sim N(\mu_\beta, \sigma_\beta), \delta \sim N(\mu_\delta, \sigma_\delta), \gamma \sim N(\mu_\gamma, \sigma_\gamma)$, $\sigma_u^{-2} \sim Gamma(a_u, b_u)$, $\sigma_e^{-2} \sim Gamma(a_e, b_e)$, $\sigma_s^{-2} \sim Gamma(a_s, b_s)$.
Hyperparameters have been chosen to have flat priors. Finally, we assume $\Sigma_w = \sigma_w^2 I$, and σ_w^2, σ_μ^2 and $(\alpha_{k',1}, \ldots, \alpha_{k',K})$ all known.
Stage 3 describes the measurement error model for both continuous and discrete covariates: we assume that the continuous observable covariates S_{ij} are modeled as Gaussian variables centered at the true unobservable value w_{ij} with variability σ_s^2. The model for the unobservable continuous variables w_{ij} is assumed normal with

unknown mean and known variance.
Thanks to the multilevel model formulation, the measurement error mechanism need not to be assumed, which is a useful characteristic of our proposal. For the discrete covariates, the misclassification mechanism is specified according to the unknown $K \times K$ matrix P; we denote its $k'-$th row by $p_{k',}$, whose entries, $p_{k'k}$, represent the probabilities $P(Z_{ij} = k|X_ij = k')$, $k = 1,\ldots K$ that the observable variable Z_{ij} takes the $k-$th category, $k = 1\ldots,K$when the true unobservable variable X_{ij} takes the $k'-$th category. We also assume that the misclassification probabilities are the same across subjects and that all the categories have the same probability $\frac{1}{K}$ to occur.
Over each $p_{k',}$, $k' = 1,\ldots,K$, we place a Dirichlet$(\alpha_{k',1},\ldots,\alpha_{k',K})$ prior distribution, with known parameter. According to the above assumptions, we can estimate the transition matrix P jointly with all the other model parameters.
Using the Bayes theorem, the posterior distribution of the unknown parameter is proportional to the product of the likelihood and the prior distributions specified in Stage 4. As the posterior distribution cannot be derived analytically in closed form, we obtain samples from the posterior distribution using Gibbs sampling.

4 Results and comments

For each country we consider the estimates of the regression parameters obtained under the two models, with and without accounting for measurement error. Previous studies show that not accounting for measurement error may lead to inaccurate estimation of regression coefficients, which in turn may affect small area predictions. Figures 1 and 2 report the posterior distributions of the regression parameters under both models for Ethiopia and Nigeria, respectively. Under the measurement error model (top panels), the covariates' effects are all consistent with expectations. The BMI increases with the wealth index category: the poorest women are more likely to be underweight than the richest ones. Although expected, such an important effect of the wealth index has not been always confirmed in previous studies. Also, in both countries more educated women show a larger BMI than less educated ones, with an effect that increases with the educational level. The model also highlights the great disparity between urban and rural areas, where the women's undernutrition problem is more severe. The number of children ever born (parity) is found to affect women's nutritional status significantly only in Ethiopia, where the BMI decreases with parity. With respect to age, the model highlights positive linear association with BMI: younger women are more likely to be underweight than older ones, as well documented in the literature. On the other hand, under the model that ignores the measurement error in wealth index and age, the strong differential effect of the wealth index is lowered or, in the case of Ethiopia, disappears (see the bottom panel of Figure 1). This is also consistent with findings in the literature, that sporadically identifies this variable as important. With respect to the other parameters, while the meaning of the coefficients is coherent with those obtained with the proposed model, the variables' effects are considerably inflated. Noticeably, for the Nigeria data parity is only significant under the unadjusted model. On the other hand, a linear effect of the wealth index on the BMI emerges quite clearly from the measurement error models in both countries, but not from the naive model. Moreover, the measurement error plays a different role in the two models: in the Ethiopia data, it strongly affects the parameters' estimates and, as expected, the posterior distribution of the P matrix is far from the diagonal one. On the other hand, in the Nigeria data the measurement error has a smaller impact on the estimates as the posterior distributions of the diagonal elements of P are concentrated around 0.9. In conclusion, the proposed model seems to be fairly robust with respect to misspecification of the measurement error mechanism.

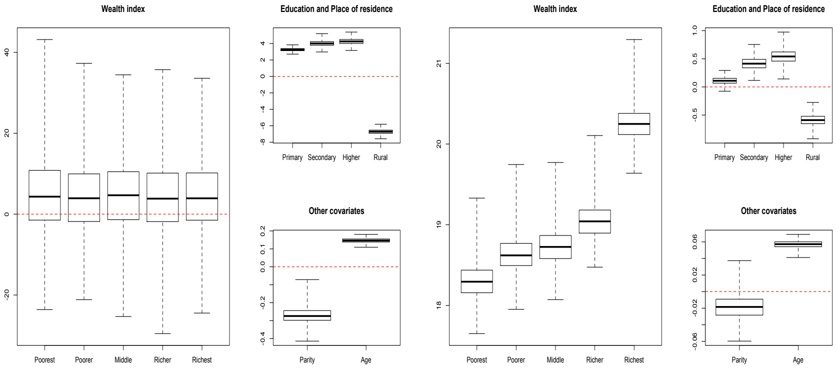


Fig. 1 Ethiopia data: posterior distributions of the model parameters under the proposed model (left panel) and under the assumption that the covariates are measured without error (right panel).

References

1. Arima, S., Datta, G.S. and Liseo, B. Bayesian Estimators for Small Area Models when Auxiliary Information is Measured with Error, Scandinavian Journal of Statistics, **42 (2)**,518–529, 2015
2. Battese, G.E., Harter, R.M. and Fuller, W.A. An error components model for prediction of county crop areas using survey and satellite data, Journal of the American Statistica Association, **83**, 28–36, 1988.
3. Datta, G.S. and Ghosh, M. Bayesian prediction in linear models: applications to small area estimation, Annals of Statistics, **19**, 1748–1770, 1991.
4. Datta, G.S., Rao, J.N.K. and Torabi, M. Pseudo-empirical Bayes estimation of small area means under a nested error linear regression model with functional measurement errors, Journal of Statistical Planning Inference, **140 (11)**, 2952–2962, 2010
5. Ghosh, M. and Rao, J.N.K. Small area estimation: an appraisal, Statistical Sciences, **9**, 55–93, 1994

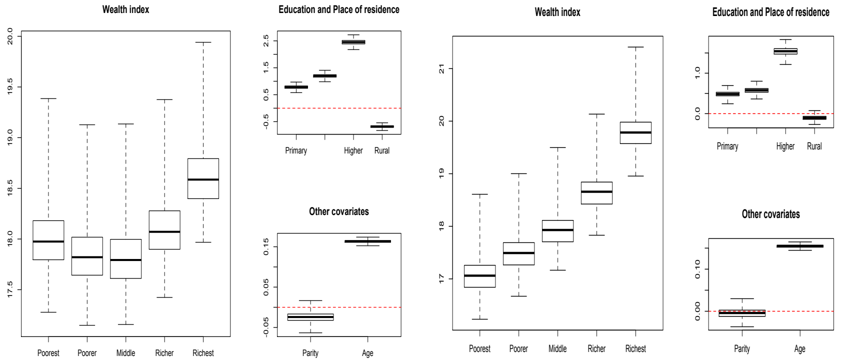


Fig. 2 Nigeria data: posterior distributions of the model parameters under the proposed model (left panel) and under the assumption that the covariates are measured without error (right panel).

6. Ghosh, M., Sinha, K. and Kim, D. Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models, Scandinavian Journal of Statistics, **33(3)**, 591-608, 2006
7. Ghosh, M. and Sinha, K. Empirical Bayes estimation in finite population sampling under functional measurement error models, Journal of Statistical Planning Inference **137**, 2759–2773, 2007
8. Prasad, N.G.N. and Rao, J.N.K. The estimation of mean squared error of small area estimators. Journal of the American Statistical Association, **85**, 163–171,1990.
9. Rao, J.N.K. and Molina, I.: Small Area Estimation, 2nd Edition, Wiley, Hoboken, New Jersey, 2015.
10. Ybarra, L.M.R. and Lohr, S.L. Small area estimation when auxiliary information is measured with error, Biometrika, **95(4)**, 919–931, 2008

SIS
2018
PALERMO
49 TH SCIENTIFIC MEETING
OF THE ITALIAN
STATISTICAL SOCIETY
20-22 JUNE

5. Contributed Sessions

Advanced Algorithms and Computation

Brexit in Italy

Text Mining of Social Media

Francesca Greco, Livia Celardo, Leonardo Salvatore Alaimo

Abstract The aim of this study is to identify how Italian people talk about Brexit on Twitter, through a text mining approach. We collected all the tweets in Italian language containing the term "Brexit" for a period of 20 days, obtaining a large corpus on which we applied multivariate techniques in order to identify the contents and the sentiments within the shared comments.

Abstract *Questo studio ha lo scopo di identificare in che modo Brexit viene discussa su Twitter dagli Italiani attraverso l'analisi automatica del testo. A questo scopo sono stati raccolti tutti i messaggi in lingua italiana contenenti i termini "Brexit" per 20 giorni, ottenendo un corpus di grandi dimensioni su cui sono state applicate delle tecniche statistiche multivariate al fine di individuare i contenuti e i sentimenti relativi al tema in esame.*

Key words: Brexit, Twitter, Emotional text mining, Co-clustering

Introduction

There is a growing increase in Euroscepticism among EU citizens nowadays, as shown by the development of the ultra-nationalist parties among the European countries. Regarding to European Union membership, public opinion is divided between Eurosceptics and pro-Europeans, as clearly shown by the results of the 2016 British referendum. Many studies about Brexit are focused on the analysis of the electoral result, trying to highlight the effects of possible determinants - such as immigration, economic crisis, socio-economic and demographic characteristics of the voting population - on the electoral choices (Gietel-Bastel, 2016; Goodwin & Heath, 2016; Clark et al., 2017; Alaimo, 2018). Other studies analyse the possible consequences of Brexit, focusing on the economic consequences for United Kingdom (Dhingra et al., 2016; Dodds, 2016; Vikers, 2017). Although Brexit has shaken the European public opinion, there are few studies about how the referendum is actually

[1] Francesca Greco; Prisma S.r.l., Sapienza University of Rome; francesca.greco@uniroma1.it
Livia Celardo; Sapienza University of Rome; livia.celardo@uniroma1.it
Leonardo Salvatore Alaimo; Sapienza University of Rome; leonardo.alaimo@uniroma1.it

perceived by the citizens of European Member States. Moreover, not many analyses are available concerning how Brexit is discussed on the social media.

The wide diffusion of the internet increases the opportunity for millions of people to surf the web, create account profiles and search or share information every day. The constant rise in the number of users of social media platforms, such as Twitter, makes a large amount of data available; these data represent one of the primary sources for exploring people's opinions, sentiments, and emotions (Ceron et al., 2013; Pelagalli et al., 2017).

Due to that, we decided to perform a quantitative study where online discourses regarding Brexit are analysed using two text analysis techniques in parallel: Content Analysis and Emotional Text Mining. The aim is to explore not only the contents but also the sentiments shared by users on Twitter. In this paper we focused only on the analysis of the sentiments and contents published in Italian, in order to understand how Brexit is treated in Italy.

Methods

In order to explore the sentiments and the contents related to Brexit, we scraped from Twitter all the messages written in Italian containing the word *Brexit*, produced from January 23rd to February 18th, 2018. The data extraction was carried out with the TwitteR package of R Statistics. From the data we extracted, we decided to create two sub-corpora: the first one including the retweets and the other one excluding all the retweets. We chose to use two different corpora for the analyses because we were studying both sentiments and contents within our texts; for the analysis of sentiments and emotions it is important to consider also retweets, while for content analysis retweets are just the repetition of the same concepts. Then, the first corpus was composed of 13,662 messengers, including 76.4% of retweets, resulted in a large size corpus of 211,205 of tokens, which underwent the Emotional Text Mining (Greco et al., 2017). A second corpus was extracted excluding all the retweets, resulting in a large size corpus of 46,458 tokens, which underwent the content analysis. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio (TTR) and the hapax percentage ($TTR_{corpus\ 1}$ = 0.04; $Hapax_{corpus\ 1}$ = 42.2%; $TTR_{corpus\ 2}$ = 0.16; $Hapax_{corpus\ 2}$ = 61.3%). According to the large size of the corpora, the lexical indicators indicate the possibility to proceed with the analyses.

Then, on the first corpus was performed a sentiment analysis, which is a technique used to investigate the sentiments of a text. In text mining, many methods exist to analyse it automatically, which are supervised and unsupervised (e.g., Carli & Paniccia, 2002; Hopkins & King, 2010; Bolasco, 2013; Ceron et al., 2016). We performed the Emotional Text Mining (ETM) (Greco, 2016; Pelagalli et al., 2017; Greco et al., 2018), which is an unsupervised method derived from the Emotional Textual Analysis of Carli and Paniccia (Carli & Paniccia, 2002; Bolasco, 2013); it is based on the idea that people emotionally symbolize an event or an object, and socially share this symbolisation. The words they choose to talk about this event or object is the product of the socially-shared unconscious symbolization. According to this, it is possible to detect the associative links between the words to infer the symbolic matrix determining the coexistence of these terms in the text. In order to perform ETM, first corpus was cleaned and pre-processed with the software T-Lab (version T-Lab Plus 2018) and keywords were selected. In particular, we used lemmas as keywords instead of types, filtering out the lemma *Brexit* and those lemmas of the low rank of frequency (Greco, 2016). Then, on the tweets per keywords matrix, we performed a cluster analysis using the bisecting *k*-means algorithm, limited to twenty partitions (Savaresi & Boley, 2004) and excluding all the tweets having less than two keywords co-occurrence. The eta squared index was used to evaluate and choose the optimal solution in terms of number of clusters. To complete the analysis, a correspondence analysis (Lebart & Salem, 1994) on the keywords per clusters matrix was made, in order to explore the relationship between clusters and identify the emotional categories. The main advantage connected with this approach is the possibility in interpreting the factorial space according to words polarization, thus identifying the emotional categories that generate Brexit representations, facilitating the interpretation of clusters and exploring their relationships within the symbolic space.

On the other hand, content analysis is a technique used to investigate the subjects treated within a text; in text mining, many methods exist to analyse the contents automatically. One of these is the Text Clustering; it consists of splitting the corpus in different subgroups based on words/documents similarities (Iezzi, 2012). In this paper, a text co-clustering approach (Celardo et al., 2016) for contents analysis is used. The objective is to simultaneously classify rows and columns, in order to identify groups of texts characterized by specific contents. To do that, data were pre-processed with Iramuteq software: we lemmatized the texts and we removed stop words and terms with frequency lower than 2. The weighted terms-documents matrix was then co-clustered through the double *k*-means algorithm (Vichi, 2001); the number of clusters for both rows and columns was identified using the Calinski-Harabasz index.

Results

The results of the ETM show that the 240 keywords selected allowed us to classify the 79% of the tweets. The eta squared index was calculated on different partitions – from 3 to 9 clusters, and the values we found showed that the optimal solution is four clusters. The correspondence analysis detected three latent dimensions (Table 1). In figure 1, we can see the emotional map of Brexit emerging from the Italian tweets. It shows how the clusters are placed in the factorial space. The first factor represents the evaluation of the Brexit deal, considering the exit from the EU a bad deal or a good one; the second factor reflects the British political strategy, differing the partisan strategy – which is led by specific political and economic interests, from the public strategy, which is carried on by the government; finally, the third factor represents the post-Brexit effects evaluation, distinguishing the forthcoming impact from the future one. The four clusters are of different sizes (Table 2), and they reflect the Italian sentiments toward Brexit. The first cluster represents the Soros scandal as a conspiracy that disregards the democratic expression of Britons' choice; the second cluster reflects Italian satisfaction in considering Brexit as a bad deal for Britons, who are punished for their betrayal; the third cluster concerns the negative European economic impact of the British policy, which is perceived as an unfair British advantage causing a loss for EU citizens; and the fourth cluster highlights the hope for a British comeback through a new referendum. By clusters interpretation, no group highlights a positive sentiment in the direction of British exit from the EU. Nevertheless, we have considered as positive or satisfactory (51.5%) the British punishment and the hope for a British comeback, and negative the other two (48.5%), the unfair British advantage and the citizen right violation.

Table 1 - *Explained inertia for each factor*

Factor	*Eigenvalues*	*%*	*Cumul. %*
1	0,705	38,3	38,3
2	0,621	33,7	72,0
3	0,515	28,0	100,0

Figure 1 - *Factorial space set by three factors*

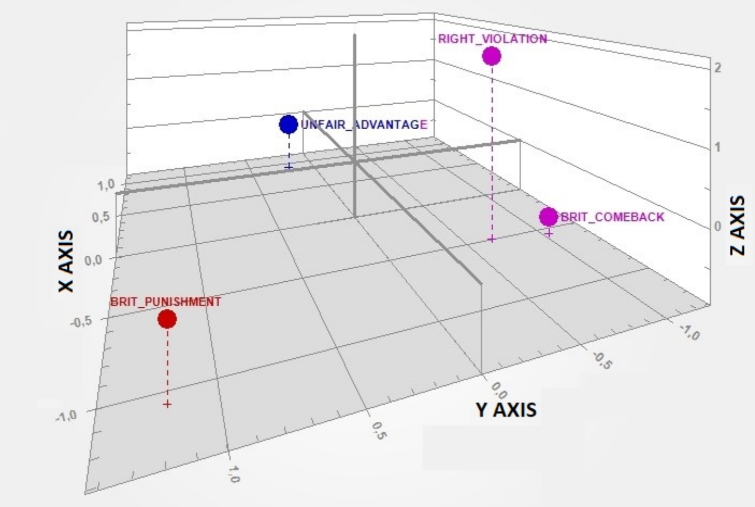


Table 2 – *Brexit representations and sentiments*

Clusters	*No. tweets classified*	*Size*	*Label*	*Keywords*	*CU*	*Sentiment*
1	1355	13.0%	Citizen right violation	Soros	792	Negative
				media	496	
				documento	425	
				telegraph	383	
				prova	372	
				cercare	359	
2	2517	24.0%	British punishment	ammettere	867	Positive
				Regno Unito	809	
				governo	789	
				peggio	692	
				britannico	606	
				Europa	436	
3	3733	35.5%	Unfair British advantage	europeo	784	Negative
				UK	717	
				Italia	618	
				occupazione	545	
				sterlina	469	
				anti-Brexit	461	
4	2894	27.5%	British comeback	UE	912	Positive
				Soros	769	
				Londra	675	
				May	378	
				voto	352	
				segreto	344	

The first six keywords of the clusters are ordered by the number of context units (CU) classified in each cluster

For the content analysis, we firstly pre-processed the corpus, removing all the noise – i.e. stop-words – cutting the messages contents by 83%. Then, on the terms-documents matrix (where documents are represented by the tweets) we calculated the Calinski-Harabasz index, in order to find the number of groups, for both the rows and the columns. The index indicated five groups for words and five for tweets; the results of the co-clustering procedure are shown in the Table 3. The first group of words identifies the common language, representative of all the messages; it is about the general conditions and aspects of Brexit. The 35% of the tweets has, in addition to this, a specific language. Almost the 30% of messages is about the impacts and the effects of Brexit on the local economies and on the political systems. A smaller share of tweets deals with the Soros scandal (5%), while just few messages are about what is going to happen in the near future in Italy.

Table 3: *Centroids matrix resulting from the co-clustering procedure*

Cluster - Label	***1***	***2***	***3***	***4***	***5***
	(65%)	(14%)	(13%)	(5%)	(3%)
1- Brexit	0.005	0.001	0.002	-0.001	0.001
2 - EU Impact	0.002	0.000	**0.049**	-0.002	-0.002
3 - Economic Effects	0.005	**0.066**	0.004	-0.003	0.002
4 - Soros Scandal	0.005	0.004	0.006	-0.007	**0.668**
5 - Forthcoming	0.009	0.004	0.001	**0.624**	0.028

Discussion and conclusion

The results of our analyses showed how “Brexit” is represented by Italians both in terms of sentiments and contents. The Emotional Text Mining revealed the presence of positive and negative sentiments in respect to the consequences of Brexit, but not directly toward the UK exit. A positive sentiment is connected to the punishment of the British choice and the hope of a rethink concerning the UK exit, which is basically perceived as a betrayal. In fact, negative opinions rely on the unfair British advantage and the disregards of the Britons’ referendum choice.

Regarding the content analysis, it identified in the first cluster the common language shared within Twitter, used to describe what happened and is continuing to happen in Europe after Brexit. The other four clusters underline the presence of a specific language. In particular, the second and third clusters focus on the current political and economic effects of Brexit in Europe.

The results of the two analyses showed that Brexit is a theme with a strong emotional charge in the Italian context. Italian people seem to focus their attention mainly toward the economic and political effects of the British choice. This subject is treated negatively from the users, probably because of the worries for the consequences of the British vote in Europe.

References

1. Alaimo, L. S.: Demographic and socio-economic factors influencing the Brexit vote. Rivista Italiana di Economia, Demografia e Statistica (RIEDS), 72(1), 17–28 (2018).
2. Bolasco S.: L'analisi automatica dei testi: fare ricerca con il text mining. Carocci, Roma (2013).
3. Carli R., Paniccia R. M.: Analisi Emozionale del Testo. Franco Angeli, Milano (2002).
4. Celardo, L., Iezzi, D. F., Vichi, M.: Multi-mode partitioning for text clustering to reduce dimensionality and noises. In: Mayaffe, D., Poudat, C., Vanni, L., Magri, V., Follette, P. (eds) JADT 2016: Statistical Analysis of Textual Data. Les Press de Fac Imprimeur, Nizza (2016).
5. Ceron, A., Curini, L., Iacus, S.M.: Social Media e Sentiment Analysis. L'evoluzione dei fenomeni sociali attraverso la Rete. Springer, Milano (2013).
6. Ceron A., Curini L., Iacus S. M.: iSA: a fast, scalable and accurate algorithm for sentiment analysis of social media content. Information Sciences, 367, 105-124 (2016).
7. Clark, H.D., Goodwin, M., Whiteley, P.: Brexit: Why People Voted to Leave the European Union. Cambridge University Press, Cambridge (2017).
8. Dhingra, S., Ottaviano, G., Sampson, T., Van Reenen, J.: The consequences of Brexit for UK trade and living standards. Centre for Economic Performance (CEP), London School of Economics and Political Science (LSE) (2016).
9. Gietel-Bastel, S.: Why Brexit? The Toxic Mix of Immigration and Austerity. Population and Development Review, 42(4), 673–680 (2016)
10. Goodwin, M. J., Heath, O.: The 2016 Referendum, Brexit and the Left Behind: An Aggregate-level Analysis of the Result. The Political Quarterly, 87(3), 323-332 (2016)
11. Greco, F.: Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale. Franco Angeli, Milano (2016).
12. Greco, F., Mascietti, D., Polli, A.: Emotional text mining of social networks: the French pre-electoral sentiment on migration. RIEDS (2018) Available from http://www.sieds.it/index.php?option=com_content&view=article&id=17:rivista-rieds&catid=26:pubblicazioni&Itemid=136
13. Hopkins D., King G.: A method of automated nonparametric content analysis for social science, American J. Pol. Sci., 54(1), 229-247 (2010).
14. Iezzi, D. F.: Centrality measures for text clustering. Communications in Statistics-Theory and Methods, 41(16-17), 3179-3197 (2012).
15. Lebart, L., Salem, A.: Statistique Textuelle. Dunod, Paris (1994).
16. Pelagalli, F., Greco, F., De Santis, E.: Social emotional data analysis. The map of Europe. In: Petrucci A., Verde R. (eds) SIS 2017. Statistics and Data Science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society, Florence 28-30 June 2017.: Firenze University Press (2017).
17. Savaresi, S.M., Boley, D.L.: A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. Intelligent Data Analysis, 8(4), 345-362 (2004).
18. Vichi, M.: Double k-means clustering for simultaneous classification of objects and variables. Advances in classification and data analysis, 43-52 (2001).
19. Vickers, J. Consequences of Brexit for Competition Law and Policy. Oxford Review of Economic Policy, 33 (2017).

Distance based Depth-Depth classifier for directional data

DD-classifier per dati direzionali basati su funzioni di distance-depths

Giuseppe Pandolfo and Giovanni C. Porzio

Sommario The DD-classifier, which has been extended to the classification of directional objects, is here investigated in the case of some new distance-based directional depths. The DD-classifier is a non-parametric techniques based on the depth vs. depth (DD) plot. Its main advantages concern the flexibility and the independence from specific distribution parameters. The new depth functions adopted here allow using them for high-dimensional directional data sets.
Sommario *Il DD-classifier, che è stato esteso alla classificazioni di dati direzionali, viene qui analizzato nel caso di due nuove funzioni di depth recentemente introdotte. Il classifier è un metodo non parametrico basato sul depth vs. depth plot. I maggiori vantaggi riguardano la flessibilità e l'indipendenza da specifiche ipotesi distribuzionali. La nuove funzioni di depth, essendo particolarmente vantaggiose in termini computazionali, permettono di utilizzare questa metodologia anche nel caso di insiemi di dati direzionali a dimensioni elevate.*

Key words: Angular data depth, Chord depth, Classification, Cosine depth.

1 Introduction

Directional data arise when observations are measured as directions or angles, and observations are represented as points on the circumference of the unit circle, or on the surface of the unit hyper-sphere. Simple examples are directions on a compass ($0^\circ - 360^\circ$), and times of the day ($0 - 24$ hours). Higher dimensional directional data arise in text mining and gene expression analysis.

Giuseppe Pandolfo
University of Naples Federico II, Napoli, e-mail: giuseppe.pandolfo@unina.it

Giovanni C. Porzio
University of Cassino and Southern Lazio, Cassino e-mail: porzio@unicas.it

They play thus an important role in many fields such as biology, meteorology, neurology and physics. Within the literature, directional data are presented and discussed in books by Mardia (2000) and Batschelet (1981). They provide a wide survey about specific features and problems we have to face when dealing with them.

Specific statistical methods are needed to analyse directional data. This is due to the periodicity and boundness of the sample space. To make it clearer, one may consider two angles of 10° and 350° which are 20° far away from each other. If treated as linear data, their arithmetic mean would be equal to 180°, while their correct directional mean is 0°. Hence, to avoid misleading results, appropriate techniques are necessary to perform directional data analysis, and this applies to classification in directional spaces as well.

In this work, the focus is on supervised classification (also called discriminant analysis), where the aim is to assign observations to classes (or groups), given a previous knowledge on their structures obtained through some preliminary observed data.

Specifically, we consider the directional non-parametric DD-classifier introduced in Pandolfo (2014, 2015), and further investigated in Pandolfo (2017).

The classifier evaluates the depth of a new observation with respect to some previously collected samples (a depth wrt each of the groups), and then uses its value as a basis for a classification rule. While Pandolfo (2014, 2015, 2017) investigated such a tool adopting different notions of depth functions already available within the literature, this work focuses on the advantages that can be obtained if some new depth functions are adopted. Pandolfo et al. (2017) recently introduced a new class of functions based on directional distances. It seems one of the main advantages of these functions is their computational feasibility. For this reason, they might be particularly suitable to deal with non-parametric classification problem in high-dimensional spaces. This essentially motivates this work, whose aim is to present a performance analysis of the directional DD-classifier under these two new depth functions.

The paper is organized as follows: Section 2 briefly recalls the definition of these new directional depth functions and presents the related classifier; Section 3 offers some concluding remarks.

2 Non-parametric classification using data depth

The literature on nonparametric discriminant analysis for directional data is not particularly extensive. Liu and Singh (1992) discussed methods for ordering directional data and suggested a possible solution for discriminant analysis. That is, an observation can be classified according to its center-outward ordering with respect to two given competing distributions. Later, following the linear approach by Stoller (1954) a circular discriminant analytical method was proposed by Ackermann (1997). This method, which looks for the pair of splitting angles which maximizes the probability of correct classification. Unfortunately, this solution requires solving an NP-complex problem, and hence it is quite computationally infeasible.

Several tools are available for classification problems in standard multivariate analysis. Assuming normality, linear discriminant analysis is probably the most used. Many other methods, both parametric and non-parametric, are available as well. Non-parametric classifiers have the important advantage to be more flexible, given that they do not require any distribution assumptions. Amongst them, this work focuses on classifiers based on data depth. Considering their full non-parametric nature, they can be used in many different contests.

The depth of a point $x \in \mathbb{R}^q$ (for dimension $q \geq 1$) is a function that measures the "centrality" or "deepness" of it with respect to a multivariate distribution F or a multivariate data cloud, and it is denoted by $D(x, F)$.

Some notions of angular data depth are available within the literature: the *angular simplicial depth*, the *angular Tukey's depth* and the *arc distance depth*. All these were given in Liu and Singh (1992). More recently Pandolfo et al. (2017) introduced a class of depth functions for directional data based on angular distances. The class is defined below.

Let $\mathscr{S}^{q-1}$ be the unit sphere in a $q-1$ dimensional space. A particular member of the class will be obtained by fixing a particular (bounded) distance $d(\cdot,\cdot)$ on $\mathscr{S}^{q-1}$. For such a distance, $d^{\sup} := \sup\{d(\theta,\psi) : \theta,\psi \in \mathscr{S}^{q-1}\}$ will denote the upper bound of the distance between any two points on $\mathscr{S}^{q-1}$. We have the following definition (Pandolfo et al., 2017).

Definition 1 (Directional distance-based depths) *Let $d(\cdot,\cdot)$ be a bounded distance on $\mathscr{S}^{q-1}$ and H be a distribution on $\mathscr{S}^{q-1}$. Then the* directional d-depth of $\theta(\in \mathscr{S}^{q-1})$ with respect to H *is*

$$D_d(\theta, H) := d^{\sup} - E_H[d(\theta, W)], \tag{1}$$

where E_H is the expectation under the assumption that W has distribution H.

Accordingly, two new easy-to-compute directional depth functions can be obtained if the distance in (1) is chosen to be the chord or the cosine distance, respectively. While the main properties of these functions have been investigated in Pandolfo et al. (2017), their use within a DD-classifier is investigated here.

2.1 Depth-based classifiers and the DD-plot

After the first suggestion in Liu and Singh (1992), the use of data depth to perform supervised classification has been suggested and investigated by many authors. Two main approaches have been adopted in the literature: the *maximum depth classifier* and the *Depth vs Depth* (*DD*) classifier. The latter (that is a refinement of the former) is based on the *DD*-plot (Depth vs Depth plot), introduced by Liu et al. (1999).

The *DD*-plot is a two-dimensional scatterplot where each data point is represented with coordinates given by its depth evaluated with respect to two distributions. A classification rule $r(\cdot)$ is then directly applied in this latter. The same procedure can be applied to any kind of data, providing that a corresponding depth function exists. For istance, DD-plot for functional data have been developed.

For directional data, the following questions arise. First, it is of interest to investigate if some depths perform better than others. Second, the classification rule to be adopted within the plot may also affect performances. A proper simulation study has been thus developed in order to suggest under which conditions one depth function should be preferred over the other.

3 Concluding remarks

The depth vs. depth classification method extended to directional data has been investigated here when some new distance based depth functions are adopted. The idea of depth provides a criterion to order a directional sample from center-outward, providing a new way to classify directional objects. The performance in terms of average misclassification rate of the classifiers is evaluated by means of a simulation study. Furthermore, their use is illustrated through a real data example. First results are promising, calling for further investigation on the performance under different directional settings.

Riferimenti bibliografici

1. Ackermann, H.: A note on circular nonparametrical classification. Biometrical J **5**, 577–587 (1997)
2. Batschelet, E.: Circular statistics in biology. Academic Press, London, (1981)
3. Liu, R.Y., Singh, K.: Ordering directional data. Concepts of data depth on circles and spheres. Ann Stat **20**, 1468–1484 (1992)
4. Liu, R.Y., Singh, K.: Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann Stat **27**, 783–1117 (1999)
5. Mardia, K.V., Jupp, E.P.: Statistics of directional data. Academic Press, London (1972)
6. Pandolfo, G.: On depth functions for directional data. Ph.D. Thesis. Department of Economics and Law, University of Cassino and Southern Lazio (2014)
7. Pandolfo, G.: A depth-based classifier for circular data. Mola F., Conversano C. eds., CLADAG 2015 BOOK of Abstracts, CUEC Editrice, 324–327 (2015)
8. Pandolfo, G., Paindaveine, D., Porzio, G.C.: Distance-based depths for directional data. Working Papers ECARES 2017 – 35 (2017) Available at https://ideas.repec.org/p/eca/wpaper/2013-258549.html
9. Stoller, D.S.: Univariate two-population distribution-free discrimination. J Am Statist Assoc **49**, 770–777 (1954)

Approximate Bayesian Computation for Forecasting in Hydrological models

Metodi Bayesian approssimati per le previsioni nei modelli idrologici

Jonathan Romero-Cuéllar, Antonino Abbruzzo, Giada Adelfio and Félix Francés

Abstract Approximate Bayesian Computation (ABC) is a statistical tool for handling parameter inference in a range of challenging statistical problems, mostly characterized by an intractable likelihood function. In this paper, we focus on the application of ABC to hydrological models, not as a tool for parametric inference, but as a mechanism for generating probabilistic forecasts. This mechanism is referred as Approximate Bayesian Forecasting (ABF). The abcd water balance model is applied to a case study on Aipe river basin in Columbia to demonstrate the applicability of ABF. The predictivity of the ABF is compared with the predictivity of the MCMC algorithm. The results show that the ABF method as similar performance as the MCMC algorithm in terms of forecasting. Despite the latter is a very flexible tool and it usually gives better parameter estimates it needs a tractable likelihood.
Abstract *In questo articolo, il metodo chiamato Approximate Bayesian Computation (ABC) viene applicato ai modelli idrologici, non come uno strumento per l'inferenza parametrica, ma come un meccanismo per generare previsioni probabilistiche, dando luogo all'Approximate Bayesian Forecasting (ABF). L'ABF è applicato a un caso studio sul bacino del fiume Aipe in Colombia. Viene considerato un modello idrologico semplice per dimostrare l'applicabilità di ABF e confrontarlo con la predittività del metodo MCMC. Nonostante i risultati mostrano che il metodo ABF e l'algoritmo MCMC non differiscono in termmini di previsioni ottenute, l'ABF è comunque uno strumento molto flessibile e fornisce risultati utili anche quando si è in presenza di una verosimiglianza intrattabile.*

Key words: Predictive uncertainty, Probabilistic post-processing approach, Bayesian forecasting, Sufficient statistics, Hydrological models, Intractable likelihood.

Jonathan Romero-Cuéllar and Félix Francés
Research Institute on Water and Environmental Engineering, Universitat Politécnica de Valéncia, Spain, e-mail: jorocue1@doctor.upv.es

Antonino Abbruzzo and Giada Adelfio
Department of Economics, Business and Statistical Sciences, Universitá degli Studi di Palermo, Italia, e-mail: antonino.abbruzzo@unipa.it

1 Introduction

In hydrological models, predictions are crucial for supporting decision-making and water management. Reliability of prediction of hydrologic outcomes is affected by several sources of uncertainty such as input or forcing data uncertainty, initial conditions, model uncertainty or epistemic error, parameters inference, output uncertainty. So several sources of uncertainty affect the full predictive uncertainty, that is the probability of occurrence of a future value of a response variable (streamflow, water level) conditional on all the covariates, usually provided by forecasting models [9]. Therefore, the forecast approaches, rather than looking for deterministic predictions, essentially aim at quantifying predictive uncertainty. Predictive uncertainty estimation in hydrological models is a challenge when dealing with intractable likelihoods. The Approximate Bayesian Computation (ABC) overcomes the likelihood-based approach via the use of sufficient statistics and simulated data [1]. The idea behind the ABC approach was first introduced in population and evolutionary genetics [6, 7]. The ABC has a wide range of application domains because it is useful when an explicit likelihood function cannot be justified [10]. The main focus, of the most of the studies about the ABC, is the quantification of uncertainty about parameters [2]. Together with the increasing applications of ABC (see [4] for recent surveys), attention has recently been paid to the theoretical properties of the method, including the asymptotic behaviour of: the ABC posterior distributions, the point estimates derived from those distributions, and the Bayes factors that condition on summaries (see for instance [4]). The ABC approach in hydrological models is introduced in [11], using the ABC to estimate posterior distributions of parameters for simulation-based models.

The aim of this paper is to introduce the ABC as an approach for generating probabilistic forecasts in hydrological models. This approach is referred to Approximate Bayesian Forecasting (ABF) [2]. A streamflow forecasting on a case study of the Aipe river basin in Columbia is used to show the potential strength of the ABF. Predictions derived from the ABF algorithm are compared to prediction derived from the MCMC algorithm.

This paper is structured as follows. In the first section, we describe a simple hydrological model. In the second section, we describe the application of the ABC for the hydrological model. In the third section, we compare the ABF and the MCMC algorithm.

2 Approxiamte Bayesian Forecasting for the hydrological model

The abcd water balance model is a hydrological model for simulating streamflow (see [8]). This model is a fairly general conceptual rainfall-runoff model which transforms rainfall and potential evapotranspiration data to streamflow at the catchment outlet. The model is comprised of two storage compartments: soil moisture and groundwater. The soil moisture gains water from precipitation and loses water to evapotranspiration, surface runoff, and groundwater recharge. The groundwater compartment gains water from recharge and loses water as discharge. The total streamflow, which is the outcome we are interested in, is the sum of surface runoff from the soil moisture and groundwater discharge. It applies the continuity equation to a control volume representing the upper soil zone, from which evapotranspiration is assumed to occur, so that

$$Sw_t + ET_t + Q_t + R_t = Sw_{t-1} + P_t, \tag{1}$$

where P_t is a total precipitation for the month, ET_t is actual evapotranspiration, R_t is recharge to groundwater storage, Q_t is upper zone contribution to runoff, and Sw_t and Sw_{t-1} represent upper soil zone soil moisture storage at the current and previous time steps respectively. For the groundwater component, the mass balance equation is

$$Sg_t + Qg_t = Sg_{t-1} + R_t, \tag{2}$$

where Qg_t is groundwater discharge, Sg_t and Sg_{t-1} represent groundwater storage at the current and previous time steps, respectively. More details about the abcd water balance model are in [8] and [5]. Equations (1) and (2) produce the streamflow output for t times. We denote this variable by $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_t, \ldots, \tilde{y}_T\}$. Starting from this result, the ABC is used as hydrologic post-processor. Hydrologic post-processing works directly on hydrologic model outputs by using a statistical model to represent the relationship between model outputs and corresponding observations. It serves the purpose of removing model biases from all upstream uncertainty sources. In this paper, we use the ABC to estimate the parameters of the linear model

$$y_t = \beta_0 + \beta_1 \tilde{y}_t + \varepsilon_t, \tag{3}$$

where y_t is the observed streamflow at time t, β_0 and β_1 are parameters, $\tilde{y}_t$ is the output form the abcd model (equations (1) and (2)). The random variable ε_t is the error term in the model, representing random fluctuations, i.e. the effect of factor outside of our control or measurement, such that $\varepsilon_t \sim N(0, \sigma^2)$, i.i.d. Specifically, the ABC produces draws from an approximation of the posterior distribution of $\theta = (\beta_0, \beta_1, \sigma^2)$, i.e.

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta),$$

where $p(\theta)$ is the prior, $p(\mathbf{y}|\theta)$ is the distribution of $\mathbf{y}$ conditional on the parameters. Even though we can use ABC with intractable likelihood $\ell(\theta|\mathbf{y})$, we must be able to simulate data from $p(\theta)$ and $p(\mathbf{y}|\theta)$. We assume flat normal priors for β_0 and β_1 and $Y_t|\theta \sim N(\mu_t = \beta_0 + \beta_1 \tilde{y}_t, \sigma^2)$. The pseudo code for the ABC is summarized in Algorithm 1. Algorithm 1 thus samples θ and pseudo-data $\mathbf{z}$ from the joint posterior:

$$p_\varepsilon(\theta, \mathbf{z}|\eta(\mathbf{y})) = \frac{p(\theta)p(\mathbf{z}|\theta)\mathbb{I}(\mathbf{z})}{\int_\Theta \int_\mathbf{z} p(\theta)p(\mathbf{z}|\theta)\mathbb{I}(\mathbf{z})d\Theta d\mathbf{z}} \tag{4}$$

The ABF produces the approximate predictive uncentanty formally defined as

Algorithm 1 ABC accept/reject algorithm

```
1: $\theta^i$, $i = 1,\ldots,N$ from $p(\theta)$
2: $\mathbf{z}^i = (z_1^i, z_2^i, \ldots, z_T^i)^\top$, $i = 1,\ldots,N$, from the likelihood, $p(\cdot|\theta^i)$
3: Select $\theta^i$ such that:
       $d\{\eta(\mathbf{y}), \eta(\mathbf{z}^i)\} \leq \varepsilon$
   where $\eta(\cdot)$ is a vector statistic, $d\{\cdot\}$ is a distance criterion, and, given $N$, the tolarence level $\varepsilon$ is chosen to be small.
```

$$g(y_{T+1}|\mathbf{y}) = \int_{\Theta} p(y_{T+1}|\theta, \mathbf{y}) p_{\varepsilon}(\theta|\eta(\mathbf{y})) d\theta, \tag{5}$$

where $p_{\varepsilon}(\theta|\eta(\mathbf{y})) = \int_{\mathbf{z}} p_{\varepsilon}(\theta, \mathbf{z}|\eta(\mathbf{y})) d\mathbf{z}$.

3 Streamflow data analysis

We use monthly data of mean areal precipitation, mean areal potential evaporation, and so the other variables of the abcd model, from the Aipe river basin at Huila, Colombia, that is a tropical basin described in the study by [3]. Fig. 1 represents some characteristics of the hydrological behavior of the Aipe river basin. We ap-

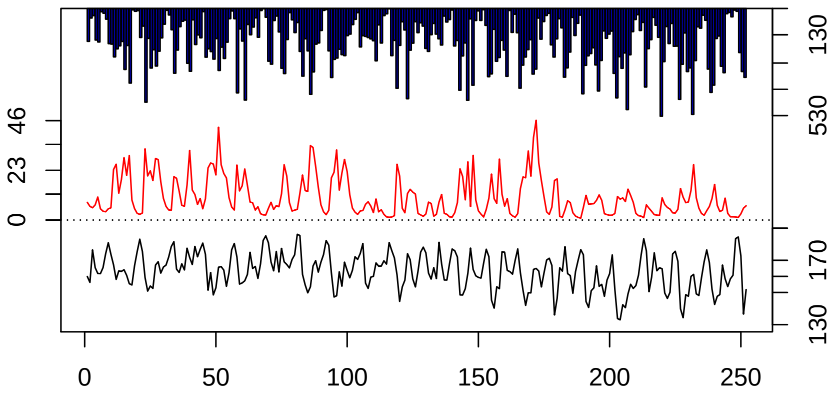


Fig. 1 Monthly time series of Aipe river basin. The blue histogram corresponds to the rainfall (mm), the red line corresponds to the runoff $(m^3 s^{-1})$ and the black line corresponds to the potential evaporation (mm).

ply both the ABC and the MCMC algorithm to obtain parameters estimation of the model (3). Moreover, the ABF is used to assess the predictive uncertainty and compared to the MCMC predictive distribution. The MCMC algorithm is used as a benchmark, since it takes advantage of the likelihood which, for the model we are dealing with, is tractable. To produce the results for the ABF we set the Euclidean distance and choose the mean and the standard deviation as sufficient statistics. In Fig. 2 we show the results of the MCMC and ABC approaches. Although the MCMC and ABC posteriors for both the elements of $\theta = (\beta_0, \beta_1)$ are quite a different one from the other (panel on the left and in the middle), the predictive distributions are quite close and similar.

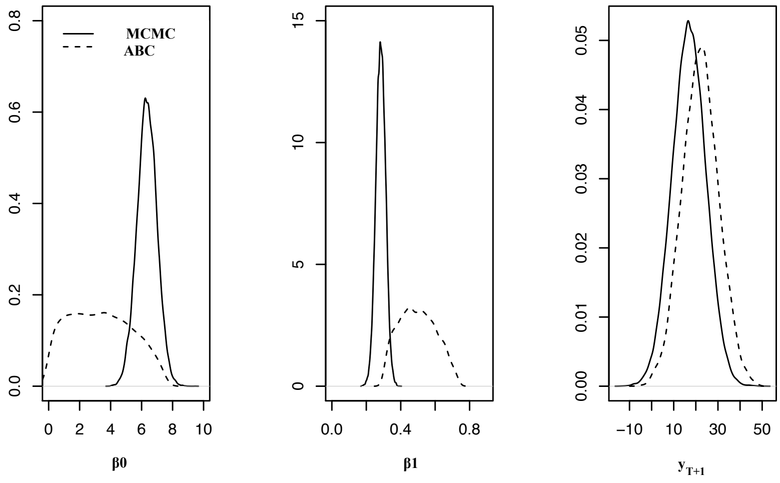


Fig. 2 Marginal posteriors fro the parameters, both considering the MCMC and ABC methods for the monthly time series of Aipe river basin (on the left and in the middle). Predictive density functions both MCMC and ABC (on the right).

4 Conclusion

In this paper, we discuss the use of the Approximate Bayesian Forecasting for hydrological models. The advantage of this tech niche is the applicability for intractable likelihood. This characteristic can make this model very appealing in hydrological forecasting. Even though the ABC seems inappropriate for parameter estimation (probably due to the choice of sufficient statistics) it shows good performance (similar to the MCMC algorithm) in terms of prediction.

References

[1] Csilléry K., François O. and Blum M.G.B. Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3(3):475–479, 2012.

[2] Frazier D. T., Maneesoonthorn W., Martin G. M. , and McCabe B. P.M.. Approximate bayesian forecasting. *arXiv preprint arXiv:1712.07750*, 2017.

[3] Labrador A. F., Zúñiga J. M, and Romero J.. Development of a model for integral planning of water resources in Aipe catchment, Huila, Colombia. *Revista Ingeniería y Región.*, 15(1):23–35, 2016.

[4] Marin J.M., Pudlo P., Robert C.P., and Ryder R.J.. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[5] Martinez G. F. and Gupta H. V.. Toward improved identification of hydrological models: A diagnostic evaluation of the "abcd" monthly water balance model for the conterminous united states. *Water Resources Research*, 46(8): W08507, 2010.

[6] Pritchard J. K., Seielstad M. T., Perez-Lezaun A., and Feldman M. W. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[7] Tavaré S., Balding D. J, Griffiths R. C, and Donnelly P.. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

[8] Harold Thomas. Improved methods for national water assessment, water resources contract: WR15249270. Technical report, Harvard University, Cambridge, 1981.

[9] Todini E.. A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, 6(2): 123–137, jun 2008.

[10] Turner B. M. and Van Zandt T.. A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, 2012.

[11] Vrugt J. A. and Sadegh M. Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, 49: 4335–4345, 2013.

Customer Churn prediction based on eXtreme Gradient Boosting classifier

Previsione della probabilità di non ritorno della clientela attraverso il metodo di classificazione eXtreme Gradient Boosting

Mohammed Hassan Elbedawi Omar and Matteo Borrotti

Abstract Nowadays, Machine Learning (ML) is a hot topic in many different fields. Marketing is one of the best sectors in which ML is giving more advantages. In this field, customer retention models (churn models) aim to identify early churn signals and recognize customers with an increased likelihood to leave voluntarily. Churn problems fit in the classification framework, and several ML approaches have been tested. In this work, we apply an innovative classification approach, eXtreme Gradient Boosting (XGBoost). XGBoost demostrated to be a powerful technique for churn modelling purpose applied to the retail sector.

Abstract *Al giorno d'oggi, il Machine Learning (ML) è un argomento estremamente importante in differenti settori. Ad esempio, il marketing rappresenta uno dei settori più vivi per l'applicazione di metodi di ML. In questo settore, i modelli di customer retention (modelli di churn) analizzano il comportamento dei clienti per individuare quali di questi non torneranno a effettuare acquisti. Questo problema può essere tradotto in un problema di classificazione e molti modelli di ML sono stati testati. In questo lavoro applicheremo un innovativo approccio di classificazione, eXtreme Gradient Boosting (XGBoost) al settore del commercio al dettaglio. Dai risultati ottenuti, si può notare che XGBoost può essere considerato come una tecnica molto efficace per i modelli di churn.*

Key words: churn, classification, XGBoost, boosting

Mohammed Hassan Elbedawi Omar
Energia Crescente S.r.l., Piazza Missori 2, Milano, e-mail: mohammedhassan.omar@en-cre.it

Matteo Borrotti
Energia Crescente S.r.l., Piazza Missori 2, Milano, e-mail: matteo.borrotti@en-cre.it
Institute of Applied Mathematics and Information Technology (IMATI-CNR), Via Alfonso Corti 12, Milano.

1 Introduction

Machine learning (ML) is gaining momentum due to a virtually unlimited number of possible uses and applications. ML is capable to produce models that can analyse bigger and more complex data and deliver accurate insights in order to identify profitable opportunities or to avoid unknown risks. As more data becomes available, more ambitious problems can be tackled. ML is widely used in many fields, such as: recommender systems, credit scoring, fraud detection, drug design, and many other applications.

An important sector where ML is widely applied is marketing. An important topic is related to customer segmentation [6]. In customer segmentation, clustering approaches are used to group customers based on their purchase behaviour. Another important ML application is customer retention. The cost of customer acquisition is much greater than the cost of customer retention and companies are interested in improving customer retention models (churn models). Churn models aim to identify early churn signals and recognize customers with an increased likelihood to leave voluntarily [8, 4].

In this work, we evaluate the performance of a novel approach, eXtreme Gradient Boosting (XGBoost) [2] classifier, on the churn problem. XGBoost is a scalable machine learning system for tree boosting. XGBoost differs from classical tree boosting algorithms for handling sparse data and a theoretically justified weighted quantile sketch procedure, which enables handling instance weights in approximate tree learning. XGBoost algorithm is compared with a classical Decision Tree classifier [7]. The two algorithms are applied to the problem of customers churning in the retail sector. XGBoost outperforms the Decision Tree classifier demonstrating to be a promising approach for customer retention models.

2 eXtreme Gradient Boosting (XGBoost)

The XGBoost algorithm [2] is based on the Gradient Boosted Decision Tree (GBDT) [3]. The following description is based on the work of [9]. XGBoost efficiently deals with sparse data and is suitable for large-scale dataset since implements distrubuted and parallel computing flexibility. XGBoost estimates the target feature by a series of decision tree and defining quantised weight for each leaf node. The prediction function is defined as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \tag{1}$$

where $\hat{y}_i$ is the predicted class of the $i-th$ observation, $\mathbf{x}_i$ is the corresponding feature vector, K is the total number of decision trees. The function f_k is defined as

$$f_k(\mathbf{x}_i) = \omega_{q_k(\mathbf{x}_i)}, \tag{2}$$

where $q_k(\mathbf{x}_i)$ is the structure function of the $k-th$ decision tree that map $\mathbf{x}_i$ to the corresponding leaf node and ω is the vector of the quantised weight.

The accuracy and complexity of the model are taken into account by a regularisation term added to the loss function. The learning process is based on the minimisation of the following loss function:

$$L_t = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

where l is the loss function. The loss function measures the differences between the observed and predicted classes. Ω denotes the complexity of the model.

3 Experimental settings

We consider a dataset composed of 7013 customers belonging to a pet shop. Sensible data were anonymised for privacy purpose. For each customer, the following features are considered: (1) Total net amount of money spent in the considered period, (2) Total amount of discount in the considered period, (3) Number of receipts in the considered period, (4) Number of days since last purchase, (5) Number of days between first and last purchases in the considered period, (6) Average number of days between purchases and (7) Label that is equal to 1 if the customer made at least a purchase in a precise period or 0 otherwise (target feature). It should be noticed that features number (1), (3) and (4) correspond to the Recency (R), Frequency (F) and Monetary (M) features of the well-known RFM model [5].

Features were computed considering the transactions from 29^{th} October 2017 to 21^{th} January 2018. The target feature was identified considering the transactions from 22^{th} January 2018 to 2^{nd} February 2018.

The dataset is divided in three parts: 65% training set, 20% validation set and 15% test set. The validation set is used for parameters optimisation purposes. The Decision Tree and XGBoost's parameters considered for optimisation are the maximum depth of a tree (*max_depth*: $\{2,6,10\}$) and the number of trees (*n_estimators*: $\{100,500,900\}$). For both algorithms, the maximum number of iterations (*nrounds*) is fixed to 100. The XGBoost's learning rate (*eta*) is fixed to 0.01 to avoid early convergence.

XGBoost and Decision Tree were implemented in Python 3.5.4 using XGBoost package version 0.7 and scikit-learn package version 0.19.1.

3.1 Performance metrics

The XGBoost and Decision Tree's performance were evaluated using two metrics: accuracy and logarithmic loss (log-loss) [10]. The confusion matrix (see Table 1) is also used to analysed the general behaviour of the approach.

Accuracy is defined as $Accuracy = \frac{T_p+T_n}{T_p+F_p+T_n+F_n}$.

Table 1 Confusion matrix.

	Observed class 1	Observed class 0
Predicted class 1	True positive (T_p)	False positive (F_p)
Predicted class 0	False negative (F_n)	True negative (T_n)

Log-loss is defined as $l_n = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}log(\hat{y}_{ij})$, where n is the number of observations to be predicted, c is the number of classes (i.e. $c = 2$), y is the observed class and $\hat{y}$ is the predicted probability class. Log-loss should be minimized.

In simple words, accuracy is the count of predicted classes correctly classified among the observed ones. Log-loss takes into account the uncertainty of the predicted class based on how much it varies from the observed class.

These metrics were used to compare XGBoost and Decision Tree on the validation set. Furthermore, we considered the Receiver Operating Characteristic (ROC) curve [1] as a performance metric on the test set. Receiver Operating Characteristic (ROC) curve is based on true positive rate, $P(Tp) = \frac{T_p}{T_p+F_n}$, and false positive rate, $P(Fp) = \frac{F_p}{T_n+F_p}$. ROC curve is used to show in a graphical way the trade-off between true positive rate and false positive rate. The area under the ROC curve is a measure of accuracy.

4 Results

XGBoost and Decision Tree were compared on the validation set. Table 2 shows the performance metrics for all the possible parameters configurations considered in this study. Decision Tree has the best performance with number of trees (*n_estimators*) equal to 500 and maximum depth of a tree (*max_depth*) equal to 10. More precisely, an accuracy of 0.834 and log-loss 0.396. Decision Tree has a similar behaviour also increasing the number of trees. XGBoost reaches the best accuracy (0.895) and log-loss (0.317) with number of trees (*n_estimators*) equal to 900 and maximum depth of a tree (*max_depth*) equal to 10. XGBoost outperforms Decision Tree in all parameters configurations except when the number of trees (*n_estimators*) is set to 100.

Given the obtained results, XGBoost were deplyed on the test set with the best configuration previously found (*n_estimators* = 900 and *max_depth* = 10). Table 3

Table 2 Performance results of Decision Tree classifier on the Validation set.

		Decision Tree		**XGBoost**	
n_estimators	*max_depth*	*Accuracy*	*Log-loss*	*Accuracy*	*Log-loss*
100	2	0.732	0.549	0.727	0.575
100	6	0.756	0.508	0.761	0.543
100	10	0.832	0.396	0.832	0.483
500	2	0.735	0.549	0.743	0.536
500	6	0.757	0.507	0.780	0.462
500	10	0.834	0.396	0.870	0.346
900	2	0.733	0.549	0.743	0.533
900	6	0.756	0.507	0.792	0.437
900	10	0.833	0.396	**0.895**	**0.317**

shows the obtained confusion matrix. The confusion matrix highlights an issue on the prediction of class 0 (not churned people). This issue arises when the two classes are unbalanced. An accuracy of 0.734 and log-loss of 0.567 were obtained. The two metrics exhibit a deterioration from validation set to test set. This deterioration shows a low generalization power of the XGBoost with the parameters configuration selected.

Table 3 Confusion matrix of the XGBoost Classifier on the Test set.

	Observed class 1	Observed class 0
Predicted class 1	557	97
Predicted class 0	169	178

The ROC curve shows a good performance in terms of area under the curve, which is equal to 0.744.

5 Conclusions and future work

In this work, XGBoost is applied to the customer retention (churn) modelling problems. Churn models are appealing tools for the marketing sector. The novel classifier is compared with the Decision Tree. The two approaches were tested on a dataset related to retail sector. From the empirical study, XGBoost outperforms Decision Tree classifier in all the parameters configurations tested except when the number of trees (*n_estimators*) is set to 100.

The best XGBoost's configuration (*n_estimators* = 900 and *max_depth* = 10) was evaluated on the Test set. XGBoost confirms the good results obtained in the validation set but some issues arise. First of all, the target feature was identified considering the transactions from 22^{th} January 2018 to 2^{nd} February 2018. A sensitive

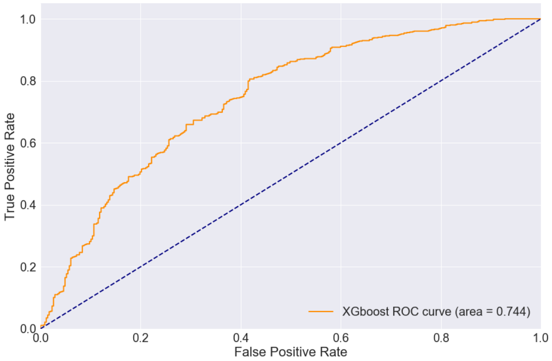


Fig. 1 ROC curve of the XGBoost classifier on the Test set.

analysis on the time window used for target feature definition should be done in oder to improve prediction capability. Additionally, XGboost should be compared with more classification approaches in order to understand the main advantages and disadvantages. A wider parameter analysis needs to be carried out to improve the generalization power. Techniques for unbalanced data should be considered to avoid misclassification problems.

Given that, XGBoost is a promising approach for customer retention modelling problems.

References

1. Brandley, A.P.: The use of the area under the ROC curve in the evaluation of Machine Learning algorithms. Pattern Recognition, **30**, 1145–1159 (1997)
2. Chen, T., Guestrin, C.: XGBoost: a Scalable Tree Boosting system. ArXiv, 1–10 (2016)
3. Friedman, J. Greedy function approximation: A Gradient Boosting machine. Annals of Statistics, **30**, 11891232 (2001)
4. García, D.L., Nebor, Á., Vellido, A.: Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems. **51**, 719–774 (2017)
5. Hughes, A.M.: Strategic database marketing. Probus Publishing Company, Chicago (1994)
6. Ngai, E.W.T., Xiu, L., Chau, D.C.K.: Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, **30**, 1145–1159 (1997)
7. Safavian, R.S., Landgrebe, D.: A survey of Decision Tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, **21**, 660–674 (2016)
8. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.Ch.: A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory. **55**, 1–9 (2015)
9. Wang, S., Dong, P., Tian, Y.: A novel method of statistical line loss estimation for distribution feeders based on feeder cluster and modified XGBoost. Energies, **10**, 1–17 (2017)
10. Whitehill, J.: Climbing the Kaggle leaderboard by exploiting the log-loss Oracle. ArXiv, 1–9 (2017)

HPC-accelerated Approximate Bayesian Computation for Biological Science

Ritabrata Dutta

Abstract Approximate Bayesian computation (ABC) provides us a rigorous tool to perform parameter inference for models without an easily accessible likelihood function. Here we give a short introduction to ABC, focusing on applications in biological science: estimation of parameters of an epidemiological spreading process on a network and a numerical platelets deposition model. Furthermore, we introduce users to a Python suite implementing ABC algorithms, with optimal use of high performance computing (HPC) facilities.

Key words: Approximate Bayesian Computation, Biological science, ABCpy

Introduction

With the recent innovations in biological science, we are increasingly facing large datasets of varied type and more realistic but complex models of natural phenomenon. This trend has led to a scenario where we do not easily have a likelihood function which is available in closed form and thus easy to evaluate at any given point (as required by most Monte Carlo and Markov chain Monte Carlo methods). Thus, traditional likelihood based inference, as Maximum likelihood or Bayesian methodology, is not possible. Still, if from the complex model, given values of the parameters that index it, we can forward simulate pseudo-dataset, a new methodology becomes available, namely Approximate Bayesian Computation (ABC). Models that have this possibility of forward simulation are known as simulator-based models and are becoming more and more popular in diverse fields of science [Martinez et al., 2016, Turchin et al., 2013, Schaye et al., 2015]; just restricting to the biological domain we can find many examples: evolution of genomes [Marttinen

Ritabrata Dutta
Institute of Computational Science, Università della Svizzera italiana, Switzerland, e-mail: duttar@usi.ch

et al., 2015], numerical model of platelet deposition [Chopard et al., 2017], demographic spread of a species [Excoffier et al., 2013] among many. Research in statistical science in the last decade or so, has illustrateted how ABC can be a tool to infer and calibrate the parameters of these models.

The fundamental rejection ABC sampling scheme iterates between three step: First a pseudo-dataset, $\boldsymbol{x}^{\text{sim}}$, is simulated from the simulator-based model $\mathscr{M}(\boldsymbol{\phi})$ for a fixed parameter value of $\boldsymbol{\phi}$. Then we compute a measure of the closeness between $\boldsymbol{x}^{\text{sim}}$ and $\boldsymbol{x}^0$, the observed dataset, using a pre-defined discrepancy measure $d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0)$. Finally, based on this discrepancy measure, ABC accepts the parameter value $\boldsymbol{\phi}$ when $d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0)$ is less than a pre-specified threshold value ε.

Following this ABC sampler, the intractable likelihood $\mathscr{L}(\boldsymbol{\phi})$ is approximated by $\mathscr{L}_{d,\varepsilon}(\boldsymbol{\phi})$ for some $\varepsilon > 0$, where

$$\mathscr{L}_{d,\varepsilon}(\boldsymbol{\phi}) \propto P(d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0) < \varepsilon) \tag{1}$$

and, as a consequence, the accepted parameters follow the posterior distribution of $\boldsymbol{\phi}$ conditional on $d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0) < \varepsilon$:

$$p_{d,\varepsilon}(\boldsymbol{\phi}|\boldsymbol{x}^0) \propto P(d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0) < \varepsilon)\pi(\boldsymbol{\phi}).$$

For a better approximation of the likelihood function, computationally efficient sequential ABC algorithms [Marin et al., 2012, Lenormand et al., 2013, Albert et al., 2015] decrease the value of the threshold ε adaptively while exploring the parameter space.

The crucial aspect for a good ABC approximation to the likelihood function is the choice of the summary statistics, as we define the discrepancy measure between $\boldsymbol{x}^{\text{sim}}$ and $\boldsymbol{x}^0$ through a distance between the extracted summary statistics from $\boldsymbol{x}^{\text{sim}}$ and $\boldsymbol{x}^0$. Knowledge domain driven summary statistics are normally chosen keeping in mind that we want to minimize the loss of information on $\boldsymbol{\phi}$ contained in the data through the choice of summary statistics. But one can also rely on automatic summary selection for ABC, thus removing a subjective component in this choice, as described in Fearnhead and Prangle [2012], Pudlo et al. [2015], Jiang et al. [2015] and Gutmann et al. [2017].

ABCpy

ABC provides a tool for statistical inference for simulator-based models, still, the necessity to simulate lots of pseudo-data, makes the algorithm extremely computationally expensive when data-simulation itself is costly. Further, the varied types of data sets available in different domain specific problems have hindered the applicability of ABC algorithms to many applied science domains. Recently, [Dutta et al., 2017a,d], have developed an High Performance Computing framework to efficiently parallelize different ABC algorithms which we believe will be extremely beneficial for inferential problems across different scientific domains.

ABC and HPC were first brought together in the ABC-sysbio package for the systems biology community, where sequential Monte Carlo ABC (ABC-SMC) [T. Toni, 2009] algorithm was efficiently parallelized using graphics processing units (GPUs). The goal of ABCpy was to overcome the need for users to have knowledge of parallel programming, as is required for using ABC-sysbio [Liepe et al., 2010], and also to make a software package available for scientists across domains. These objectives were partly addressed by parallelization of ABC-SMC using MPI/OpenMPI Stram et al. [2015], and by making ABC-SMC available for the astronomical community Jennings and Madigan [2016]. Regardless of these advances, a recent ABC review article Lintusaari et al. [2017] highlights the depth and breadth of available ABC algorithms, which can be optimally efficient only via parallelization in an HPC environment Kulakova et al. [2016], Chiachio et al. [2014]. These developments emphasized the need of a generalized HPC supported platform for efficient ABC algorithms, which can be parallelized on multi-processor computers or computing clusters and is accessible to a broad range of scientists.

ABCpy addressed this need for an user-friendly scientific library of ABC algorithms, which is written in Python and designed in a highly modular fashion. Existing ABC software suites are mainly domain-specific and optimized for a narrower class of problems. Modularity of ABCpy makes it intuitive to use and easy to extend. Further, it enables users to run ABC sampling schemes in parallel without too much re-factoring of existing code. ABCpy includes likelihood free inference schemes, both based on discrepancy measures and approximate likelihood, providing a complete environment to develop new ABC algorithms.

Illustrative Applications

To highlight the versatility of ABC and ABCpy in diverse applied problems, we point the interested reader to two recent research papers with biological applications in mind: a) estimation of parameters of an epidemiological spreading process on a contact network[Dutta et al., 2017c] and b) estimation of parameters of a numerical platelets deposition model [Dutta et al., 2017b].

Epidemics on a Contact Network

Infectious diseases are studied to understand their spreading mechanisms, to evaluate control strategies and to predict the risk and course of future outbreaks. Because people only interact with a small number of individuals, and because the structure of these inter- actions matters for spreading processes, the pairwise relationships between individuals in a population can be usefully represented by a network. For modeling the spread of infections on a human contact network, we consider a simple spreading process, i.e., the standard susceptible-infected (SI) process with unit infectivity on a fixed network [Zhou et al., 2006, Staples et al., 2016]. In this model, there are only two states, susceptible and infected, and this process is suitable for modeling the spread of pathogens in contact networks because a single successful exposure can be sufficient for transmission. In this process, at each time step, each infected node chooses one of its neighbors with equal probability regardless of their status (susceptible or infected), and if this neighboring node is susceptible, the node successfully infects it with probability θ. We denote this model by $\mathscr{M}_S$ and parametrize it in terms of the spreading rate θ and of the seed node (the node representing the first infected person) n_{SN}. For given values of these two parameters, $n_{\text{SN}} = n^*_{\text{SN}}$ and $\theta = \theta^*$, we can forward simulate the evolving epidemic over time using the $\mathscr{M}_S$ model as

$$\mathscr{M}_S[n^*_{\text{SN}}, \theta^*] \rightarrow \{\mathbb{N}_{\mathbb{I}}(t), t = 0, \ldots, T\}, \tag{2}$$

where $\mathbb{N}_{\mathbb{I}}(t)$ is a list of infected nodes at time t. We simulated an epidemic of a disease using the above simple contagion process in an Indian village contact network[Banerjee et al., 2013]. The network has 354 nodes and 1541 edges, representing 354 villagers and reported contacts and social relationships among them. The epidemic is simulated using $\theta^0 = 0.3$, $n^0_{\text{SN}} = 70$, and the observed dataset $\boldsymbol{x}^0$ is the infected nodes $\mathbb{N}_{\mathbb{I}}(t)$ for $t = t_0, \ldots, T$ with $t_0 = 20$ and $T = 70$. The marginal posterior distributions and the Bayes estimates of (θ, n_{SN}) are show in Figure 1. The inferred posterior distributions for the epidemics on the Indian village contact network, is concentrated around the true parameter values. The Bayes estimates are also in a very small neighborhood of the true value, specifically the estimated seed-node $(\hat{n}_{\text{SN}})$ has a shortest path distance of 1 from n^0_{SN} in both the cases.

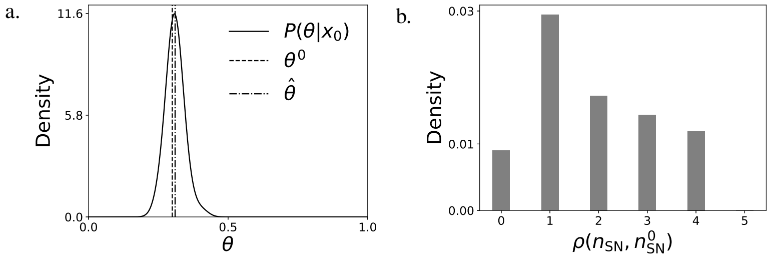


Fig. 1: **Simple contagion model on Indian village contact network.** Panel **a** shows the density of the inferred marginal posterior distribution and Bayes estimate of θ, given $\boldsymbol{x}^0$, the epidemics on the Indian village contact network. Panel **b** displays the average marginal posterior distribution at different distances from the true seed-node n^0_{SN}. The shortest path length distance between $n^0_{\text{SN}} = 70$ and $\hat{n}_{\text{SN}} = 59$ is 1.

For details on how the inference was performed via ABC and ABCpy, we direct readers to Dutta et al. [2017]. We can further extend this inferential approach to any complex spreading processes on a network, e.g. inference of parameters of complex contagion model representing a disinformation campaign on a social network is reported in Dutta et al. [2017].

Platelet Deposition Model

Chopard et al. [2015], Chopard et al. [2017] has recently developed a numerical model that quantitatively describes how platelets in a shear flow adhere and aggregate on a deposition surface. Five parameters specify the deposition process and are relevant for a biomedical understanding of the phenomena. An experimental observations can be collected from a patient, at time intervals, on the average size of the aggregation clusters, their number per mm^2, the number of platelets and the ones activated per $\mu\ell$ still in suspension. In Dutta et al. [2017b], we have demonstrated that approximate Bayesian computation (ABC) can be used to automatically explore the parameter space of this numerical model. To illustrate the performance of ABC, in Figure 2, we show the inferred posterior distribution of the parameters (adhesion rate p_{Ad}, the aggregation rates p_{Ag} and p_T, the deposition rate of albumin p_F, and the attenuation factor a_T) of the platelet deposition model. For details on the specific model and on how the inference was performed via ABC, we direct readers to Chopard et al. [2017] and Dutta et al. [2017b] correspondingly.

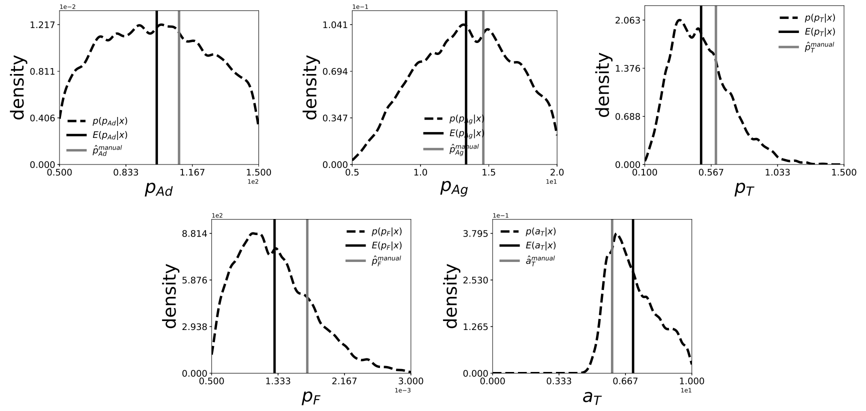


Fig. 2: Marginal posterior distribution (black-dashed) and Bayes Estimate (back-solid) of $(p_{Ad}, p_{Ag}, p_T, p_F, a_T)$ for collective data set generated from of 7 patients. The smoothed marginal distribution is created by a Gaussian-kernel density estimator on 5000 samples drawn from the posterior distribution using Simulated annealing approximate Bayesian computation [Albert et al., 2015]. The (gray-solid) line indicates the manually estimated values of the parameters as in [Chopard et al., 2017].

The proposed approach can be applied patient per patient, in a systematic way, without the bias of a human operator. In addition, the approach is computationally fast enough to provide results in an acceptable time for contributing to a new medical diagnosis, by giving data that no other known method can provide.

Conclusion

We would like to stress here the fact that ABC inference scheme provides not only a point estimate of the parameters of interest but also their entire (approximated) posterior distribution thus allowing for uncertainty quantification: the higher the variability of the posterior distribution the higher the uncertainty inherent in the inferential scheme. Via the ABC approximated posterior one can then construct credible intervals and perform hypothesis testing. Furthermore ABC allows to compare possible alternative models by simply adding, to the three steps Rejection ABC scheme illustrated above, an additional initial layer where first a model index is sampled from the model prior distribution and then, once a model has been selected a regular ABC scheme within that model is performed. For details on ABC model selection via random forest approach see Pudlo et al. [2015].

Acknowledgements The research was supported by Swiss National Science Foundation Grant No. 105218 163196 (Statistical Inference on Large-Scale Mechanistic Network Models). We also thank Dr. Marcel Schoengens, CSCS, ETH Zurich for helps regarding HPC and Swiss National Super Computing Center for providing computing resources.

References

Carlo Albert, Hans R. Künsch, and Andreas Scheidegger. A simulated annealing approach to approximate Bayesian computations. *Statistics and Computing*, 25: 1217–1232, 2015.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.

Manuel Chiachio, James L. Beck, Juan Chiachio, and Guillermo Rus. Approximate Bayesian computation by subset simulation. *SIAM J. Sci. Comput.*, 36(3):A1339–A1358, 2014.

B. Chopard, D. Ribeiro de Sousa, J. Latt, F. Dubois, C. Yourassowsky, P. Van Antwerpen, O. Eker, L. Vanhamme, D. Perez-Morga, G. Courbebaisse, and K. Zouaoui Boudjeltia. A physical description of the adhesion and aggregation of platelets. *ArXiv e-prints*, 2015.

Bastien Chopard, Daniel Ribeiro de Sousa, Jonas Lätt, Lampros Mountrakis, Frank Dubois, Catherine Yourassowsky, Pierre Van Antwerpen, Omer Eker, Luc Vanhamme, David Perez-Morga, et al. A physical description of the adhesion and aggregation of platelets. *Royal Society Open Science*, 4(4):170219, 2017.

R. Dutta, A. Mira, and J.-P. Onnela. Bayesian Inference of Spreading Processes on Networks. *ArXiv e-prints*, September 2017.

R Dutta, M Schoengens, J.P. Onnela, and Antonietta Mira. ABCpy: A user-friendly, extensible, and parallel library for approximate Bayesian computation. In *Proceedings of the Platform for Advanced Scientific Computing Conference*. ACM, June 2017a.

Ritabrata Dutta, Bastien Chopard, Jonas Lätt, Frank Dubois, Karim Zouaoui Boudjeltia, and Antonietta Mira. Parameter estimation of platelets deposition: Approximate bayesian computation with high performance computing. *arXiv preprint arXiv:1710.01054*, 2017b.

Ritabrata Dutta, Antonietta Mira, and Jukka-Pekka Onnela. Bayesian inference of spreading processes on network. *arXiv preprint arXiv:1709.08862*, 2017c.

Ritabrata Dutta, Marcel Schoengens, Avinash Ummadisingu, Jukka-Pekka Onnela, and Antonietta Mira. Abcpy: A high-performance computing perspective to approximate bayesian computation. *arXiv preprint arXiv:1711.04694*, 2017d.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C Sousa, and Matthieu Foll. Robust demographic inference from genomic and snp data. *PLoS genetics*, 9(10):e1003905, 2013.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74 (3):419–474, 2012.

Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, pages 1–15, 2017.

E. Jennings and M. Madigan. astroABC: An Approximate Bayesian Computation Sequential Monte Carlo sampler for cosmological parameter estimation. *ArXiv:1608.07606*, 2016.

Bai Jiang, Tung-yu Wu, Charles Zheng, and Wing H Wong. Learning summary statistic for approximate Bayesian computation via deep neural network. *arXiv preprint arXiv:1510.02175*, 2015.

Lina Kulakova, Panagiotis Angelikopoulos, Panagiotis E. Hadjidoukas, Costas Papadimitriou, and Petros Koumoutsakos. Approximate Bayesian computation for granular and molecular dynamics simulations. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '16, pages 4:1–4:12. ACM, 2016. doi: 10.1145/2929908.2929918.

Maxime Lenormand, Franck Jabot, and Guillaume Deffuant. Adaptive approximate bayesian computation for complex models. *Computational Statistics*, 28 (6):2777–2796, 2013.

Juliane Liepe, Chris Barnes, Erika Cule, Kamil Erguler, Paul Kirk, Tina Toni, and Michael P.H. Stumpf. ABC-SysBio – approximate Bayesian computation in Python with GPU support. *Bioinformatics*, 26(14):1797–1799, 2010. doi: 10.1093/bioinformatics/btq278.

Jarno Lintusaari, Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1):e66–e82, 2017. doi: 10.1093/sysbio/syw077. URL https://doi.org/10.1093/sysbio/syw077.

Jean-Michel Marin, Pierre Pudlo, ChristianP. Robert, and RobinJ. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6): 1167–1180, 2012. ISSN 0960-3174. doi: 10.1007/s11222-011-9288-2. URL http://dx.doi.org/10.1007/s11222-011-9288-2.

Esteban A. Martinez, Christine A. Muschik, Philipp Schindler, Daniel Nigg, Alexander Erhard, Markus Heyl, Philipp Hauke, Marcello Dalmonte, Thomas Monz, Peter Zoller, and Rainer Blatt. Real-time dynamics of lattice gauge theories with a few-qubit quantum computer. *Nature*, 534(7608):516–519, 2016. doi: 10.1038/nature18318.

Pekka Marttinen, Nicholas J Croucher, Michael U Gutmann, Jukka Corander, and William P Hanage. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, 1(5), 2015.

Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till Sawala, Peter A. Thomas, and James Trayford. The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554, 2015. doi: 10.1093/mnras/stu2058.

Patrick Staples, Mélanie Prague, De Gruttola Victor, and Jukka-Pekka Onnela. Leveraging contact network information in clustered randomized trials of infectious processes. *arXiv preprint arXiv:1610.00039*, 2016. URL https://arxiv.org/abs/1610.00039.

Alexander H. Stram, Paul Marjoram, and Gary K. Chen. al3c: high-performance software for parameter inference using Approximate Bayesian Computation. *Bioinformatics*, 31(21):3549–3551, 2015. doi: 10.1093/bioinformatics/btv393.

M. S. T. Toni. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 31(6):187–202, 2009.

Peter Turchin, Thomas E. Currie, Edward A. L. Turner, and Sergey Gavrilets. War, space, and the evolution of old world complex societies. *Proceedings of the National Academy of Sciences*, 110(41):16384–16389, 2013. doi: 10.1073/pnas.1308825110.

Tao Zhou, Jian-Guo Liu, Wen-Jie Bai, Guanrong Chen, and Bing-Hong Wang. Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity. *Physical Review E*, 74(5):056109, 2006. doi: 10.1103/PhysRevE.74.056109. URL https://doi.org/10.1103/PhysRevE.74.056109.

PC Algorithm for Gaussian Copula Data

L'algoritmo PC per dati generati da copula gaussiana

Vincenzina Vitale and Paola Vicard

Abstract The PC algorithm is the most popular algorithm used to infer the structure of a Bayesian network directly from data. For Gaussian distributions, it infers the network structure using conditional independence tests based on Pearson correlation coefficients. Here, we propose two modified versions of PC, the R-vine PC and D-vine PC algorithms, suitable for elliptical copula data. The correlation matrix is inferred by means of the estimated structure and parameters of a regular vine. Simulation results are provided, showing the very good performance of the proposed algorithms with respect to their main competitors.

Abstract *L'algoritmo PC è l'algoritmo piu diffuso per l'apprendimento della struttura di una rete bayesiana direttamente dai dati. Quando i dati sono gaussiani, esso apprende la struttura della rete per mezzo di test di indipendenza condizionata basati sui coefficienti di correlazione di Pearson. In questo lavoro, proponiamo due versioni modificate del PC, gli algoritmi R-vine PC e D-vine PC, validi per dati generati da copule ellittiche. La matrice di correlazione è calcolata sulla base della struttura e dei parametri stimati di un regular vine. Vengono forniti i risultati delle simulazioni che mostrano l'ottima performance degli algoritmi qui proposti rispetto ai loro principali competitor.*

Key words: Structural learning, Bayesian networks, R-vines, Gaussian copulae, PC algorithm.

Vincenzina Vitale
Dipartimento di Economia - Università Roma Tre, Via Silvio D'Amico, 77 - 00145 Roma, e-mail: vincenzina.vitale@uniroma3.it

Paola Vicard
Dipartimento di Economia - Università Roma Tre, Via Silvio D'Amico, 77 - 00145 Roma, e-mail: paola.vicard@uniroma3.it

1 Introduction

A Bayesian network (BN,[6]) is a multivariate statistical model satisfying sets of (conditional) independence statements encoded in a *Directed Acyclic graph* (DAG). Each node in the graph represents a random variable while the edges between the nodes represent probabilistic dependencies among the corresponding variables. If there is an arrow from X_i to X_j, X_j is said *child* of X_i and X_i is said *parent* of X_j. The set of parents of X_j in the graph G is denoted by $pa(X_j)$. In a BN each node is associated with a conditional distribution given its parents and the joint distribution can be factorized according to the DAG structure as:

$$p(X_1 \dots X_p) = \prod_{j=1}^{p} p(X_j | pa(X_j)) \tag{1}$$

BN structure can be elicited by the expert knowledge or learnt directly from data by means of structural learning techniques [16]. Among the *constraint-based* algorithms, the most known is the PC algorithm [18]: it estimates the Markov equivalence class of a DAG performing three main steps: i) *skeleton*[1] *identification* by recursively testing marginal and conditional independencies using Pearson correlation coefficients, for a fixed significance level α; ii) v-*structures identification*: an unshielded triple $i - k - j$, such that the pairs i and k and j and k are connected while i and j is not, is oriented as $i \rightarrow k \leftarrow j$ if k is not in the separation sets of nodes X_i and X_j; iii) *orientation of some of the remaining edges* without producing additional *v*-structures and/or directed cycles.

In the BN framework, when the analysed variables are continuous, joint normality[2] is assumed. Unfortunately, in many applied context the normality assumption may not be reasonable. In such cases, copula modeling has become very popular and, recently, there is a growing literature where the theory of copulae and Bayesian networks are combined [10, 14, 8, 15, 11, 2]. Here, a modified version of the PC algorithm suitable for Gaussian and Student t copula distributions is proposed. In particular, we work in the theoretical framework of pair-copula constructions with reference to the subclass of regular vines [4, 12]. From the estimated structure and parameters of a regular vine, we infer the corresponding marginal correlation matrix valid under the assumption that data are drawn from a Gaussian copula family. The correlation coefficients are then used as sufficient statistics in the conditional independence tests implemented in the PC algorithm. Simulations are carried out in order to evaluate the performance of the proposed algorithms and to compare them with the PC and the Rank PC (RPC) algorithms. RPC has been recently introduced by [15] to overcome the normality assumption limitation and can be used for Gaussian copula data.

[1] The skeleton of a DAG is the undirected graph obtained replacing arrows with undirected edges.

[2] In the mixed continous and discrete case, the conditional Gaussian distribution is assumed.

The paper is organized as follows. In Section 2 pair copula construction and regular vines are introduced; in Section 3 the new algorithms are illustrated and simulation results are shown and discussed.

2 Pair copula construction and regular vines

Let F be a n-dimensional distribution function of the random vector $\mathbf{X} = (X_1, \cdots X_n)$ with univariate marginals $F_1 \ldots F_n$. A $n-variate\ copula$ is a multivariate cumulative distribution function (cdf) $C : [0,1]^n \to [0,1]$ with $n \in N$ and uniformly distributed marginals $U(0,1)$ on the interval $[0,1]$. By Sklar theorem [17], every cdf F with marginals $F_1 \ldots F_n$ can be written as:

$$F(\mathbf{x}) = C(F_1(x_1), \ldots, F_n(x_n)) \tag{2}$$

for some appropriate n-dimensional copula C. By copulas, multivariate distribution modeling is split into univariate marginals and dependence structure modeling. While for bidimensional case there is an exaustive literature on bivariate copula families, their extension to multivariate case is not straightforward[3]. In [3, 4] and [14] the decomposition of the multivariate copulae into the product of bivariate ones, known in literature as *pair-copula construction* (PCC), is proposed. Each pair-copula can be selected independently from the others allowing for a great flexibility in dependence modeling. Since in higher dimensions the number of possible pair-copulae constructions grows up significantly, in [3, 4] a graphical representation (called *regular vine*) to organize them, is introduced.

Generally speaking, a regular vine (R-vine) is a sequence of trees whose edges correspond to bivariate copulae; see [14] for a formal definition. A n-dimensional R-vine is a set of $n-1$ trees such that the first tree comprises n nodes, identifying $n-1$ pairs of variables and $n-1$ corresponding edges. Each subsequent tree is derived so that all the edges of tree i turn into nodes of the tree $i+1$; furthermore, two edges in T_i, becoming nodes in T_{i+1}, are joined by an edge in T_{i+1} only if these edges share a common node in T_i. The graphical structures of R-vines allow the specification of all bivariate copulae of the pair copula construction. In particular, each edge corresponds to a bivariate copula density. The copulae defined in the first tree are unconditional copulae while the others are all conditional [4]. The importance of these results arises from the fact that all bivariate copulae can belong to different families and their parameters can be specified independently from each other. Many applications concern a special case of R-vines, the Drawable vines (D-vines), see [1]. D-vines only need the ordering definition of their first tree sequence to completely

[3] Standard multivariate copulae such as Gaussian or Student t lack the flexibility of accurately modeling the dependence structure in higher dimensions.

[4] The copulae of the second tree have only one node as conditioning set, the third two nodes and so on.

identify the structure. Differently, R-vines suffer from the fact that many possible tree sequences can be specified.

Three separate steps have to be done to specify the vine structure and distribution:

1. selecting the structure with all its trees;
2. selecting the appropriate bivariate copula family for each of the $n(n-1)/2$ pair copulae associated with the vine structure;
3. estimating the parameters for each bivariate copula identified at the previous step.

Since the number of possible R-vines on n variables increases exponentially with n, a sequential method has been proposed by [7] and implemented in the `VineCopula R` package. Among the possible copula types, the independence copula can also be chosen by means of a preliminary independence test based on Kendall's tau [9]. The concept of independence copula is strictly connected to that of conditional independence: it allows to reduce the number of parameters to be estimated. In [5] the truncated R-vines, for which independence is assumed for the k last trees, is proposed. More recently, [11] have deeply analysed the truncation procedure showing the relationship between the truncated R-vines and the decomposable graphs.

3 The R-vine and D-vine PC algorithms: simulations and results

Here, we take advantage of the use and properties of regular vines to estimate the BN dependence structure under the assumption of data coming from a Gaussian copula distribution. More precisely, we infer the R-vine (and D-vine) structure together with its copula parameters in order to extract the corresponding marginal correlation matrix. Note that the copula family is limited to Gaussian or Student t case for which the correlation coefficients can be estimated. The last step consists in using the marginal correlation coefficients as sufficient statistics for Pearson correlation tests implemented in the classical version of the PC algorithm. According to these definitions and purposes, we propose four algorithms: the R-vine PC algorithm, the D-vine PC algorithm and their truncated versions respectively. They work along the following four steps:

1. transforming data in pseudo-observations;
2. fitting a R-vine (or D-vine) to the transformed data based on the AIC criterion and the maximum likelihood estimation of copula parameters;
3. inferring the marginal correlation matrix from the estimated R-vine (or D-vine);
4. running the PC algorithm providing, in input, the estimated marginal correlation matrix of the previous step.

All functions used in the first three steps are implemented in the `VineCopula R` package. Regarding the D-vine, the `TSP R` package has been used to determine the order of the nodes of its first tree. The fourth step functions are implemented in the `pcalg R` package.

We argue that the two non truncated algorithms allow the specification of the independence copula as proposed by [9][5]. The procedure of truncation applied in this work follows the approach of [5]. To choose the optimal truncation level, a R function has been written in order to recursively perform the likelihood ratio based test between different levels.

Two random DAGs, one not decomposable and the other decomposable, with sparsity parameter $s = 0.4$ and $s = 0.3$ respectively, are simulated according to the procedure ensuring faithfulness [13]. 250 datasets are drawn from each DAG following a Gaussian copula distribution, fixing $n = 500$ and $\alpha = 0.01$. The *structural Hamming distance* (SHD, [19]) has been computed in order to evaluate the different performances of the algorithms.

Table 1 Simulation results by algorithms

Graph (n=500)	Algorithm	SHD (mean)	SHD(Median)	SHD (s.d.)	SHD (IQR)
Decomposable graph	PC	8,29	8	1,62	2
	RPC	6,43	6	1,14	1
	R-vine PC	6,14	6	1,24	2
	Truncated R-vine PC	6,38	6	1,05	1
	D-vine PC	6,23	6	1,21	2
	Truncated D-vine PC	6,56	6	1,34	1
Non decomposable graph	PC	5,87	6	2,02	2
	RPC	3,69	3	2,35	4
	R-vine PC	2,04	1	1,66	2
	Truncated R-vine PC	2,36	2	1,71	2
	D-vine PC	2,88	3	1,92	3
	Truncated D-vine PC	4,07	4	2,51	4

The simulation results, shown in Tab. 1, are very promising. As expected, under the assumption of Gaussian copula data, the performance of the PC algorithm, in terms of mean value, is worse than all the others. For a non decomposable graph, the performance of R-vine PC algorithm, followed by that of its truncated version is extremely good. The mean value of the errors is the smallest, about 2; if its median value is taken into account, it is equal to 1. As far as variability is concerned, its standard deviation is also very small. With the exception of the truncated D-vine PC algorithm, the proposed algorithms outperform their competitors.

For data generated from a decomposable graph, the differences in performance with respect to the PC algorithm are still evident. The distance between our proposals and the RPC algorithm is less remarkable. The R-vine and D-vine PC algorithms show the smallest mean values but larger variability. The truncated R-vine PC algorithm seems to balance these two aspects. Simulation results clearly show that the undirected graph R-vine is able to capture the underlying dependence structure of data. It considerably increases the capability of the PC to detect the best fitting

[5] The hypothesis test for the independence of pseudo-observations u_1 and u_2 is performed before bivariate copula selection. The independence copula is chosen for a (conditional) pair if the null hypothesis of independence cannot be rejected.

network. The main limitation of the proposed algorithms is that they are applicable only to elliptical copula distributions, restricting the choice of possible copula families. Future research will necessarily concern the definition of a new class of algorithms suitable for non normal data without any restriction to the class of copula families.

References

1. Aas, K., Czado, C., Frigessi, A., Bakken, H.: Pair-copula constructions of multiple dependence. Insurance: Mathematics and economics **44**(2), 182–198 (2009)
2. Bauer, A., Czado, C.: Pair-copula bayesian networks. J. Comput. Graph. Stat. **25**(4), 1248–1271 (2016)
3. Bedford, T., Cooke, R.M.: Probability density decomposition for conditionally dependent random variables modeled by vines. Ann. Math. Artif. Intell. **32**(1), 245–268 (2001)
4. Bedford, T., Cooke, R.M.: Vines: A new graphical model for dependent random variables. Ann. Stat. **30**(4), 1031–1068 (2002)
5. Brechmann, E.C., Czado, C., Aas, K.: Truncated regular vines in high dimensions with application to financial data. Can. J. Stat. **40**(1), 68–85 (2012)
6. Cowell, R.G., Dawid, P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic Networks and Expert Systems. Springer-Verlag, New York. (1999)
7. Dissmann, J., Brechmann, E.C., Czado, C., Kurowicka, D.: Selecting and estimating regular vine copulae and application to financial returns. Comput. Stat. Data. Anal. **59**, 52–69 (2013)
8. Elidan, G.: Copula bayesian networks. In: Advances in neural information processing systems, pp. 559–567 (2010)
9. Genest, C., Favre, A.: Everything you always wanted to know about copula modeling but were afraid to ask. J. Hydrol. Eng. **12**(4), 347–368 (2007)
10. Hanea, A.M., Kurowicka, D., Cooke, R.M.: Hybrid method for quantifying and analyzing bayesian belief nets. Qual. Reliab. Eng. Int. **22**(6), 709–729. (2006)
11. Hobæk Haff, I., Aas, K., Frigessi, A., Lacal, V.: Structure learning in bayesian networks using regular vines. Comput. Stat. Data Anal. **101**(C), 186–208 (2016)
12. Joe, H.: Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. Lecture Notes-Monograph Series pp. 120–141 (1996)
13. Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the pc-algorithm. J. Mach. Learn. Res. **8**, 613–636. (2007)
14. Kurowicka, D., Cooke, R.: Uncertainty Analysis with High Dimensional Dependence Modelling. John Wiley & Sons, Ltd. (2006)
15. Naftali, H., Drton, M.: Pc algorithm for nonparanormal graphical models. J. Mach. Learn. Res. **14**, 3365–3383. (2013)
16. Neapolitan, R.E.: Learning Bayesian Networks. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. (2003)
17. Sklar, A.: Fonctions de répartition ǹ dimensions et leurs marges. Publications de l'Institut de Statistique de L'Université de Paris **8**, 229–231. (1959)
18. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT press, Cambridge, Massachusetts. (2000)
19. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Mach. Learn. **65**(1), 31–78 (2006)

Advances in Clustering Techniques

On the choice of an appropriate bandwidth for modal clustering

Scelta di un appropriato parametro di lisciamento per il clustering modale

Alessandro Casa, José E. Chacón and Giovanna Menardi

Abstract In *modal* clustering framework groups are regarded as the domains of attraction of the modes of probability density function underlying the data. Operationally, to obtain a partition, a nonparametric density estimate is required and kernel density estimator is commonly considered. When resorting to these methods a relevant issue regards the selection of the smoothing parameter governing the shape of the density and hence possibly the modal structure. In this work we propose a criterion to choose the bandwidth, specifically tailored for the clustering problem since based on the minimization of the distance between a partition of the data induced by the kernel estimator and the whole-space partition induced by the true density.
Abstract Nell'ambito del clustering, l'approccio modale associa i gruppi ai domini di attrazione delle mode della funzione di densità sottostante i dati. L'individuazione dei gruppi richiede una stima non parametrica della densità, spesso basata su metodi kernel. Un problema rilevante, a tale scopo, riguarda la selezione del parametro di lisciamento che governa la forma della densità e, di conseguenza, la struttura modale. In questo lavoro si propone un criterio per la selezione del parametro di lisciamento, specificamente orientato al problema del clustering non parametrico e basato sulla minimizzazione di una misura di distanza tra la partizione dei dati indotta da uno stimatore kernel e la partizione dello spazio indotta dalla vera funzione di densità.

Key words: modal clustering, distance in measure, bandwidth selection, kernel density estimator

Alessandro Casa, Giovanna Menardi
Dipartimento di Scienze Statistiche, Università degli Studi di Padova
via C. Battisti 241, 35121, Padova; e-mail: casa@stat.unipd.it, menardi@stat.unipd.it
José E. Chacón
Departamento de Matemáticas, Universidad de Extremadura,
E-06006 Badajoz, Spain; e-mail: jechacon@unex.es

1 Introduction

Distance-based clustering is probably the most common approach to the unsupervised problem of obtaining a partition of a set of data into a number of groups. In spite of an intuitive interpretation and conceptual simplicity, this approach lacks of a 'ground truth', thus preventing the possibility to resort to formal statistical procedures. The density-based approach to cluster analysis overcomes such drawback by providing a formal definition of cluster, based on some specific features of the probability density function assumed to underlie the data. This approach has been developed following two distinct directions. The parametric one hinges on modelling the density function by means of a mixture distribution, where clusters are associated to the mixture components. Readers can refer to [3] for a recent review. This work focuses on the nonparametric - or *modal* - formulation, which is built on the concept of clusters as the "domains of attraction" of the modes of the density underlying the data [7]. The local maxima of the density are regarded to as the archetypes of the clusters, which are represented by the sorrounding regions (see Figure 1 for an illustration). These concepts have been translated into a formal definition of cluster by [1], resorting to notions and tools borrowed from Morse theory (see [2] for an introduction). Operationally, modal clustering has been pursued by two different strands of methods, both based on a preliminary nonparametric estimate of the density. The first strand looks directly for the modes of the estimated density and associates each cluster to the set of points along the steepest ascent path towards a mode, while the second one associates the clusters to the estimated density level sets of the sample space. For a detailed review see [4].

Modal clustering is appealing for several reasons. The outlined notion of cluster is close to the intuition of groups as dense regions; consistently, clusters are not constrained to have some particular pre-determined shape and resorting to nonparametric tools allow to mantain this flexibility. Also, since clusters are the domains of attraction of the density modes, the number of clusters is an intrinsic property of the data generator mechanism and its determination is itself an integral part of the estimation procedure. Furthermore, the existence of a formalized notion of cluster, based on the features of the density, allows to define an ideal population clustering goal, and frames the clustering problem into a standard inferential context.

Despite enjoying these relevant strenghts, when resorting to the nonparametric formulation, some criticalities have to be faced, mostly related to the estimation of the density underlying the data. Firstly obtaining nonparametric density estimates is usually computationally burdensome. This issue gets worse when working in high-dimensional spaces where nonparametric estimators suffer of the "curse of dimensionality". A relevant issue is that, regardless of the specific choice of the nonparametric density estimator, the selection of a smoothing parameter is required. Choosing this parameter turns out to be crucial since an inaccurate choice could lead to a misleading resulting estimate: too large values may lead to cover interesting structures, while too small values may lead to the appearance of spurious modes. If a kernel density estimator, the most common choice in the considered framework, is employed, the selection of the smoothing parameter is based on some reference

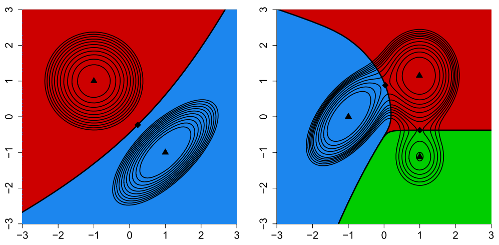


Fig. 1 Partitions induced by the modes of the density function in two examples of mixtures of bivariate normal densities.

rule or on criteria attempting to estimate properly the underlying density function. Even if these criteria have proved to produce appropriate clustering results in different situations, we believe that the clustering problem, being of a different nature with respect to the estimation of the density, would require a different rationale. In this work, a possible way to choose the optimal amount of smoothing, hinging on the specific clustering aim, is discussed. After formally defining a convenient loss function to measure the distance between data and population clustering, in the following we obtain its asymptotic expansion which, through a minimization, allows a focused selection of the smoothing parameter. Implications of this selection are finally discussed.

2 Kernel density estimation

According to the nonparametric formulation of density-based clustering the observed data $\mathscr{X} = \{x_i\}_{i=1,\dots,n}, x_i \in \mathbb{R}$, are supposed to be sampled from a random variable $\mathbb{X}$ with unknown density f. Note that, initially, we restrict our attention to the univariate case to allow a more rigorous treatment of the problem, with the intention to generalize the results to higher dimensional situations. To obtain a partition of the data adopting a nonparametric clustering perspective, regardless of its operational formulation (level set-based or mode seeking-based), an estimate $\hat{f}$ of the true density f is needed. In the rest of the paper we focus on the kernel density estimator

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) , \tag{1}$$

where K is the kernel, usually a symmetric density function, and $h > 0$ is the bandwidth, controlling the smoothness of the resulting estimate. A large value for h will tend to oversmooth the density, possibly covering some revelant features, while a small value will lead to an undersmoothed estimate where spurious modes (i.e. clusters) could arise.

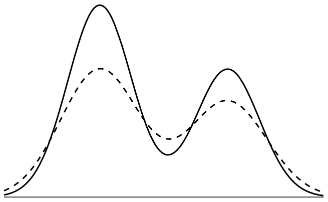


Fig. 2 Two density functions that are not close but induce exactly the same clustering.

The usual approach to select the bandwidth consists in minimizing some specific optimality criterion: the most used one is the *Mean Integrated Squared Error* (MISE)

$$MISE(h) = \mathbb{E}\int_{\mathbb{R}} \{\hat{f}_h(x) - f(x)\}^2 dx \tag{2}$$

which employs the $L2$-distance to assess the performance of the estimator. However, this expression does not have a tractable closed form and its asymptotic approximation -AMISE- is usually minimized, instead. Since both the MISE and the AMISE depend on the unknown f, different approaches to estimate them have been pursued, such as the ones based on *least square cross validation*, *biased cross validation* or *plug-in bandwidth selectors*. For a more comprehensive review and comparison see, e.g., [8].

3 The proposed selector

The selectors introduced in Section 2 are designed to choose the bandwidth so that it induces an appropriate estimate of the density. Nonetheless density estimation and clustering are two different problems with different requirements. It has been shown (e.g. [1]) that, even if two density functions are not really close, they can produce exactly the same partition of the data; see Figure 2 for an example. Furthermore, modal clustering strongly depends on some specific characteristics of the density function (the gradient for the mode seeking-based formulation and the high-density regions for the level set-based one) while, minimizing criteria such as the AMISE, an appropriate estimate is required in a global sense.

Our contribution hence finds its motivation in the lack of bandwidth selectors specifically conceived for nonparametric cluster analysis. In a similar fashion, even without specifically referring to clustering, [6] develop a plug-in type bandwidth selector that is appropriate for estimation of the highest density regions thus focusing on the modal regions, particularly relevant in the clustering formulation outlined above.

When density estimate is employed to subsequently partition the data, a more specifically tailored and appropriate performance measure than the $L2$-distance should be considered to select the amount of smoothing. Recalling [1], a natural way to quantify the performance of a data-based clustering is to consider the distance in measure between sets, where the measure has to be intended as the probability distribution with density f.

Formally, let $\mathscr{C} = \{C_1, \dots, C_r\}$ and $\mathscr{D} = \{D_1, \dots, D_r\}$ be two clusterings with the same number of groups r, their distance in measure can be measured by

$$d_1(\mathscr{C}, \mathscr{D}) = \min_{\nu \in \mathscr{P}} \sum_{i=1}^{r} \mathbb{P}(C_i \Delta D_{\nu(i)}) \, , \tag{3}$$

where $\mathscr{P}$ is the set of the permutation of $\{1, \dots, r\}$, and $C\Delta D = (C \cap D^c) \cup (C^c \cap D)$. In the following we will actually consider

$$d_P(\mathscr{C}, \mathscr{D}) = \frac{1}{2} \min_{\nu \in \mathscr{P}} \left\{ \sum_{i=1}^{r} \mathbb{P}(C_i \Delta D_{\nu(i)}) + \sum_{i=r+1}^{s} \mathbb{P}(D_{\nu(i)}) \right\} \, , \tag{4}$$

accounting for the intrinsic redundancy in (3) and for the possibility of having clustering with different number of groups. This distance can be seen as the minimal probability mass that needs to be moved to transform one clustering into the other.

Consider $\mathscr{C}_0$ as the ideal population clustering induced by the true density f and $\hat{\mathscr{C}}_n$ a data-based partition obtained from the sample $\mathscr{X}$. The idea is to quantify the quality of $\hat{\mathscr{C}}_n$ by measuring its distance in measure from $\mathscr{C}_0$. For large n, since the estimated number converges to the true number of clusters, it can be shown [1, Theorem 4.1] that (4) could be written as

$$d_P(\hat{\mathscr{C}}_n, \mathscr{C}_0) = \sum_{j=1}^{r-1} |F(\hat{m}_j) - F(m_j)| \, , \tag{5}$$

where F is the distribution function associated with f while $m_1, \dots, m_{r-1}$ and $\hat{m}_1, \dots, \hat{m}_{r-1}$ denote respectively the local minima (i.e. cluster boundaries in the univariate setting) of f and $\hat{f}$. Through two Taylor expansions, under some regularitiy conditions [1, Theorem 4.1], we obtain

$$|F(\hat{m}_j) - F(m_j)| \simeq \frac{f(m_j)}{f^{(2)}(m_j)} |\hat{f}^{(1)}(m_j)| \, , \tag{6}$$

where $f^{(j)}$ is the $j-th$ derivative of f. To obtain an asymptotic expression for (6) we have to study further the limit behavior of $\hat{f}^{(1)}(m_j)$. Considering that, if $h \to 0$ and $nh^{2r+1} \to \infty$, it is known that

$$(nh^{2r+1})^{1/2} \{\hat{f}^{(r)}(x) - K_h * f^{(r)}(x)\} \sim \mathscr{N}(0, R(K^{(r)}) f(x)) \, , \tag{7}$$

where $R(K^{(r)}) = \int_{\mathbb{R}} (K^{(r)}(x))^2 dx$ and $(h * g)(x) = \int h(x-y) g(y) dy$. For a detailed treatment of the behaviour of kernel estimators at the critical points of a density, see [5]. Studying appropriately the bias term in (7), considering $r = 1$ and focusing on the local minima (i.e. $x = m_j$) we end up obtaining the limit distribution for the quantity of interest

$$n^{2/7}\hat{f}^{(1)}(m_j) \sim \mathscr{N}\left(\frac{\beta^2 f^{(3)}(m_j)\mu_2(K)}{2}, \frac{R(K^{(1)})f(m_j)}{\beta^3}\right),$$

where $\mu_2(K) = \int_{\mathbb{R}} x^2 K(x)dx$ and $\beta = n^{1/7}h$.
Thus, considering the property of a *folded normal distribution* and after some algebra, the asymptotic *expected distance in measure* (EDM) between a data clustering $\hat{\mathscr{C}}_n$ and the ideal population clustering $\mathscr{C}_0$ is given by

$$\begin{aligned}\mathbb{E}(d_P(\hat{\mathscr{C}}_n, \mathscr{C}_0)) &= \sum_{j=1}^{r-1} \frac{f(m_j)}{f^{(2)}(m_j)} \mathbb{E}(|\hat{f}^{(1)}(m_j)|) \\ &\simeq \sum_{j=1}^{r-1} \frac{f(m_j)}{f^{(2)}(m_j)} n^{-2/7}\{2\sigma^2\phi_\sigma(\mu) + \mu[1-2\Phi_\sigma(-\mu)]\}, \qquad (8)\end{aligned}$$

where ϕ_σ and Φ_σ denote respectively the density and the distribution function of a $\mathscr{N}(0,\sigma^2)$ random variable, $\mu = \beta^2 f^{(3)}(m_j)\mu_2(K)/2$ and $\sigma^2 = R(K^{(1)})f(m_j)/\beta^3$.
The optimal bandwidth h_{d_P} for modal clustering purposes can be then obtained as $h_{d_P} = argmin_h \mathbb{E}(d_P(\hat{\mathscr{C}}_n, \mathscr{C}_0))$ by means of numerical optimization, after obtaining a suitable estimate of the unknown quantities $f(\cdot)$ and $f^{(2)}(\cdot)$ in the guise of the MISE/AMISE minimization. Another viable solution would be to work with a more manageable upper bound of (8) in order to obtain an explicit formula for the minimizer.

Further work is required to evaluate the performance of the proposed bandwidth selector as well as its comparison with some alternatives. There is much room for proceeding, and a multivariate extension of the discussed selector is needed to provide it with a concrete usability in more realistic settings.

References

1. Chacón, J.E.: A population background for nonparametric density-based clustering. Stat Sci, 30(4): 518-532 (2015).
2. Matsumoto, Y.: An introduction to Morse Theory. Amer. Math. Soc. (2002)
3. McNicholas, P.D.: Model-based clustering. J Classif, 33(3): 331-373 (2016).
4. Menardi, G.: A review on modal clustering. Int Stat Rev, 84(3): 413-433 (2016).
5. Romano, J.P.: On weak convergence and optimality of kernel density estimates of the mode. Ann Stat, 16(2):629-647 (1988).
6. Samworth, R.J & Wand, M.P.: Asymptotics and optimal bandwidth selection for highest density region estimation. Ann Stat, 38(3): 1767-1792 (2010).
7. Stuetzle, W.: Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. J Classif, 20(1): 25-47 (2003).
8. Wand, M.P. & Jones, M.C.: Kernel smoothing. Chapman & Hall (1994)

Unsupervised clustering of Italian schools via non-parametric multilevel models

Classificazione non supervisionata delle scuole italiane per mezzo di modelli a effetti misti non parametrici

Chiara Masci, Francesca Ieva and Anna Maria Paganoni

Abstract This work proposes an EM algorithm for the estimation of non-parametric mixed-effects models (NPEM algorithm) and shows its application to the National Institute for the Educational Evaluation of Instruction and Training (INVALSI) dataset of 2013/2014, as a tool for unsupervised clustering of Italian schools. Among the main novelties, the NPEM algorithm, when applied to hierarchical data, allows the covariates to be group specific and assumes the random effects to be distributed according to a discrete distribution with an (a priori) unknown number of support points. In doing so, it induces an automatic clustering of the grouping factor at higher level of hierarchy. In the application to INVALSI data, the NPEM algorithm enables the identification of latent groups of schools that differ in their effects on student achievements.

Abstract *Questo lavoro propone un algoritmo EM per la stima di modelli a effetti misti non parametrici (algoritmo NPEM) e mostra la sua applicazione ai dati dell'Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione (INVALSI) 2013/2014, con l'obiettivo di fare classificazione non supervisionata delle scuole italiane. Tra i principali vantaggi, l'algoritmo NPEM, applicato a dati gerarchici, permette alle covariate di essere specifiche del gruppo e assume che gli effetti casuali seguano una distribuzione discreta, con un numero di masse non noto a priori. Questa assunzione induce un clustering automatico del fattore di raggruppamento al piú alto livello della gerarchia. Nell'applicazione ai dati INVALSI, l'algoritmo NPEM permette l'idetificazione di gruppi latenti di scuole, che differiscono nel loro effetto sul rendimento scolastico degli studenti.*

Chiara Masci
Politecnico di Milano, via Bonardi 9, 20133 Milan e-mail: chiara.masci@polimi.it

Francesca Ieva
Politecnico di Milano, via Bonardi 9, 20133 Milan e-mail: francesca.ieva@polimi.it

Anna Maria Paganoni
Politecnico di Milano, via Bonardi 9, 20133 Milan e-mail: anna.paganoni@polimi.it

Key words: EM algorithm, Non-parametric mixed-effects models, student achievements.

1 Introduction

Administrative educational databases are often characterized by a hierarchical structure, in which students are nested within classes, that are in turn nested within schools. Given this, mixed-effects models are increasingly used in several educational applications. Mixed-effects models include parameters associated with the entire population (fixed effects) and subject/group specific parameters (random effects). They provide both estimates for the entire population's model and for each group's one, where the random effects represent a deviation from the common dynamics of the population. In this work, we develop random effects models, for applying them to educational data, whose random effects have a different meaning: they describe the common dynamics of different clusters of subjects/groups. Indeed, the mixed-effects models that we develop provide estimates for each cluster specific model and they may be considered as an unsupervised clustering tools for hierarchical data. The difference with respect to classical parametric mixed-effects models is that the random effects, instead of being Normal distributed, follow a discrete distribution that we call P^* [16]. Most of the mixed-effects models used in the educational field are parametric linear multilevel models [6], but parametric assumptions sometimes result to be too restrictive to describe very heterogeneous populations. Moreover, when the number of measurements for group is small, predictions for random effects are strongly influenced by the parametric assumptions. For these reasons, we opt for a nonparametric (NP) framework, which allows P^* to live in an infinite dimensional space and that also provides, in a natural way, a classification tool. Hierarchical models have been already applied to educational data in the Italian literature: [1], [2], [11] and [17] apply multilevel linear models in order to disentangle the portion of variability in students' scores given to different levels such as the family, the class or the school. Differently, our algorithm aims at identifying clusters of schools that perform in similar ways and, in a second step, at characterizing these clusters in terms of similarities within/between groups [13]. To the best of our knowledge, this is one of the first times that this kind of algorithm has been applied in the educational context [7]. Our method is strictly related to the branch of literature about growth mixture models (GMM) [15], latent class analysis (LCA) [14] and finite mixture models [18], which also aim at the identification of latent subpopulations, but with the main difference that all these models need to fix a priori the number of latent subpopulations. The choice of the number of latent classes (mass points) is not trivial when the sample is very big or the knowledge about possible different trends across the individuals (groups) is limited. For this reason, our approach brings a significant value-added with respect to the existing literature. In particular, our algorithm is inspired by both the one proposed in [3] and [4] and the one proposed in [5], but with some substantial changes. Contrarily to the algorithm described in [3] and [4], we do not need to fix the number of groups a priori but the algorithm identifies it by itself, standing on given tolerance values. While referring to the algorithm in [5], we adjust it in order to consider the linear case, to allow the covariates to be group-specific and to compute the optimization of the Maximization step in closed-form. We apply this algorithm to INVALSI data of year 2013/2014, in which we consider students nested within schools. Each group is identified by a school and the aim is to cluster schools standing on their different effects on their student performance trends. In this way, it is possible to identify clusters of schools that perform in different ways, trying to find out which are the determinants of different school effects.

2 The Dataset

The INVALSI database [8] contains information about more than 6,500 Italian students attending the third year of junior secondary school in the year 2013/2014, nested within about 500 schools. At pupil's level, we have reading and mathematics INVALSI test scores at grade 8 (RS and MS) and also, reading and mathematics INVALSI test scores at grade 6, two years before, of the same students. It is well known from the literature that education is a cumulative process, where achievement in the period t exerts an effect on results of the period t + 1. These variables take values between 0 and 100. Moreover, the following information is available: gender, immigrant status, if the student is early/late-enrolled, information about the family's background and socioeconomic status of the student (ESCS). At school's level, we have variables about three different areas: (i) the school-body composition (school-average characteristics of students, such as the proportion of immigrants, early and late-enrolled students, etc); (ii) school principal's characteristics; (iii) managerial practices of the school. Two dummies are also included to distinguish (i) private schools from public ones, and (ii) "Istituti Comprensivi", which are schools that include both primary and lower-secondary schools in the same building/structure. This latest variable is relevant to understand if the "continuity" of the same educational environment affects (positively or negatively) students results. Some variables about size (number of students per class, average size of classes, number of students of the school) are also included to take size effects into account. Lastly, regarding geographical location, we include two dummies for schools located in Central and Southern Italy and the district in which the school is located; some previous literature, indeed, pointed at demonstrating that students attending the schools located in Northern Italy tend to have higher achievement scores than their counterparts in other regions, all else equal. As we have the anonymous student ID, we have also the encrypted school IDs that allow us to identify and distinguish schools.

3 Methodology

We consider the case of a non-parametric two-level model with one covariate with fixed effect, one covariate with random slope and one random intercept. The model takes the following form:

$$\begin{aligned} &\mathbf{y}_i = \beta \mathbf{x}_i + c_{0l} + c_{1l}\mathbf{z}_i + \varepsilon_i \quad i = 1,\dots,N, l = 1,\dots,M \\ &\varepsilon_i \overset{ind}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{1}_n) \end{aligned} \tag{1}$$

where, in our application, N is the total number of schools; $\mathbf{y}_i$ is the n_i-dimensional vector of student achievements at grade 8 in school i; $\mathbf{x}_i$ is the n_i-dimensional vector of ESCS of students in school i; $\mathbf{z}_i$ is the n_i-dimensional vector of the same students achievements at grade 6 (two years before) in school i. We use these three variables at student level and we make this choice of random and fixed effects because we are interested in modeling the association between student achievements at grade 6 and 8, across different schools, adjusting the model for the effect of the ESCS, that, standing on the Italian literature, [2], [11], [12], results to be one of the most influential variable, with an homogeneous effect in the whole country. $\mathbf{c} \in R^2$ is the vector containing the coefficients of random effects. $\mathbf{c}$ follows a discrete distribution P^* with M support points, where M is not known a priori. P^* can then be interpreted as the mixing distribution that generates the density of the stochastic model in (1). The ML estimator $\hat{P}^*$ of P^* can be obtained following the theory of mixture likelihoods in [9] and [10], where the author proves the existence, discreteness and uniqueness of the non-parametric maximum likelihood estimator of a mixing distribution, in the case of exponential family densities.The ML estimator of the random effects distribution can be expressed as a set of points $(c_1,\dots,c_M)$, where $M \leq N$ and $c_l \in R^2$ for $l = 1,\dots,M$, and a set of weights $(w_1,\dots,w_m)$, where $\sum_{l=1}^{M} w_l = 1$ and $w_l \geq 0$ for each $l = 1,\dots,M$. Given this, we develop an algorithm for the joint estimation of σ^2, β, $(c_1,\dots,c_M)$ and $(w_1,\dots,w_M)$, that is performed through the maximization of the likelihood, mixture by the discrete distribution of the random effects,

$$L(\beta,\sigma^2,c_l,w_l|y) = \sum_{l=1}^{M} \frac{w_l}{(2\pi\sigma^2)^{\sum_{i=1}^{N} n_i}} exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} (y_{ij} - \beta x_{ij} - c_{0l} - c_{1l} z_{ij})^2 \Big\} \tag{2}$$

with respect to σ^2, β and (c_l, w_l), for $l = 1,\dots,M$. Each school i, for $i = 1,\dots,N$ is therefore assigned to a cluster l, for $l = 1,\dots,M$. The EM algorithm is an iterative algorithm that alternates two steps: the expectation step (E step) in which we compute the conditional expectation of the likelihood function with respect to the random effects, given the observations and the parameters computed in the previous iteration; and the maximization step (M step) in which we maximize the conditional expectation of the likelihood function. Moreover, given N starting support points, during the iterations of the EM algorithm, we reduce the support of the discrete distribution standing on both two criteria: the former is that we fix a threshold D and if two points are closer than D they collapse to a unique point; the latter is that we remove points with very low weight ($w_l \leq \tilde{w}$, being $\tilde{w}$ a given threshold on weights) and that are not associated to any school. When two points $\mathbf{c}_l$ and $\mathbf{c}_k$ collapse to a unique point, because their Euclidean distance is smaller than D, we obtain a new mass point $\mathbf{c}_{l,k} = \frac{\mathbf{c}_l + \mathbf{c}_k}{2}$ with weight $w_{l,k} = w_l + w_k$. The thresholds D and $\tilde{w}$ are two complexity parameters that affect the estimation of the nonparametric distribution: the higher is D, the lower is the number of clusters. The choice of the values for D and $\tilde{w}$ depends on how much we want to be sensitive to the differences among clusters (D) and which is the minimum number of groups (schools) that we allow within each clusters ($\tilde{w}$). Anyway, different results obtained using different set of tuning parameters can be compared in terms of AIC or BIC in order to choose the best one. Notice that the number of support points M is computed by the algorithm as well and we do not have to fix it a priori. Since we do not have to specify a priori the number of support points, the NP mixed-effects model could be interpreted as an unsupervised clustering tool for longitudinal data.

4 Results

The algorithm cluster the Italian schools within 5 clusters, whose estimated parameters are shown in Table 1.

	$\hat{\beta}$	$\hat{c}_0$	$\hat{c}_1$	$\hat{w}$
Cluster 1	1.417	46.028	0.454	12.2%
Cluster 2	1.417	22.579	0.707	39.6%
Cluster 3	1.417	30.293	0.648	37.5%
Cluster 4	1.417	31.207	0.393	8.8%
Cluster 5	1.417	25.359	0.027	1.9%

Table 1 Coefficients of Eq. (1) estimated by the NPEM algorithm. Each row corresponds to a cluster l. The intercept and the coefficient of z differ across groups (c_0 and c_1 respectively), while the coefficient of x (β) is fixed. $\hat{w}$ represents the weight assigned to each cluster.

Each cluster is characterized by an intercept, a slope of the grade 6 test score variable and by the fixed coefficient of the ESCS. We identify two main clusters (Cluster 2 and Cluster 3 in Table 1), that contain about the 77% of the total population of schools, while the remaining 23% is distributed across the other three clusters. From an interpretative point of view, with respect to Cluster 2 and Cluster 3 that form the reference cluster, while Cluster 5 contains the “worse” set of Italian schools. Indeed, it is characterized by both low intercept and slope and this means that there is a kind of equality in student achievements, but with on average very low scores at grade 8, even if the results at grade 6 were on average higher. In a second step, we apply a multinomial logit model at school level, by treating the five clusters as the categorical outcome variable and all the school level characteristics as covariates, with the aim of characterizing the identified clusters by means of school level variables. It emerges that the dummy for private/public school, the percentage of disadvantaged students and the geographical area are associated to heterogeneity across groups.

References

1. Agasisti, T., Vittadini, G.: Regional economic disparities as determinants of student's achievement in Italy. Research in Applied Economics, 4(2), 33 (2012).
2. Agasisti, T., Ieva, F., Paganoni, A.M.: Heterogeneity, school-effects and the North/South achievement gap in Italian secondary education: evidence from a three-level mixed model. Statistical Methods & Applications, 26(1), 157-180 (2017).
3. Aitkin M.: A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and computing, 6(3), 251-262 (1996).
4. Aitkin M.: A general maximum likelihood analysis of variance components in generalized linear models. Biometrics, 55(1), 117-128 (1999).
5. Azzimonti, L., Ieva, F., Paganoni, A.M.: Nonlinear nonparametric mixed-effects models for unsupervised classification. Computational Statistics, 28(4), 1549-1570 (2013).
6. Fox, J.: Linear mixed models. Appendix to An R and S-PLUS Companion to Applied Regression (2002).
7. Gnaldi, M., Bacci, S., Bartolucci, F.: A multilevel finite mixture item response model to cluster examinees and schools. Advances in Data Analysis and Classification, 10(1), 53-70 (2016).
8. INVALSI website
http://www.invalsi.it/
9. Lindsay, B.: The geometry of mixture likelihoods: a general theory. The Annals of Statistics, 11(1), 86-94 (1983).
10. Lindsay, B.: The geometry of mixture likelihoods, part II: the exponential family. The Annals of Statistics, 11(3), 783-792 (1983).
11. Masci, C., Ieva, F., Agasisti, T., Paganoni, A.M.: Does class matter more than school? Evidence from a multilevel statistical analysis on Italian junior secondary school students. Socio-Economic Planning Sciences, 54,47-57 (2016).
12. Masci, C., Ieva, F., Agasisti, T., Paganoni, A.M.: Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. Journal of Applied Statistics, 44(7), 1296-1317 (2017).
13. Masci, C., Ieva, Paganoni, A.M.: Non-parametric mixed-effects models for unsupervised classification of Italian schools. MOX-report 63/2017.
14. McCulloch, C. E., Lin, H., Slate, E. H., Turnbull, B. W.: Discovering subpopulation structure with latent class mixed models. Statistics in medicine, 21(3), 417-429 (2002).
15. Muthén, B., Shedden, K.: Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics, 55(2), 463-469 (1999).
16. Skrondal, A., Rabe-Hesketh, S.: Multilevel and related models for longitudinal data. In Handbook of multilevel analysis (pp. 275-299). Springer, New York, NY (2008).
17. Sulis, I., Porcu, M.: Assessing divergences in mathematics and reading achievement in italian primary schools: A pro posal of adjusted indicators of school effectiveness. Social Indicators Research, 122(2), 607-634 (2015).
18. Tutz, G., Oelker, M. R.: Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. International Statistical Review, 85(2), 204-227 (2017).

An INDCLUS-type model for occasion-specific complementary partitions

Un modello INDCLUS per partizioni complementari specifiche per occasione

Laura Bocci and Donatella Vicari

Abstract This paper presents an INDCLUS-type model for partitioning the units in three-way proximity data taking into account the systematic differences among the occasions. Specifically, the proximity structure of each occasion is assumed to underlie two complementary partitions: the first, common to all occasions, defines a partitioning of a subset of units and the second, occasion-specific, defines a partitioning of the remaining units. The model is fitted in a least-squares framework and an efficient ALS algorithm is given.

Abstract *Si presenta un modello di tipo INDCLUS per partizionare le unità nel caso di dati di prossimità a tre vie tenendo conto delle differenze sistematiche esistenti tra le similarità a coppie rilevate in diverse occasioni. In particolare, si assume che la struttura di prossimità di ciascuna occasione si componga di due partizioni complementari: un sottogruppo di unità definisce dei gruppi comuni a tutte le occasioni, mentre le rimanenti unità sono allocate a dei gruppi specifici per ogni occasione. Il modello, formulato seguendo l'approccio dei minimi quadrati, è stimato utilizzando un algoritmo ALS.*

Key words: three-way proximity data, INDCLUS, clustering

1 Introduction

[1] Laura Bocci, Department of Communication and Social Sciences, Sapienza University of Rome; email: laura.bocci@uniroma1.it

Donatella Vicari, Department of Statistical Sciences, Sapienza University of Rome; email: donatella.vicari@uniroma1.it

In many research domains, there is an increasing interest in how the perception or evaluation of several units (objects or stimuli) differs across several sources or occasions such as subjects, experimental conditions, situations, scenarios or times.

Such kind of information can be arranged in three-way data represented as a cube having multiple data matrices stacked as slices along the third dimension which represents the data sources. Our concern here is three-way two-mode data, so that the frontal slices consist of square symmetric matrices $\mathbf{S}_h$ $(h = 1, \dots, H)$ of pairwise proximities of a set of *N* units coming from H occasions.

Clustering three-way data is a complex task since each proximity data matrix $\mathbf{S}_h$ generally subsumes a (more or less) different clustering of units due to the occasion heterogeneity. Therefore, a three-way clustering should represent a consensus classification of the ones obtained from each occasion, but such a consensus is actually representative only when the classifications from each frontal slice of the three-way data do not contain systematic differences.

However, since this hypothesis is often not appropriate in real situations, several methodologies for clustering three-way proximity data have been proposed both in a least-squares (Carroll and Arabie (1983), Gordon and Vichi (1998), Vichi (1999), Vicari and Vichi (2009), Giordani and Kiers (2012), Bocci and Vicari (2017)) and in a maximum likelihood approach (Basford and McLachlan (1985), Hunt and Basford (1999), Bocci, Vicari and Vichi (2006)).

Carroll and Arabie (1983) introduced the INDCLUS (INdividual Differences CLUStering) model to extract (overlapping) clusters of *N* units from a set of proximity matrices provided in H occasions which are supposed to differently weigh each group of units. Therefore, all occasions in the INDCLUS model are assumed to employ the same clustering, but with different patterns of weights indicating the effect of each cluster on the proximities.

A fuzzy variant of the INDCLUS model, called FINDCLUS, has been proposed by Giordani and Kiers (2012), where a fuzzy membership matrix, still common to all occasions, is assigned to the classification of units.

Bocci and Vicari (2017) proposed a Generalized INDCLUS model, called GINDCLUS, for clustering both units and occasions incorporating (possible) information from external variables not directly involved in the collection process of the similarities.

In the context of clustering and multidimensional scaling for three-way data, the CLUSCALE (simultaneous CLUstering and SCAL[E]ing) model (Chaturvedi and Carroll, (2006)) combines additively INDCLUS and INDSCAL (Carrol and Chang, (1970)) searching for a common both (discrete) clustering representation and continuous spatial configuration of the units.

In order to extract more information from the three-way proximities accounting for and taking advantage of the heterogeneity of the occasions, we present an INDCLUS-type model where the pairwise proximities subsume two complementary partitions of the units: one is common to all occasions while the other is occasion-specific. Therefore, we assume that a subset of the *N* units defines a partition shared by all occasions, while the remaining units are differently clustered in each occasion. The model is formalized in a least-squares framework and an appropriate ALS algorithm is given.

2 The model

In order to present our INDCLUS-type model, let's formally recall the INDCLUS model (Carroll and Arabie, (1983))

$$\mathbf{S}_h = \widetilde{\mathbf{P}}\widetilde{\mathbf{W}}_h\widetilde{\mathbf{P}}' + \tilde{c}_h\mathbf{1}_N\mathbf{1}_N' + \mathbf{E}_h, \qquad h = 1, \dots, H, \tag{1}$$

where $\widetilde{\mathbf{P}} = [\tilde{p}_{ij}]$ ($\tilde{p}_{ij} = \{0,1\}$ for $i = 1, \dots, N$ and $j = 1, \dots, J$) is a $N \times J$ binary matrix defining the clustering of N units, $\widetilde{\mathbf{W}}_h$ is a non-negative diagonal weight matrix of order J for occasion h, $\tilde{c}_h$ is a real-valued additive constant for occasion h, $\mathbf{1}_N$ denotes the column vector with N ones and $\mathbf{E}_h$ is the square matrix of errors which represents the part of $\mathbf{S}_h$ not accounted for by the model.

On the other hand, when the occasions present systematic differences, it is reasonable to think that a subset of the N units shares the same partition and weights in all occasions but for each occasion there exists a different partitioning and weighing of the remaining units. This assumption specifies a new model

$$\begin{aligned}\mathbf{S}_h = {} & (\mathbf{P} + \mathbf{M}_h)\mathbf{W}(\mathbf{P} + \mathbf{M}_h)' \\ & + \mathbf{M}_h\mathbf{R}_h(\mathbf{P} + \mathbf{M}_h)' + \mathbf{P}\mathbf{R}_h\mathbf{M}_h' + c_h\mathbf{1}_N\mathbf{1}_N' + \mathbf{E}_h, \qquad h = 1, \dots, H,\end{aligned} \tag{2}$$

where $\mathbf{P} = [p_{ij}]$ ($p_{ij} = \{0,1\}$ for $i = 1, \dots, N$ and $j = 1, \dots, J$) is an $N \times J$ binary matrix defining the common partition of N_C ($\leq N$) units in J groups, $\mathbf{M}_h = [m_{ijh}]$ ($m_{ijh} = \{0,1\}$ for $i = 1, \dots, N$ and $j = 1, \dots, J$) is an $N \times J$ binary matrix defining the partition of $N_I = N - N_C$ units in J groups for occasion h ($h = 1, \dots, H$), $\mathbf{W} = [w_{ij}]$ is a non-negative diagonal weight matrix of order J, $\mathbf{R}_h = [r_{ijh}]$ is a non-negative diagonal matrix of order J containing the *differential* weights for occasion h ($h = 1, \dots, H$).

Therefore, on one hand, we suppose that there exists a subset N_C of the N units whose pairwise proximities for each occasion subsume the same partition **P** and the same weights **W** for its clusters; on the other hand, the partition of the remaining N_I units is supposed to be different for each occasion, so that H occasion-specific complementary partition $\mathbf{M}_h$ ($h = 1, \dots, H$) can be recognised.

Within this framework, the i-th unit ($i = 1, \dots, N$) at occasion h ($h = 1, \dots, H$) is assigned to one of the J groups in either **P** or $\mathbf{M}_h$ ($h = 1, \dots, H$): if $p_{ij} = 1$ for some $j = 1, \dots, J$ then $m_{ijh} = 0$ for all $j = 1, \dots, J$, while if $p_{ij} = 0$ for all $j = 1, \dots, J$ then $m_{ij'h} = 1$ for some $j' = 1, \dots, J$.

Therefore, conditionally on occasion h, the binary membership matrices **P** and $\mathbf{M}_h$ ($h = 1, \dots, H$) specify pairs of *incomplete and complementary* partitions which together define the occasion-specific clustering structure of the N units in J groups. Note that $\mathbf{P} + \mathbf{M}_h$ ($h = 1, \dots, H$) is a membership matrix defining a *complete* occasion-specific partition where each group is the union of the corresponding groups from **P** and $\mathbf{M}_h$ ($h = 1, \dots, H$).

The J groups identified in **P** can be considered as *core* clusters in the sense that they represent the part common to all occasions, while the *complementary* partitions identified in $\mathbf{M}_h$ $(h = 1, \dots, H)$ capture the heterogeneity of the occasions.

Moreover, the H clustering structures $\mathbf{P} + \mathbf{M}_h$ $(h = 1, \dots, H)$ are weighted by two patterns of weights: **W** which is common to all occasions and $\mathbf{R}_h$ which is occasion-specific and expresses how the *complementary* partition $\mathbf{M}_h$ is *differently* weighted with respect to the *core* partition **P** in occasion *h*.

3 The algorithm

In model (2), the incomplete membership matrices **P** and $\mathbf{M}_h$ $(h = 1, \dots, H)$, the weight matrices **W** and $\mathbf{R}_h$ $(h = 1, \dots, H)$ and constants c_h $(h = 1, \dots, H)$ can be estimated by solving the following least-squares fitting problem:

$$\min F(\mathbf{P}, \mathbf{W}, \mathbf{M}_h, \mathbf{R}_h, c_h) = \sum_{h=1}^{H} \| \mathbf{S}_h - (\mathbf{P} + \mathbf{M}_h)\mathbf{W}(\mathbf{P} + \mathbf{M}_h)' - \mathbf{M}_h\mathbf{R}_h(\mathbf{P} + \mathbf{M}_h)' - \mathbf{P}\mathbf{R}_h\mathbf{M}_h' + c_h \mathbf{1}_N \mathbf{1}_N' \| \quad (3)$$

subject to

$$p_{ij} = \{0,1\} \quad (i = 1, \dots, N; j = 1, \dots, J) \text{ and } \textstyle\sum_{j=1}^{J} p_{ij} \leq 1 \quad (i = 1, \dots, N), \quad \text{(3a)}$$

$$m_{ijh} = \{0,1\} \ (i = 1, \dots, N; j = 1, \dots, J; h = 1, \dots, H) \text{ and}$$
$$\textstyle\sum_{j=1}^{J} m_{ijh} \leq 1 \quad (i = 1, \dots, N; h = 1, \dots, H), \quad \text{(3b)}$$

$$\text{if } \textstyle\sum_{j=1}^{J} p_{ij} = 1 \text{ then } \sum_{j=1}^{J} m_{ijh} = 0 \quad (i = 1, \dots, N), \quad \text{(3c)}$$

$$\text{if } \textstyle\sum_{j=1}^{J} p_{ij} = 0 \text{ then } \sum_{j=1}^{J} m_{ijh} = 1 \quad (i = 1, \dots, N), \quad \text{(3d)}$$

$$w_{jj} \geq 0 \quad (j = 1, \dots, J), \quad \text{(3e)}$$

$$r_{jjh} \geq 0 \quad (j = 1, \dots, J; h = 1, \dots, H). \quad \text{(3f)}$$

For the sake of clarity, the set of constraints (3a)-(3d) specify that in the partitions defined by **P** and $\mathbf{M}_h$ $(h = 1, \dots, H)$ each units is assigned to only one clusters in either **P** or $\mathbf{M}_h$ $(h = 1, \dots, H)$, while constraints (3e)-(3f) specify the non-negativity of the weights **W** and $\mathbf{R}_h$ $(h = 1, \dots, H)$, respectively.

Problem (3) can be solved using an appropriate coordinate descent algorithm also known as Alternating Least-Squares (ALS) algorithm, which alternates the estimation of a set of parameters when all the other are fixed.

The algorithm proposed here estimates in turn:

a) the *core* membership matrix **P** by sequentially solving assignment problems for the different rows of **P**, where each unit can possibly remain not assigned;
b) the occasion-specific membership matrix $\mathbf{M}_h$ $(h = 1, \dots, H)$ by sequentially solving h $(h = 1, \dots, H)$ assignment problems for the rows of $\mathbf{M}_h$ corresponding to the N_I units not assigned in **P**;
c) the common weight matrix **W** by solving a regression problem using nonnegative least squares (Lawson and Hanson, 1974);

d) the occasion-specific weight matrix $\mathbf{R}_h$ $(h = 1, \dots, H)$ by solving h $(h = 1, \dots, H)$ regression problems using nonnegative least squares (Lawson and Hanson, 1974);
e) the additive constant c_h $(h = 1, \dots, H)$ by successive residualizations of the three-way data matrix.

The five main steps are alternated and iterated until convergence and the best solution over different random starting classification matrices is retained to prevent from local minima.

Results from applications to real and artificial data will be presented to show the performance of the algorithm and the capability of the model to capture the heterogeneity of the occasions.

References

1. Basford, K.E., McLachlan, G.J.: The Mixture Method of Clustering Applied to Three-way Data. J. Classif. **2**, 109--125 (1985).
2. Bocci, L., Vicari, D., Vichi, M.: A mixture model for the classification of three-way proximity data. Comput. Stat. Data An. **50**, 1625--1654 (2006).
3. Bocci, L., Vicari, D: GINDCLUS: Generalized INDCLUS with external information. Psychometrika **82**, 355--381 (2017).
4. Carroll, J.D., Arabie, P.: INDCLUS: an Individual Differences Generalization of ADCLUS model and the MAPCLUS algorithm. Psychometrika **48**, 157--169 (1983).
5. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-generalization of the Eckart-Young decomposition. Psychometrika **35**, 283--319 (1970).
6. Chaturvedi, A., Carroll, J.D.: CLUSCALE (CLUstering and multidimensional SCAL[E]ing): a three-way hybrid model incorporating clustering and multidimensional scaling structure. J. Classif. **23**, 269--299 (2006).
7. Giordani, P., Kiers, H.A.L.: FINDCLUS: Fuzzy INdividual Differences CLUStering. J. Classif. **29**, 170--198 (2012).
8. Gordon, A.D., Vichi, M.: Partitions of Partitions. J. Classif. **15**, 265--285 (1998).
9. Hunt, L.A., Basford, K.E.: Fitting a Mixture Model to Three-mode Three-way Data with Categorical and Continuous Variables. J. Classif. **16**, 283--296 (1999).
10. Lawson, C.L., Hanson, R.J.: Solving least squares problems. Prentice-Hall, Englewood Cliffs, NJ (1974).
11. Vichi, M.: One Mode Classification of a Three-way Data Set. J. Classif. **16**, 27--44 (1999).
12. Vicari, D., Vichi, M.: Structural Classification Analysis of Three-Way Dissimilarity Data. J. Classif. **26**, 121--154 (2009).

Robust Reduced k-Means and Factorial k-Means by trimming

Un approccio robusto a Reduced k-Means e Factorial k-Means

Luca Greco and Antonio Lucadamo and Pietro Amenta

Abstract In this paper we propose a robust version of Reduced and Factorial k-means, based on a trimming strategy. Reduced and Factorial k-means are data reduction techniques for simultaneous dimension reduction and clustering. The occurrence of data inadequacies can invalidate standard analyses. An appealing approach to develop robust counterparts of Reduced and Factorial k-means is given by impartial trimming. The idea is to discard a fraction of observations that are selected as the most distant from the centroids.
Abstract *In questo lavoro viene proposta una versione robusta di Reduced e Factorial k-means, basata su una procedura di trimming. Reduced e Factorial k-means sono tecniche che simultaneamente realizzano una riduzione della dimensionalitá e della numerosità, mediante analisi in componenti principali e k-means, rispettivamente. La presenza di contaminazione nei dati puó invalidare le analisi standard. Un approccio utile per sviluppare una procedura robusta alla presenza di valori anomali é rappresentato dal trimming, che si basa sull'idea di eliminare le osservazioni piú distanti dai centroidi stimati.*

Key words: Clustering, Factorial k-means, Reduced k-means, Trimmed k-means

Luca Greco
DEMM University of Sannio, Piazza Arechi II Benevento, e-mail: lucgreco@unisannio.it

Antonio Lucadamo
DEMM University of Sannio, Piazza Arechi II Benevento, e-mail: antonio.lucadamo@unisannio.it

Pietro Amenta
DEMM University of Sannio, Piazza Arechi II Benevento, e-mail: amenta@unisannio.it

1 Introduction

Reduced (De Soete and Carroll, 1994) and Factorial k-means (Vichi and Kiers, 2001) (RKM and FKM, respectively, hereafter) are data reduction techniques aimed at performing principal components and k-means clustering simultaneously. The main idea is that cluster centroids are located in a low dimensional subspace determined by the most relevant features.

Let $\mathbf{X}$ be the $n \times p$ zero centered data matrix, where n denotes the number of objects and p the number of variables, k be the number of clusters and $q < p$ the number of components, with $k \geq q+1$. We denote by $\mathbf{U}$ the $n \times k$ membership matrix whose i^{th} row has a one corresponding to the cluster assignment of the i^{th} object and zero otherwise, whereas $\mathbf{A}$ is the $p \times q$ loadings matrix and $\mathbf{Y} = \mathbf{XA}$ is the $n \times q$ scores matrix. RKM looks for centroids in a low dimensional subspace that minimize the distance of the data points from such centroids. The optimization problem connected with RKM can be expressed as

$$\min_{A,\bar{Y}} ||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}^{\mathrm{T}}||^2 = \min_{A,\bar{Y}} \sum_{i=1}^{n} \min_{c=1,\dots,k} \sum_{j=1}^{p} \left(x_{ij} - \sum_{j'=1}^{q} \bar{y}_{cj'} a_{j'j} \right)^2 , \qquad (1)$$

where $\bar{\mathbf{Y}}$ is the $k \times q$ matrix collecting centroids $\bar{\mathbf{y}}_{\mathbf{c}} = (\bar{y}_{c1}, \dots, \bar{y}_{cq})$. In a complementary fashion, FKM finds low dimensional centroids such that the scores, rather than the original data, are closest to them, that is

$$\min_{A,\bar{Y}} ||\mathbf{XA} - \mathbf{U}\bar{\mathbf{Y}}||^2 = \min_{A,\bar{Y}} \sum_{i=1}^{n} \min_{c=1,\dots,k} \sum_{j=1}^{q} (y_{ij} - \bar{y}_{cj})^2 \ . \qquad (2)$$

Both RKM and FKM are built on conventional k-means that can be badly affected by the occurrence of contamination in the data (the reader is pointed to Farcomeni and Greco (2016) for a recent account on robustness issues in data reduction). In this paper, we aim at developing a robust counterpart of RKM and FKM stemming from trimmed k-means (Cuesta-Albertos et al, 1997). Let us assume that a fraction α of data points is prone to contamination and therefore discarded. The remaining part of clean objects is then used to solve the k-means optimization problem. Trimmed data are not assigned to any cluster and do not contribute to centroids estimation. The key feature is that trimming and estimation are performed simultaneously: this approach is usually referred as impartial trimming (Gordaliza, 1991). Here, in a similar fashion, it is suggested to introduce impartial trimming into problems (1) and (2). The interesting features of the proposed methodologies rely on the ability of the method, on the one hand, to detect the anomalous data and rightly assign the remaining ones to clusters and, on the other hand, to estimate clusters' centroids in the presence of outliers. These objectives are shown in the following illustrative examples. Since the complementary nature of the RKM and FKM models, we only consider the latter model both for simulated and real data.

Figure 1 displays the result of applying trimmed FKM (tFKM) to three simulated datasets of size 400 × 8. All the panels display the scores lying in a two dimensional subspace, where the data are assumed to be clustered. A sample of 320 genuine scores has been drawn from a mixture of three bivariate normal with standardized components and null correlation, that is we are dealing with spherical clusters. Then 80 outliers were added, that are simulated from a different random mechanism. Genuine data have been generated according to the scheme described in Timmerman et al (2010). Three different scenarios have been considered. In the top left panel, the anomalous data have been randomly generated from a bivariate normal centered at the mean of the centroids corresponding to the three original groups with a dispersion large enough to produce both inner and outer contaminations. In the top middle panel, outliers are clustered to form a well separated group from the genuine observations. In the top right panel, outliers are clustered along some linear structures. In all scenarios, we observe the capability of the proposed method both in detecting outliers and adapting to the true underlying clustering structure. On the contrary, in the first and third data configuration, the standard procedure allocates outliers in the three clusters leading to biased centroids' estimates, inflated within-group variances and lack of separation among them, whereas in the second scenario, two well separated clusters are wrongly merged together.

2 Trimmed RKM and FKM

The optimization problems connected with trimmed RKM and FKM (tRKM and tFKM, hereafter) can be expressed as follows, respectively:

$$\min_{z \in Z} \min_{A,\bar{Y}} \sum_{i=1}^{n} z_i \min_{c=1,\ldots,k} \sum_{j=1}^{p} \left(x_{ij} - \sum_{j'=1}^{q} \bar{y}_{cj'} a_{j'j} \right)^2 \tag{3}$$

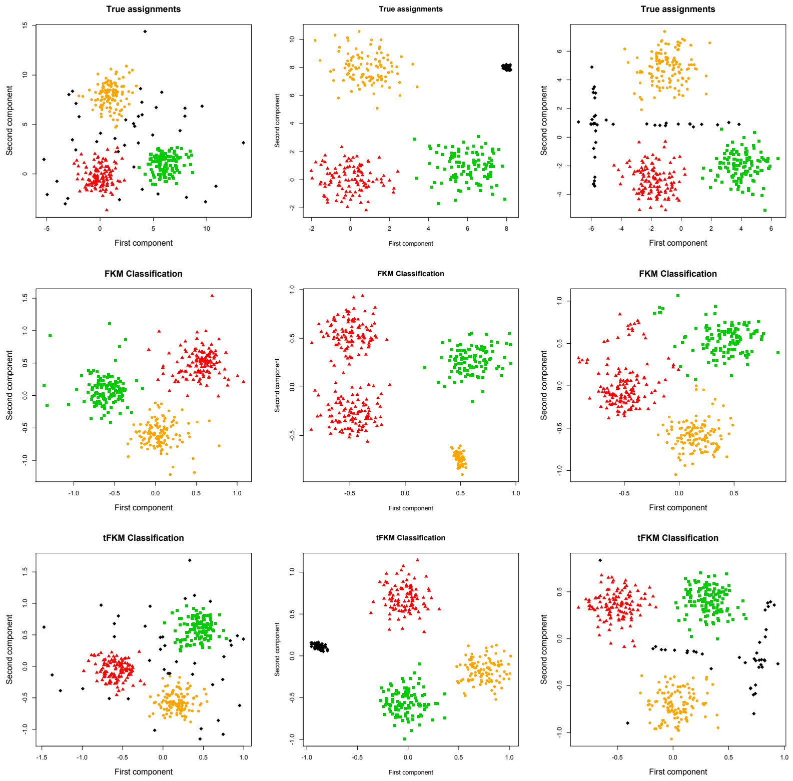
and

$$\min_{z \in Z} \min_{A,\bar{Y}} \sum_{i=1}^{n} z_i \min_{c=1,\ldots,k} \sum_{j=1}^{q} \left(y_{ij} - \bar{y}_{cj} \right)^2, \tag{4}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_n)^{\mathrm{T}}$ is a binary vector and $Z = \{z : \sum_{i=1}^{n} z_i = \lfloor (n(1-\alpha) \rfloor \}$. The objective functions (3) and (4) are such that data points for which $z_i = 0$ do not contribute to their minimization. It is worth noting that classical RKM and FKM are included in the above definitions as limiting cases when $\alpha = 0$.

Minimization of the objective functions (3) and (4) requires an iterative algorithm characterized by the introduction of concentration steps. The concentration step (Rousseeuw and Driessen, 1999; Gallegos and Ritter, 2005; Farcomeni, 2009) is meant to detect the $n\alpha$ largest distances with respect to the current closest centroid and trim the corresponding observations. Then, the loss functions in (1) or (2) are minimized based on the observations not flagged for trimming. The final $\mathbf{U}$ obtained at convergence simultaneously identifies the optimal cluster for each

Fig. 1 Three simulated data sets with true assignments (top row), classical classification from FKM (middle row), robust classification from tFKM (bottom row). The symbol ♦ is used to denote true outliers in the top row and trimmed observations in the other rows.



observation that is not trimmed and trimmed observations, which correspond to a constant row of zeros. Both algorithms have been implemented into the statistical environment `R` by combining the main features of the functions `cluspca` from package `clustrd` and `tkmeans` from package `tclust`.

3 Selecting the number of clusters, components and the trimming level

The selection of the number of clusters k can be pursued by paralleling the approach described in García-Escudero et al (2011): the strategy could be to display the objective function at convergence against the trimming level α for different choices of

k. Then, the number of clusters should be set equal to the minimum k for which there is no substantial improvement in the loss function when adding one group more.

In order to choose the number of components q, it is suggested to explore the quality of the fitted model by varying q from 1 to $(k-1)$. For instance, this task could be exploited by looking at the rate of explained robust variance or the adjusted Rand index. Then, the number of components can be augmented until there is no more significant increase in the selected criterion.

The selection of α could be based on the inspection of the G-statistic or of the generalized G-statistic introduced in Farcomeni (2009). Parameters' estimates or the objective function itself are monitored by varying α. Then, we select a trimming level above which the differences in parameters' estimates or in the objective function become negligible.

4 Macroeconomic data

This is a 20×6 data set, already analyzed in Vichi and Kiers (2001), concerning the macroeconomic performance of national economies in September 1999. Six main economic indicators, that measure the percentage change from the previous year, have been considered: gross domestic product (GDP), leading indicator (LI), unemployment rate (UR), interest rate (IR), trade balance (TB), net national savings (NNS). A classification of the countries into $k = 3$ groups is considered, that is expected to reflect the striking features of economic development and to take into account the differences in growth among them. The G-statistic leads to select $\alpha = 0.15$ (i.e. 3 trimmed observations). Figure 2 gives the classification resulting from FKM and tFKM. There are remarkable differences between the classical and the robust analysis, mainly due to the three outlying countries that have been detected. The cluster profiles and the raw score measurements for the outlying countries are given in Table 1. It can be seen that the three clusters are well separated, even if Cluster 2 and Cluster 3 are separated only w.r.t. the second component. The first component is mainly determined by NNS (positive sign), UR (positive sign) and TB (negative sign), whereas the second is dominated by GDP. The first cluster is composed by 4 countries that are characterized by the largest values on the first component, that is countries showing large values of NNS or UR and small values of TB. The second cluster is composed by 5 countries. It is characterized by the largest values on the second component, that is countries with large GDP. The third cluster is composed by 8 countries, that are those exhibiting the lowest growth in GDP. The outlying countries Sweden and Japan are easily explained. Sweden is well spotted on the left side and it actually shows the lowest NNS; Japan, on the contrary, is well detected along the first component since it exhibits the minimum growth in GDP. The explanation for Denmark is more complicated: it could be included in Cluster 3 but it shows an unexpected low growth in UR and NNS.

Fig. 2 Macroeconomic data: classification from FKM (left) and tFKM (right). Trimmed observations in the right panel are denoted by ♦.

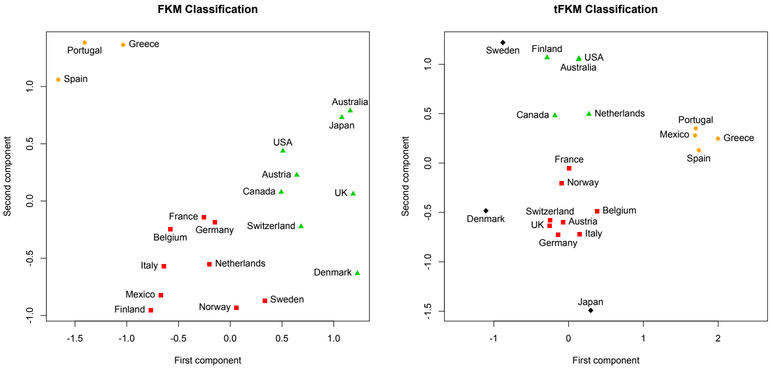


Table 1 Macroeconomic data: cluster profiles and raw scores for the outlying countries.

	Comp. 1	Comp. 2		Comp.1	Comp. 2
Cluster 1	1.782	0.252	Denmark	-1.106	-0.483
Cluster 2	0.014	0.830	Japan	0.296	-1.491
Cluster 3	-0.034	-0.500	Sweden	-0.879	1.220

References

Cuesta-Albertos J, Gordaliza A, Matrán C (1997) Trimmed *k*-means: An attempt to robustify quantizers. The Annals of Statistics 25(2):553–576

De Soete G, Carroll JD (1994) K-means clustering in a low-dimensional euclidean space. In: New approaches in classification and data analysis, pp 212–219

Farcomeni A (2009) Robust double clustering: a method based on alternating concentration steps. Journal of Classification 26(1):77–101

Farcomeni A, Greco L (2016) Robust methods for data reduction. CRC press

Gallegos M, Ritter G (2005) A robust method for cluster analysis. Annals of Statistics pp 347–380

García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2011) Exploring the number of groups in robust model-based clustering. Statistics and Computing 21(4):585–599

Gordaliza A (1991) Best approximations to random variables based on trimming procedures. Journal of Approximation Theory 64(2):162–180

Rousseeuw P, Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223

Timmerman M, Ceulemans E, Kiers HA, Vichi M (2010) Factorial and reduced k-means reconsidered. Computational Statistics & Data Analysis 54(7):1858–1871

Vichi M, Kiers H (2001) Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis 37(1):49–64

Dirichlet processes, posterior similarity and graph clustering

Processo di Dirichlet, similarità a posteriori e classificazione su grafi

Stefano Tonellato

Abstract This paper proposes a clustering method based on the sequential estimation of the random partition induced by the Dirichlet process. Our approach relies on the Sequential Importance Resampling (SIR) algorithm and on the estimation of the posterior probabilities that each pair of observations are generated by the same mixture component. Such estimates do not require the identification of mixture components, and therefore are not affected by label switching. Then, a similarity matrix can be easily built, allowing for the construction of a weighted undirected graph. A random walk can be defined on such a graph, whose dynamics is closely linked to the posterior similarity. A community detection algorithm, the map equation, can then be implemented in order to achieve a clustering by minimising an information theoretic criterion.

Abstract *Si propone un metodo di classificazione basato sulla stima sequenziale della partizione indotta dal processo di Dirichlet. Il metodo si basa su un algoritmo stocastico sequenziale (SIR) e sulla probabilità a posteriori che ciascuna coppia di osservazioni sia generata da una componente della mistura. Il metodo non richiede l'identificazione delle singole componenti e non risente degli inconvenienti del label switching. Una matrice di similarità può quindi essere stimata a posteriori. Questo consente la costruzione di un grafo pesato e la definizione, su di esso, di una passeggiata aleatoria. Un algoritmo utilizzato nella classificazione di reti, chiamato map equation, che ricerca la partizione che minimizza un criterio di informazione, può quindi essere facilmente implementato.*

Key words: Dirichlet process, sampling importance resampling, community detection, map equation

Stefano Tonellato
Ca' Foscari University of Venice, e-mail: stone@unive.it

1 Dirichlet process mixtures and clustering

A very important class of models in Bayesian nonparametrics is based on the Dirichlet process and is known as Dirichlet process mixture [1]. In this model, the observable random variables, X_i, $i = 1, \ldots, n$, are assumed to be exchangeable and generated by the following hierarchical model:

$$\begin{aligned} X_i|\theta_i &\overset{ind}{\sim} p(\cdot|\theta_i),\ \theta_i \in \Theta \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0), \end{aligned}$$

where $DP(\alpha, G_0)$ denotes a Dirichlet process (DP) with base measure G_0 and precision parameter $\alpha > 0$. Since the DP generates almost surely discrete random measures on the parameter space Θ, ties among the parameter values have positive probability, leading to a batch of clusters of the parameter vector $\theta = [\theta_1, \ldots, \theta_n]^T$. Exploiting the Pólya urn representation of the DP, the model can be rewritten as

$$X_i|s_i, \theta^*_{s_i} \overset{iid}{\sim} p(\cdot|\theta^*_{s_i}),\ \theta^*_{s_i} \in \Theta \tag{1}$$

$$\theta^*_{s_i} \overset{iid}{\sim} G_0 \tag{2}$$

$$p(s_i = j|\mathbf{s}_{<i}) = \begin{cases} \frac{\alpha}{\alpha+i-1} & j = k \\ \frac{n_j}{\alpha+i-1} & j \in \{k-1\}, \end{cases} \tag{3}$$

$$s_i \perp \theta^*_j \qquad \forall i, j, \tag{4}$$

where $\{k\} = \{1, \ldots, k\}$, $\mathbf{s}_{<i} = \{s_j,\ j \in \{i-1\}\}$ (in the rest of the paper, the subscript $< i$ will refer to those quantities that involve all the observations $X_{i'}$ such that $i' < i$), $s_j \in \{k\}$ for $j \in \{k-1\}$, and n_j is the number of θ_i's equal to θ^*_j. In this model representation, the parameter θ can be expressed as $(\mathbf{s}, \theta^*)$, with $\mathbf{s} = \{s_i : s_i \in \{k\},\ i \in \{n\}\}$, $\theta^* = [\theta^*_1, \ldots, \theta^*_k]^T$ with $\theta^*_j \overset{iid}{\sim} G_0$, and $\theta_i = \theta^*_{s_i}$. Consequently, the marginal distribution of X_i is a mixture with k components, where k is an unknown random integer.

In the case of finite mixtures with k components, with k fixed and known, under a frequentist perspective it would be quite straightforward to cluster the data by maximising the probability of the allocation of each datum to one of the k components, conditionally on the observed sample [6]. Under a Bayesian perspective, the same results can be achieved, provided that either some identifiability constraints on the parameters are introduced, or a suitable risk function is minimised [12]. Unfortunately, under the assumptions we made, such computations are not feasible even numerically, due to the well known label switching problem [3] that persists when the number of mixture components is not known, nor finite, as in the case of Dirichlet process mixtures. Nevertheless, equations (1)–(4) are very helpful in estimating posterior pairwise similarities and building hierarchical clustering algorithms as in [7, 8]. In section 2, a sequential estimation algorithm analogous to the one in [5] is developed. In section 3, individuals are represented as nodes of a weighted undirected graph on which a random walk is built, with transition probabilities proportional to the posterior similarities. Nodes can then be classified by minimising the entropy through the map algorithm introduced in [10, 11]. The approach proposed in sections 2 and 3 has a double benefit. On one hand, the sequential estimation algorithm guarantees a fast estimation of pairwise similarities. On the other hand, the construction of the random walk on the graph mentioned above, allows us to choose the optimal partition by a minimum description length algorithm, so avoiding the subjective choice of a cut of the dendrogram usually associated to hierarchical clustering algorithms. Furthermore, as a byproduct, the entropy of any partition of the data can be computed and it is closely linked to the fitted model. This allows for a model based comparison of any pair of partitions.

2 Sampling importance resampling

Under the assumptions we introduced above, following the arguments of [5], we can write the conditional posterior distribution of s_i given $x_1,\ldots,x_i$, as

$$p(s_i = j|\mathbf{s}_{<i},\theta^*,\mathbf{x}^{(j)}_{<i},x_i) = \begin{cases} \frac{n_j}{\alpha+i-1} p(x_i|\theta^*_j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) & j \in \{k\} \\ \frac{\alpha}{\alpha+i-1} p(x_i|\theta^*_{k+1}) & j = k+1, \end{cases}$$

where $\mathbf{x}^{(j)}_{<i} = \{x_{i'} : i' < i, s_{i'} = j\}$, $j = 1,\ldots,k$, and $\mathbf{x}^{(k+1)}_{<i} = \emptyset$, since $\forall i' < i$, $s_{i'} \in \{k\}$.

We can marginalise the conditional posterior of s_i with respect to θ^*, obtaining

$$p(s_i = j|\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i},x_i) = \begin{cases} \frac{n_j}{\alpha+i-1} p(x_i|s_i = j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) & j \in \{k\} \\ \frac{\alpha}{\alpha+i-1} p(x_i|s_i = k+1,\mathbf{s}_{<i},\mathbf{x}_{<i}) & j = k+1, \end{cases}$$

where

$$p(x_i|s_i = j,\mathbf{s}_{<i},\mathbf{x}_{<i}) = \int_\Theta p(x_i|\theta,s_i = j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) p(\theta|s_i = j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) d\theta \tag{5}$$

and

$$p(x_i|s_i = k+1,\mathbf{s}_{<i},\mathbf{x}_{<i}) = \int_\Theta p(x_i|\theta) dG_0(\theta). \tag{6}$$

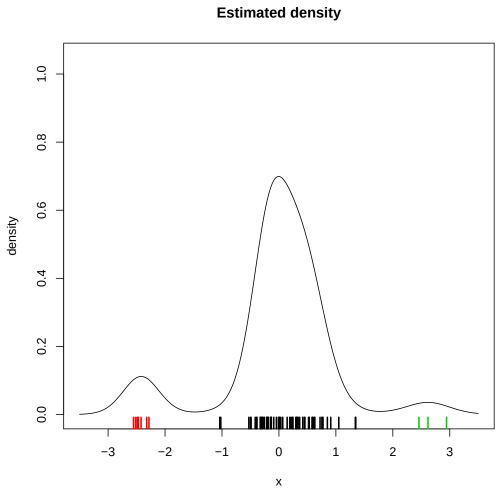
Notice that when G_0 is a conjugate prior for (1), the computation of (5) and (6) is often straightforward.

The following importance sampler has been introduced in [5].

SIR algorithm. For $i = 1,\ldots,n$, repeat steps (A) and (B)

(A) Compute

Fig. 1 Estimated posterior density function and clustering



$$g(x_i|\mathbf{s}_{<i},\mathbf{x}_{<i}) \propto \sum_{j=1}^{k+1} \frac{n_j}{\alpha+i-1} p(x_i|s_i=j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}),$$

with $n_{k+1}=\alpha$.

(B) Generate s_i from the multinomial distribution with

$$p(s_i=j|\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i},x_i) \propto \frac{n_j}{\alpha+i-1} p(x_i|s_i=j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}).$$

Taking R independent replicas of this algorithm we obtain $s_i^{(r)}$, $i=1,\ldots,n$, $r=1,\ldots,R$, and $\theta_j^* \sim p(\theta|\mathbf{x}^{(j)})$, with $\mathbf{x}^{(j)}=\{x_i : i \in \{n\}, s_i=j\}$, and compute the importance weights

$$w_r \propto \prod_{i=1}^{n} g(x_i|\mathbf{s}_{<i},\mathbf{x}_{<i})$$

such that $\sum_{r=1}^{R} w_r = 1$. Should the variance of the importance weights be too small, the efficiency of the sampler could be improved by resampling as follows [2]. Compute $N_{\text{eff}} = (\sum_{r=1}^{R} w_r^2)^{(-1)}$. If $N_{\text{eff}<\frac{R}{2}}$, draw R particles from the current particle set with probabilities equal to their weights, replace the old particle with the new ones and assign them constant weights $w_r = \frac{1}{R}$.

3 Pairwise similarities and community detection

Intuitively, we can state that two individuals, i and j, are similar if x_i and x_j are generated by the same mixture component, i.e. if $s_i = s_j$. Label switching prevents

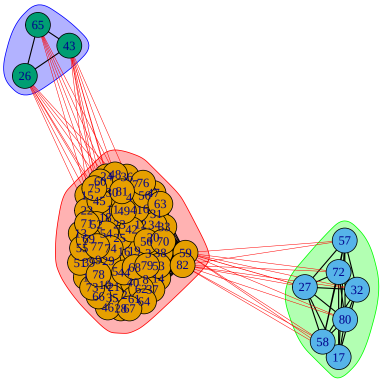


Fig. 2 The graph induced by the posterior similarity and the clusters detected by the map equation algorithm

us from identifying mixture components, but not from assessing similarities among individuals. In fact, the algorithm introduced in the previous section may help us in estimating pairwise similarities between individuals. The posterior probability that x_i and x_j are generated by the same component, i.e. the posterior probability of the event $\left\{s_i = s_j\right\}$, can be estimated as

$$\hat{p}_{ij} = \sum_{r=1}^{R} w_r I\left(s_i^{(r)}, s_j^{(r)}\right),$$

where $I(x,y) = 1$ if $x = y$ and $I(x,y) = 0$ otherwise. We can then define a similarity matrix S with ij-th element $s_{ij} = \hat{p}_{ij}$.

The matrix S can be used to build the weighted undirected graph $G = (V,E)$, where each node in the set V represents an individual in the sample, i.e. $V = \{n\}$, and the set E contains all the edges in G. Furthermore, the weight of the generic edge (i,j) is given by $w_{ij} = s_{ij}$ if $i \neq j$, and $w_{ij} = 0$ otherwise. We can then define a random walk $\mathscr{X}$ on G, with state space V. Let $d_i = \sum_{j=1}^{n} w_{ij}$, $i = 1,\ldots,n$ and $D = \mathrm{diag}(d_1,\ldots,d_n)$. We define the transition matrix of $\mathscr{X}$ as $P = D^{-1}W$. If G is connected, $\mathscr{X}$ has π as invariant distribution, with $\pi_i = \frac{d_i}{\sum_{i,j} w_{ij}}$ [4]. The random walk we have just defined represents an artificial stochastic flow such that the probability of moving from i to j is proportional to w_{ij}, i.e. to the similarity between i and j. Such a dynamics induces some high density subsets of V, i.e. subsets where the random walker spends a long time before moving to other clusters, separated by low weight edges. In such a context, community detection algorithms attempt to identify an optimal partition of V. We shall refer, in particular, to the so called map equation [10, 11] that attempts to find a partition of V such that the length of the code describing the behaviour of $\mathscr{X}$ is minimised. Let M be a partition of V. The map equation computes the entropy $L(M)$, which is strictly related to P and π. The optimal partition minimises $L(M)$. As stated in [10], "the map equation calculates the minimum description length of a random walk on the network for a two-level code that separates the important structures from the insignificant details based on the partition M".

As an example, figures 1 and 2 show the results of an application to the well known galaxy data set [9].

References

1. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian non parametric problems. Annals of Statistics. **2**, 1152–1174 (1974).
2. Cappé, O., Moulines, E., and T., Rydén: Inference in Hidden Markov Models. Springer, New York (2005)
3. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer: Berlin (2006)
4. Lovász, L.: Random walks on graphs: a survey. In: Combinatorics, Paul Erdös is eighty, pp. 353–397. János Bolyai Math. Soc., Budapest (1993)
5. MacEachern, S.N., Clyde, M., and Liu, J.S: Sequential importance sampling for nonparametric Bayes models: The next generation. The Canadian Journal of Statistics, **27**, 251–267 (1999)
6. McLachlan, G., and Peel, D.: Finite Mixture Models. Wiley, New York (2000)
7. Medvedovic, M., and Sivaganesan, S.: Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics, **18**, 1194–1206 (2002)
8. Medvedovic, M., Yeung, K.Y., and Bumgarber, R.E.: Bayesian mixture model based clustering of replicated microarray data. Bioinformatics, **20**, 1222–1232 (2004)
9. Roeder, K.: Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. JASA **85** 617–624
10. Rosvall, M. and Bergstrom C. T.: Maps of random walks on complex networks reveal community structure. PNAS. **105**, 1118–1123 (2008) doi: www.pnas.org/cgi/doi/10.1073/pnas.0706851105
11. Rosvall, M., Axelsson, D., and Bergstrom C.T.: The map equation. Eur. Phys. J. Special Topics 178, 1323 (2009) doi: 10.1140/epjst/e2010-01179-1
12. Stephens, M.: Dealing with label switching in mixture models. J. R. Statistic. Soc. B. **62**, 795–809 (2000)

Bootstrap ClustGeo with spatial constraints

Bootstrap ClustGeo con vincoli spaziali

Veronica Distefano[1,2], Valentina Mameli[1], Fabio Della Marra[1,2]

Abstract The aim of this paper is to introduce a new statistical procedure for clustering spatial data when an high number of covariates is considered. In particular, this procedure is obtained by coupling the agglomerative hierarchical clustering method that ha been recently proposed for spatial data, referred as *ClustGeo* (*CG*) method , with the bootstrap technique. The proposed procedure, which we call *Bootstrap ClustGeo* (*BCG*), is developed and tested on a real dataset. The results that we achieve show that *BCG* outperforms *CG* in terms of accuracy of some cluster evaluation measures.

Abstract Il presente lavoro propone una nuova procedura di clustering per dati spaziali, che denoteremo *Bootstrap ClustGeo* (*BCG*). In particolare, questa nuova procedura coniuga il metodo di clustering agglomerativo gerarchico, recentemente proposto per dati spaziali sotto il nome di *ClustGeo* (*CG*), con il metodo bootstrap. I risultati ottenuti dimostrano una migliore performance dell'approccio *BCG* secondo un insieme di misure di valutazione di clustering.

Key words: Agglomerative hierarchical clustering, Bootstrap technique, *ClustGeo* method, Geographical data, Hamming distance, Spatial data.

1 Introduction

Addressing a study on sustainable development of geographical areas is becoming crucial to derive geopolitical policy. As noted by UN-GGIM (2012), "all of

[1] European Centre for Living Technology, Ca' Foscari University of Venice, Italy.
[2]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy.
e-mail: veronica.distefano@unive.it, e-mail: valentina.mameli@unive.it, e-mail: fabio.dellamarra@unive.it

the issues impacting sustainable development can be analyzed, mapped, discussed and modeled within a geographic context. Whether collecting and analyzing satellite images or developing geopolitical policy, geography can provide the integrative framework necessary for global collaboration and consensus decision-making" [11]. In this context, the geo-spatial clustering procedures represent an important area of research in data analysis, and a growing interest in clustering spatial data is emerging in several application fields. Indeed, the geo-spatial data and accurate clustering approaches in selecting constraints and parameters can provide better and more meaningful results.
In this study, we address the problem of clustering n spatial locations into K disjoint clusters when an high number of covariates is considered. Most approaches in the literature have been developed to derive clusters containing only contiguous locations. These procedures are based on the assumption that, within each cluster, there exists a connecting path for any couple of locations [9, 10, 5, 2]. This assumption involves then a strict-spacial constraint, i.e. locations characterized by social-economic variables with very similar values, but not close to each other in space, will be likely to be grouped into different clusters. Very recently, a non-strict constrained procedure has been developed, in which the condition of spatial closeness is relaxed [4]. In [6], the authors propose a hierarchical clustering method with non-strict spatial constraints, which is referred as *ClustGeo*. The method is based on two dissimilarity matrices: a matrix with dissimilarities derived from the "covariate-space"and a matrix with the dissimilarities derived from the "non-strict constraint space". In our work, we extend the aforementioned approach proposed in [6] by developing a procedure based on the generation of multiple bootstrap clustering partitions combined by using the Hamming distance. The novel procedure is called *Bootstrap ClustGeo*.

The paper is organized as follows. In Section, 2 we review the *ClustGeo* method and we then introduce the novel procedure *Bootstrap ClustGeo*. In Section 3, we evaluate this procedure on a real and known database which includes 303 French municipalities characterized by a set of 10 socio-economic covariates.

2 Methods

In classical cluster analysis, similar observations can be grouped into clusters. There exist different types of clustering algorithms which have been developed for different structure of data to be analyzed [8, 3]. In this paper, we select the agglomerative hierarchical clustering approach for analyzing geo-spatial data. The structure of an agglomerative hierarchical clustering approach can be summerised as follows. At the initialization, each cluster contains a single observation. Then, at each step of the agglomerative process, the two clusters with the smallest distance are merged into a new one. This procedure is iterated until a single cluster containing all the observations is obtained. The results of the agglomerative algorithm are usually represented with a tree or a dendrogram.

Formally, let $\{x_i = (x_{i1}, \ldots, x_{ip})\}_{i=1,\ldots,n}$ be the set of n observations (municipalities in the dataset), each of which is described by p covariates. Let $\{w_i\}_{i=1,\ldots,n}$ be the weights associated to the i-th observation selected by the researcher. $\{w_i\}_{i=1,\ldots,n}$ can be proportional to the inverse of the total poluation of the municipality x_i. Consider $D_0 = (d_{0ij})_{i,j=1,\ldots,n}$ a $n \times n$ dissimilarity matrix associated with the n observations, where $d_{0ij} = d_0(x_i,x_j)$ is the dissimilarity measure between observations i and j, which could be Euclidean or non-Euclidean. In this paper we focus only on the non-Euclidean case. The matrix D_0 is usually referred as the dissimilarity of the "covariate-space". Consider also the $n \times n$ dissimilarity matrix $D_1 = \{(d_1(x_i,x_j))\}_{i,j=1,\ldots,n} = (d_{1ij})_{i,j=1,\ldots,n}$ containing spatial constraints between the observations, which is refereed as the dissimilarity of the "non-strict geo-spatial constraint space". In this research we consider the geographical distances as spatial constraints.

Let $\alpha \in [0,1]$ be a parameter which controls the importance of the spatial constraints in the clustering procedure. Suppose that the dataset, $X = (x_1,\ldots,x_n)^T$, is partitioned into K clusters C_k^α with $k = 1,\ldots,K$, which form the partition $\mathscr{P}_K^\alpha = (C_1^\alpha,\ldots,C_K^\alpha)$. In the agglomerative clustering process for spatial data, the distance between clusters could be determined by using the *ClustGeo* (*CG*) method as proposed in [6], which is based on the minimization of the pseudo within-cluster inertia of the partition $\mathscr{P}_K^\alpha$. The pseudo within-cluster inertia is defined as

$$W(\mathscr{P}_K^\alpha) = \sum_{k=1}^{K} I_\alpha(C_k^\alpha),$$

where

$$I_\alpha(C_k^\alpha) = (1-\alpha)\sum_{i\in C_k^\alpha}\sum_{j\in C_k^\alpha}\frac{w_i w_j}{2\mu_k^\alpha}d_{0ij}^2 + \alpha\sum_{i\in C_k^\alpha}\sum_{j\in C_k^\alpha}\frac{w_i w_j}{2\mu_k^\alpha}d_{1ij}^2,$$

with $\mu_k^\alpha = \sum_{i\in C_k^\alpha} w_i$. Starting from a given partition $\mathscr{P}_K^\alpha$, the *CG* method aggregates the two clusters with the smallest within-cluster inertia in order to obtain a new partition with $K-1$ clusters.

In order to analyze the database with an high number of covariates, in this paper we develop a novel procedure based on the *CG* method. This novel procedure is derived by generating multiple bootstrap clustering partitions with the *CG* method and combining the results by using the Hamming distance. We will refer to the new methodology as the *Bootstrap ClustGeo* (*BCG*) method. The main features of this novel methodology are described in the following.

At first, we take B bootstrap sample from the n data points x_i. Each bootstrap sample sample will be denoted by X_b, for $b = 1,\ldots,B$. For each bootstrap sample X_b, we will set the number of clusters K_b by drawing it from a discrete uniform distribution, i.e. $K_b \sim DUnif$. Then, we use the *CG* method to find a clustering partition $\mathscr{P}_{K_b}^\alpha$ of size K_b for each dataset X_b. At the end of the B bootstrap replications, we construct the incidence matrix

$$\mathscr{I}^{\alpha} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1B} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nB} \end{bmatrix},$$

where r_{ib} represents the index of the cluster in the partition $\mathscr{P}^{\alpha}_{K_b}$ to which x_i belongs. At last, we derive a new dissimilarity matrix

$$D^{\alpha}_{B} = (d^{\alpha}_{B}(x_i, x_j))_{i,j=1,\dots,n},$$

in which

$$d^{\alpha}_{B}(x_i, x_j) = \frac{1}{B} \sum_{b=1}^{B} \delta(r_{ib}, r_{jb}),$$

where δ is the Hamming distance. The new dissimilarity matrix D^{α}_{B} will be used to find a new clustering partition by exploiting the *CG* method.

3 Results

In this section, we compare the performances of *CG* and *BCG* methods on a real and known database. We perform a geo-spatial clustering analysis with geographical constraints in the administrative region of the Nouvelle Aquitanie, which is located in the southwest of France. The available dataset consists of n = 303 municipalities and $p = 10$ indicators, which includes social and economic indicators, as shown in Table 1. The data source of the indicators is the INSEE (*National Institute of Statistics and Economic Studies*).

The number of clusters has been estimated by visual inspection of the clustering tree generated by the *CG* algorithm when only D_0 is taken into account. The optimal value of the parameter α has been chosen by using the criterion proposed in [6]. From Table 1, we can see that the means of the vast majority of the variables within clusters 1 and 2 do not show significant differences across the two clustering methods. We can notice more relevant differences in the remaining clusters (3, 4, and 5). For example, the mean of the variable “Ratio of the agricultural area”assumes a small value (2.90) in cluster 5 by using the *CG* method with respect to the mean of the same variable (35.16) in cluster 3 with the *BCG* method.

From the results presented in Fig. 1, we can note that the clusters obtained by using the *BCG* are spatially more compact than those obtained by the *CG* method. Indeed, the municipalities located in cluster 2 for the *CG* method are grouped into a greater number of clusters than by the *BCG* method. Neither of the two methods requires any strict-contiguity assumption, as shown in Fig 1. Hence, cluster 1 contains municipalities which are not connected in space.

To assess the quality of the two clustering partitions obtained with the *CG* and the *BCG* algorithms, we consider some known evaluation measures, such as the Con-

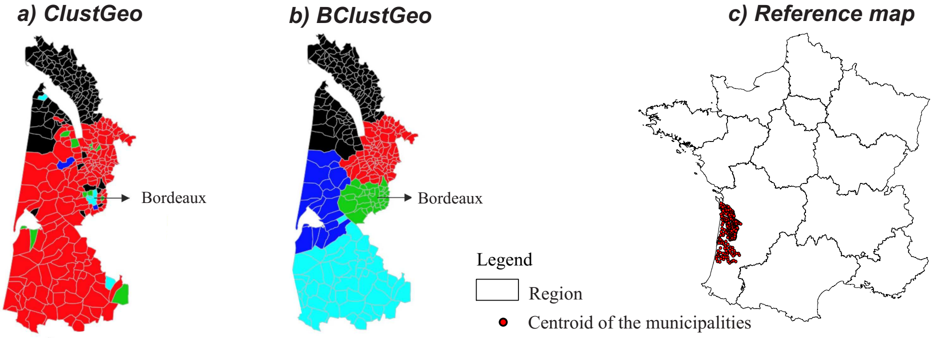
Indicators	Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
(S) Employment rate	*CG*	27.92 (11.09)	25.45 (13.41)	24.94 (12.48)	32.04 (0.42)	61.35 (20.98)
	BCG	28.34 (11.23)	24.28 (11.09)	24.94 (12.78)	31.70 (16.71)	61.35 (14.97)
(S) Level of Education	*CG*	15.43 (3.23)	14.88 (2.91)	17.12 (1.91)	16.43 (1.82)	17.28 (2.45)
	BCG	14.91 (3.08)	15.13 (2.68)	17.12 (2.69)	16.49 (3.27)	17.2 (3.15)
(S) Ratio of apartment housing	*CG*	6.37 (7.81)	5.11 (6.22)	33.32 (13.61)	10.38 (20.07)	74.39 (11.21)
	BCG	5.42 (9.24)	5.59 (11.85)	33.32 (18.39)	10.35 (13.37)	74.39 (9.63)
(S) Ratio of the agricultural area	*CG*	63.53 (25.18)	51.90 (29.33)	12.77 (8.15)	21.44 (1.30)	4.99 (2.19)
	BCG	64.45 (24.32)	45.85 (27.17)	12.77 (32.26)	19.60 (29.13)	4.99 (21.17)
(S) Average density of the population	*CG*	132.82 (182.43)	102.67 (98.32)	1480.56 (285.86)	195.83 (153.02)	4995.70 (330.07)
	BCG	109.15 (510.10)	110.97 (414.29)	1480.56 (1223.27)	201.23 (381.15)	4995.70 (662.32)
(E) Share of workplaces on business, transportations and in financial activities	*CG*	28.65 (19.83)	26.33 (19.99)	48.96 (8.20)	30.25 (10.32)	48.50 (6.31)
	BCG	24.48 (19.44)	29.09 (20.93)	48.96 (22.05)	30.74 (18.80)	48.50 (18.51)
(E) Share of workplaces on public administration, education, health and social action	*CG*	31.08 (19.06)	27.19 (16.91)	27.14 (10.53)	30.35 (8.49)	45.60 (10.39)
	BCG	29.28 (18.78)	27.14 (14.91)	27.14 (13.81)	30.47 (16.13)	45.60 (21.06)
(E) Share of employees on the total number of jobs	*CG*	25.13 (10.51)	25.11 (12.21)	26.57 (5.91)	25.85 (0.00)	29.80 (1.54)
	BCG	25.13 (10.64)	25.11 (11.82)	26.57 (9.11)	25.85 (9.25)	29.80 (13.97)
(E) Share of workers on the total number of jobs	*CG*	25.84 (13.25)	29.45 (13.98)	21.54 (5.11)	28.52 (1.41)	12.00 (4.32)
	BCG	27.00 (13.15)	30.01 (13.15)	21.54 (11.97)	28.47 (7.50)	12.00 (15.94)
(E) Share of owners in the main residences	*CG*	76.30 (8.30)	74.64 (9.01)	54.95 (6.29)	68.75 (9.19)	32.00 (11.59)
	BCG	76.68 (8.40)	73.45 (11.27)	54.95 (13.73)	68.57 (9.61)	32.00 (9.43)

Table 1: Comparison of the CG and the BCG methods in derived clusters for Economic (E) and Social (S) indicators. In each cluster we present the mean value and the standard deviation within brackets.

nectivity, the average Silhouette, and the Dunn indices [7]. The Connectivity represents the strength of connectedness of the clusters, lies in the range between 0 and infinity and should be minimized. Both the Silhouette and the Dunn indices measure the quality of the compactness and the separation of the the clusters and should be maximized. The Silhouette value lies in the interval $[-1,1]$, and the Dunn index assumes values between 0 and infinity. As we can see from Table 2, the *BCG* algorithm outperforms the *CG* algorithm for all the indices. These results show that the novel methodology proposed is capable of exploiting the spatial constraints to achieve better clustering accuracy in comparison with the *CG*. The use of the bootstrap sampling and the Hamming distance for categorical variables permit to accurately obtain information for defining the Clustering structure. The results show the BCG to be more accurate then the CG.

		Indices		
		Silhouette	Dunn	Connectivity
Methods	*CG*	0.07	0	117
	BCG	0.79	0.64	0

Table 2: Evaluation measures to validate the clustering methods.



d) Comparison of the CG and BCG methods

	Nr.municipalities	
	ClustGeo	*BClustGeo*
⬛ Cluster 1	108	112
🟥 Cluster 2	178	90
🟩 Cluster 3	11	32
🟦 Cluster 4	2	19
▢ Cluster 5	4	50

Fig. 1: Study area and maps generated by the *CG* and the *BCG* algorithms.

References

1. Amiri, S., Clarke, B.S., Clarke, J.L.: Clustering categorical data via ensembling dissimilarity matrices. J. Comput. Graph. Statist. 1–14 (2017).
2. Becue-Bertaut, M., Alvarez-Esteban, R., Sanchez-Espigares, J.A., Xplortext: Statistical Analysis of Textual Data R package. https://cran.r-project.org/package=Xplortext. R-package version 1.0 (2017).
3. Benassi, F., Bocci, C. and Petrucci, A.. Spatial data mining for clustering: an application to the Florentine Metropolitan Area using RedCap. Classification and Data Mining, pp. 157-164. Springer, Berlin, Heidelberg (2013)
4. Bourgault, G., Marcotte, D., Legendre, P.: The Multivariate (co) Variogram as a Spatial Weighting Function in Classification Methods. Mathematical Geology 24(5): 463–478 (1992).
5. Carvalho, A. X. Y., Albuquerque, P. H. M., de Almeida Junior, G. R., Guimaraes, R. D.: Spatial hierarchical clustering. Revista Brasileira de Biometria, 27(3), 411–442 (2009).
6. Chavent, M., Kuentz-Simonet, V., Labenne, A., Saracco, J.: ClustGeo: an R package for hierarchical clustering with spatial constraints. Computational Statistics, 1-24 (2018).
7. Brock, G., Pihur, V., Datta, S., Datta, S.: clValid: An R Package for Cluster Validation. Journal of Statistical Software 25: 1–22 (2008)
8. Everitt, B., Landau, S., Leese, M., Stahl, D.:Cluster analysis. 5th edn, Wiley, Chichester (2011).
9. Lance, G.N., Williams, W.T.: A General Theory of Classicatory Sorting Strategies 1. Hierarchical Systems. The Computer Journal 9: 373–380 (1967).
10. Murtagh, F.: Multidimensional clustering algorithms. Compstat Lectures, Vienna: Physika, Verlag (1985).
11. UN-GGIM. 2012. Monitoring Sustainable Development: Contribution of Geospatial Information to the Rio Processes. New York: United Nations. Accessed January 17, (2016).

Advances in Statistical Models

Regression modeling via latent predictors

Regressione basata su predittori latenti

Francesca Martella and Donatella Vicari

Abstract A proposal for multivariate regression modeling based on latent predictors (LPs) is presented. The idea of the proposed model is to predict the responses on LPs which, in turn, are built as linear combinations of disjoint groups of observed covariates. The formulation naturally allows to identify LPs that best predict the responses by jointly clustering the covariates and estimating the regression coefficients of the LPs. Clearly, in this way the LP interpretation is greatly simplified since LPs are exactly represented by a subset of covariates only. The model is formalized in a maximum likelihood framework which is intuitively appealing for comparisons with other methodologies, for allowing inference on the model parameters and for choosing the number of subsets leading to LPs. An Expectation Conditional Maximization (ECM) algorithm is proposed for parameter estimation and experiments on simulated and real data show the performance of our proposal.

Abstract *In questo lavoro si propone un nuovo modello di regressione basato su predittori latenti (PL) che, a loro volta, sono modellizzati come combinazioni lineari di gruppi disgiunti di covariate osservate. Tale formulazione permette di identificare direttamente i migliori PL che predicono le risposte attraverso: la classificazione delle covariate e la stima dei coefficienti di regressione. In questo modo l'interpretazione dei PL é notevolmente semplificata poiché i PL sono rappresentati solamente da sottoinsiemi di covariate. Il modello é formalizzato in un contesto di massima verosimiglianza e viene presentato un algoritmo Expectation Conditional Maximization (ECM) per la stima dei parametri. Esperimenti su dati simulati e reali mostrano l'utilitá e la validitá della proposta.*

Francesca Martella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Rome, e-mail: francesca.martella@uniroma1.it

Donatella Vicari
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Rome e-mail: donatella.vicari@uniroma1.it

Key words: Regression model, Clustering, Latent Predictors, Maximum Likelihood

1 Introduction

Considering ordinary repression modeling of several dependent variables (responses) on a large set of covariates is not always the best choice. In fact, in such a case, difficulties in interpretation of the (many) regression coefficients and the presence of multicollinearity among predictors may arise. Many strategies can be adopted in order to reduce such problems such as standard variable selection methods, penalized (or shrinkage) techniques and dimensionality reduction methods (DRMs). The latter attempt to build small set of linear combinations of the predictors, used as input to the regression model, and differ in how the linear combinations are built. See among others, principal component regression (PCR, [7]), factor analysis regression (FAR, [2]), canonical correlation regression (CCR, [5]), partial least squares regression (PLSR, [12]), [10] for the continuum regression (unified regression technique embracing OLS, PLSR and PCR), reduced rank regression (RRR, [1], [6]), redundancy analysis (RA, [11]). [13] proposed a general formulation for dimensionality reduction and coefficient estimation in multivariate linear regression, which includes many existing DRMs as specific cases. Finally, [3] proposed a new formulation to the multiblock setting of latent root regression applied to epidemiological data and [4] investigated a continuum approach between MR and PLS. [8] proposed a multivariate regression model based on the optimal partition of predictors (MRBOP). A drawback of DRMs is that they may generally suffer from a possible difficulty of interpretability of the resulting linear combinations which are often overcome through rotation methods. Here, we propose to build latent predictors (LPs) as linear combinations of disjoint groups of covariates that best predict the responses where such groups identify block-correlated covariates. Actually, we simultaneously perform clustering of the covariates and estimation of the regression coefficients of the LPs. This turns out to be a relevant gain in the interpretation of the regression analysis, since LPs are formed by disjoint groups of covariates and, therefore easily interpretable. Clearly, in this way the LP interpretation is greatly simplified since LPs are exactly represented by a subset of covariates only. An Expectation Conditional Maximization algorithm (ECM, [9]), for maximum likelihood estimation of the model parameters is described. The performance of the proposed model is confirmed by the application on simulated and real data sets. The results are encouraging and would deserve further discussion.

2 Model

Consider $\mathbf{x}_i$ be a J-dimensional data vector representing the covariates and $\mathbf{y}_i$ be a M-dimensional data vector of the responses observed on the i-th unit in a sample of size n. Without loss of generality, $\mathbf{x}_i$ and $\mathbf{y}_i$ are assumed to be centered to zero mean vector. Our proposal can be summarized by two models: the first is a regression model formalizing the relations between responses and latent predictors, while the second one represents a dimensional reduction model where the latent predictors synthesize the relations among the covariates. In formula, we have

$$\mathbf{y}_i = \mathbf{C}'\mathbf{f}_i + \mathbf{e}_i \tag{1}$$

and

$$\mathbf{f}_i = \mathbf{V}'\mathbf{W}\mathbf{x}_i + \xi_i \tag{2}$$

where $\mathbf{C}$ is the $(Q \times M)$ regression coefficient matrix, $\mathbf{f}_i$ is the Q-th dimensional LP vector, $\mathbf{e}_i$ is the M-th dimensional noise term vector, $\mathbf{V}$ is the $(J \times Q)$ binary membership matrix defining a partition of the covariates in Q non-empty groups $(Q \leq J)$, $\mathbf{W}$ is the $(J \times J)$ diagonal matrix which gives weights to the J covariates, ξ_i is the Q-th dimensional noise term vector $(i = 1, \ldots, n)$.
Moreover, we assume that the noise terms are independent and follow multivariate Normal distributions: $\mathbf{e}_i \sim \mathrm{MVN}(\mathbf{0}, \Sigma_{\mathbf{e}})$ with $\Sigma_{\mathbf{e}}$ diagonal matrix and $\xi_i \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}_Q)$. Then, we derive that $\mathbf{f}_i \sim \mathrm{MVN}(\mathbf{V}'\mathbf{W}\mathbf{x}_i, \mathbf{I}_Q)$, $\mathbf{y}_i \sim \mathrm{MVN}(\mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i, \mathbf{C}'\mathbf{C} + \Sigma_{\mathbf{e}})$ and, conditional on $\mathbf{f}_i$, results in $\mathbf{y}_i | \mathbf{f}_i \sim \mathrm{MVN}(\mathbf{C}'\mathbf{f}_i, \Sigma_{\mathbf{e}})$. Thus, the log-likelihood function $l(\Theta)$, being $\Theta = \{\mathbf{C}, \mathbf{V}, \mathbf{W}, \Sigma_{\mathbf{e}}\}$, is given by

$$\begin{aligned} l(\Theta) = &- \sum_{i=1}^{n} \left[(2\pi)^{M/2} |\mathbf{C}'\mathbf{C} + \Sigma_{\mathbf{e}}|^{1/2} \right] \\ &+ \sum_{i=1}^{n} \left\{ \frac{1}{2} \left(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i \right)' (\mathbf{C}'\mathbf{C} + \Sigma_{\mathbf{e}})^{-1} \left(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i \right) \right\}. \end{aligned} \tag{3}$$

Finding the maximum likelihood estimates for $\mathbf{C}, \mathbf{W}, \Sigma_{\mathbf{e}}, \mathbf{V}$ is more problematic. We propose an ECM algorithm which iteratively computes the expected value of the complete-data log-likelihood and maximizes the expected complete-data log-likelihood over one of the parameters while holding the other fixed until convergence is achieved. Similarly to the factor analysis context, we take $\mathbf{y}$ as the observed data and $\mathbf{f}$ as the missing data, by assuming, therefore, that the complete data vector consist of $\mathbf{z}_i = (\mathbf{y}_i', \mathbf{f}_i')'$ $(i = 1, \ldots, n)$. Therefore, the complete-data log-likelihood is given by:

$$\begin{aligned} \ell_C(\Theta) &= \sum_{i=1}^{n} \log\left[\phi(\mathbf{y}_i | \mathbf{f}_i, \Theta)\right] + \sum_{i=1}^{n} \log\left[\phi(\mathbf{f}_i | \Theta)\right] \\ &= \sum_{i=1}^{n} \log\left[\phi(\mathbf{y}_i | \mathbf{f}_i, \Theta)\right] + \sum_{i=1}^{n} \log\left[\phi(\mathbf{f}_i | \mathbf{W}, \mathbf{V})\right] \end{aligned} \tag{4}$$

since the distribution of $\mathbf{f}_i$ is independent of $\mathbf{C}$ and $\Sigma_{\mathbf{e}}$. The expected value of $\mathbf{f}_i$ conditional on $\mathbf{y}_i$ and the current model parameters is

$$\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta) = \mathbf{V}'\mathbf{W}\mathbf{x}_i + \beta(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i) \tag{5}$$

and

$$\begin{aligned}\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta) = {}& (\mathbf{I}_Q - \beta\mathbf{C}') + [\mathbf{V}'\mathbf{W}\mathbf{x}_i + \beta(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i)]\\ & [\mathbf{V}'\mathbf{W}\mathbf{x}_i + \beta(\mathbf{y}_i - \mathbf{C}'\mathbf{V}'\mathbf{W}\mathbf{x}_i)]'\end{aligned} \tag{6}$$

where $\beta = \mathbf{C}(\mathbf{C}'\mathbf{C} + \Sigma_{\mathbf{e}})^{-1}$. Therefore the expected complete-data log-likelihood Q is

$$\begin{aligned}Q = K - \frac{n}{2}\log|\Sigma_{\mathbf{e}}| - \frac{1}{2}\sum_{i=1}^{n}\{&\mathbf{y}_i'\Sigma_{\mathbf{e}}^{-1}\mathbf{y}_i - 2\mathbf{y}_i'\Sigma_{\mathbf{e}}^{-1}\mathbf{C}'\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)\\ &+ \mathrm{Tr}\left[\mathbf{C}\Sigma_{\mathbf{e}}^{-1}\mathbf{C}'\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta)\right] + \mathrm{Tr}\left[\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta)\right]\\ &- 2\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)'\mathbf{B}_i\mathbf{w} + \mathrm{Tr}\left[\mathbf{w}'\mathbf{B}_i'\mathbf{B}_i\mathbf{w})\right]\}.\end{aligned} \tag{7}$$

where K is a constant, $\mathbf{w}$ is the J-dimensional vector of the diagonal elements of $\mathbf{W}$ (i.e. $\hat{\mathbf{W}} = \mathrm{diag}(\hat{\mathbf{w}})$), and $\mathbf{B}_i$ is the $(Q \times J)$ matrix having the j-th column equal to $\mathbf{v}_j x_{ij}$ with $\mathbf{v}_j$ being the j-th row of $\mathbf{V}$ $(i = 1,\ldots,n)$.
Differentiating Q with respect to each parameter in Θ and setting to zero the corresponding score functions, we obtain

$$\hat{\mathbf{C}} = \left[\sum_{i=1}^{n}\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta)\right]^{-1}\left[\sum_{i=1}^{n}\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)\mathbf{y}_i'\right], \tag{8}$$

$$\hat{\Sigma}_{\mathbf{e}} = \frac{1}{n}\mathrm{diag}\left[\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' - \sum_{i=1}^{n}\mathbf{y}_i\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)'\mathbf{C}\right], \tag{9}$$

$$\hat{\mathbf{w}} = \left[\sum_{i=1}^{n}\mathbf{B}_i'\mathbf{B}_i\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{B}_i'\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)\right]. \tag{10}$$

In order to estimate the membership matrix of the covariates $\hat{\mathbf{V}}$, we proceed as follows:

- For each covariate j and group q, compute the log-likelihood values

$$l_{jq} = l(\cdot, v_{jq} = 1|\mathbf{C}, \Sigma_{\mathbf{e}}, \mathbf{W}, \{v_{hs}\}_{h=1,\ldots,J,h\neq j;s=1,\ldots,Q,s\neq q});$$

- Fix j and compute the maximum of this set $\{l_{jq}\}$ over $q = 1,\ldots,Q$; denote this term by l_j^{max};
- Allocate the j-th covariate to the q-th group ($\hat{v}_{jq} = 1$) iff $l_{jq} = l_j^{max}$ $q = 1,\ldots,Q$.

The ECM algorithm for the proposed model therefore becomes:

- E-step: Compute the expected values $\mathrm{E}(\mathbf{f}_i|\mathbf{y}_i,\Theta)$ and $\mathrm{E}(\mathbf{f}_i\mathbf{f}_i'|\mathbf{y}_i,\Theta)$ for all data ($i = 1,\dots,n$).
- CM-steps: Maximize Q over one of the parameters Θ while holding the other fixed.

The log-likelihood function, l, is computed for the current parameter values. The two steps are repeatedly alternated until convergence, which is reached when:

$$l_{(r)} - l_{(r-1)} < \varepsilon, \qquad \varepsilon > 0 \tag{11}$$

where r is the current iteration and ε is a small tolerance value.

3 Conclusions

A new mutivariate regression model based on latent predictors is presented. The latter are built as linear combinations of disjoint groups of observed covariates which best predict the responses. In this way, we jointly cluster covariates and estimate regression coefficients of the LPs. The model is particularly appropriate in a regression context where the reduction of the number of covariates is required for interpretability reasons or multicollinearity problems. In fact, in situations where the covariates are block-correlated, the assumptions on the covariances of the error terms, which are supposed diagonal, are fulfilled and lead to a gain in terms of parsimony and interpretability. We describe an EM algorithm for estimating model parameters and we will discuss the performance of the proposed approach on both simulated and real datasets. The results are encouraging and would deserve further discussion.

References

1. Anderson, T.W.: Estimating linear restrictions on regression coefficients for multivariate distributions. Ann. Math. Stat. **22**, 327–351 (1951)
2. Basilevsky, A.: Factor analysis regression. The Canadian Journal od Statistics, **9(1)**, 109–117 (1981)
3. Bougeard, S., Hanafi, M., Qannari, E.M.: Multiblock latent root regression: application to epidemiological data. Comput. Stat. **22(2)**, 209–222 (2007)
4. Bougeard, S., Hanafi, M., Qannari, E.M.: Continuum redundancy-PLS regression: a simple continuum approach. Comput. Stat. Data Anal. **52(7)**, 3686–3696 (2008)
5. Hotelling, H.: The most predictable criterion. J. Educ. Psychol. **25**, 139–142 (1935)
6. Izenman, A.J.: Reduced-rank regression for the multivariate linear model. J. Multivar. Anal. **5**, 248–262 (1975)
7. Kendall, M.G.: A Course in Multivaraite Analysis. Griffin, London (1957)
8. Martella, F., Vicari D., Vichi, M.: Partitioning predictors in multivariate regression models. Stat. Comput. **25(2)**, 261–272 (2013)
9. Meng, Xiao-Li, Rubin, D,: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika. **80(2)**, 267-278 (1993).
10. Stone, M., Brooks, R.J.: Continuum regression: cross-validated squares partial least squares and principal components regression. J. R. Stat. Soc. b **52(2)**, 237–269 (1990)

11. Van Den Wollenberg, A.L.: Redundancy analysis an alternative for canonical analysis. Psychometrika **42(2)**, 207–219 (1977)
12. Wold, H.: Estimation of principal components and relates models by iterative least squares. in Krishnaiaah, P.R. (ed.) Multivariate Analysis, 391–420. Academic Press, New York (1966)
13. Yuan, M., Ekici, A., Lu, Z., Monteiro, R.: Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc., Ser. B, Stat. Methodol. **69**, 329–346 (2007)

Analysis of dropout in engineering BSc using logistic mixed-effect models

Analisi dell'abbandono universitario nelle lauree di primo livello di ingegneria utilizzando modelli logistici a effetti misti

Luca Fontana and Anna Maria Paganoni

Abstract The main goal of this report is to apply a logistic mixed-effect model to analyse the relationship between the success probability in getting the BSc engineering degree in Politecnico di Milano and different sets of covariates, using as grouping factor the engineering programme attended. The dataset of interest contains detailed information about more than 18,000 students enrolled in BSc from 2010 to 2013. This analysis is performed within the Student Profile for Enhancing Tutoring Engineering (SPEET) ERASMUS+ project that involves Politecnico di Milano and five other european engineering universities, aimed at opening a new perspective to university tutoring systems.

Abstract *L'obiettivo di questo report è quello di applicare un modello logistico a effetti misti per analizzare la possibile dipendenza tra la probabilità di concludere con successo la Laurea di Primo Livello in Ingegneria al Politecnico di Milano e diversi insiemi di covariate, raggruppando gli studenti per corso di studi frequentato. Il dataset utilizzato contiene informazioni dettagliate riguardo più di 18,000 studenti immatricolati tra il 2010 e il 2013. Questa analisi è parte del progetto Student Profile for Enhancing Tutoring Engineering (SPEET), una collaborazione ERASMUS+ tra il Politecnico di Milano e cinque altri atenei europei di ingegneria.*

Key words: academic data; engineering programmes; university tutoring systems; generalized linear mixed-effects model; dropout prediction.

Luca Fontana
MOX, Dipartimento di Matematica, Politecnico di Milano,
P.za Leonardo da Vinci 32, 20133 Milano (Italy)
e-mail: `luca11.fontana@mail.polimi.it`

Anna Maria Paganoni
MOX, Dipartimento di Matematica, Politecnico di Milano,
P.za Leonardo da Vinci 32, 20133 Milano (Italy)
e-mail: `anna.paganoni@polimi.it`

1 Introduction

The present work is a first step of statistical analysis of academic data related to engineering students attending Bachelor of Science Degree in Politecnico di Milano. This analysis is performed within the Student Profile for Enhancing Tutoring Engineering (SPEET) ERASMUS+ project that involves Politecnico di Milano, University of Galati (Romania), Escola d'Enginyeria UAB (Spain), Instituto Politecnico de Braganca (Portugal), Opole University of Technology (Poland) and Universidad de Leon (Spain). The project novelty emerges from the potential synergy among the huge amount of academic data actually existing at the academic departments and the maturity of data science algorithms and tools to analyse and extract information from those data. SPEET project aims to process the academic data to identify different student profiles to provide them with a personal tutoring service [6]. Despite many possibilities we have chosen to start the project by analyzing the distinction between students who complete their programme and those who instead decide to abandon studies [2]. This choice is based on the fact that across all SPEET partners almost a student out of two resigns his engineering studies before obtaining the BSc degree, and this phenomenon marks Politecnico di Milano in a remarkable way. The student profiles we are referring to within the SPEET project scope are:

- *dropout*: careers permanently finished for any reason other than the achievement of the BSc degree;
- *graduate*: careers definitely closed with the achievement of academic qualification, sooner or later.

Mixed-effects models are commonly employed in the analysis of grouped or clustered data, where observations in a cluster cannot reasonably be assumed to be independent of one-another: we use a mixed model in which engineering students are nested within the programmes they are attending.

2 Dataset and Model

The dataset consists of $18,612$ careers that began from A.Y. 2010/2011 to A.Y. 2013/2014, from 19 different engineering programmes at Politecnico di Milano. Collected data include degree information, collateral information regarding the students' background (nationality, previous studies, age, ...) as well as student performance on every subject of his study plan (subject score, subject year, subject semester, ...). The variables we are including in the model are described in table 1.

We now analyze the relationship between the success probability in getting the degree and a set of explainatory variables. Our response variable is the career `status`, a two-level factor we coded as a binary variable:

- `status` $= 1$ for careers definitely closed with the achievement of academic qualification (factor level = *graduate*)

- `status` $= 0$ if the career is permanently finished for any reason other than the achievement of the BSc degree (factor level = *dropout*).

Since those careers belong to 19 different engineering programmes, the choice of a Logit Mixed-Effects Model in which students are nested within programmes is justified [1]. The influence of the programme on the linear predictor is modeled through a single random effect on the intercept. Let y_{ij} denote observation j in group i $(i = 1:19, j = 1:n_i)$. The model formulation in extended form is:

$$\begin{aligned} Y_{ij} &\sim Bernoulli(p_{ij}) \\ p_{ij} &= E[Y_{ij}|b_i] = P(y_{ij}=1|b_i) \\ logit(p_{ij}) &= \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \sum_{h=0}^{p} x_{ijh}\beta_k + b_i \\ b_i &\sim N(0,\sigma^2) \\ b_i, b_{i'} &\text{ are independent for } i \neq i' \end{aligned} \tag{1}$$

where x_{ijh} represents the values of explanatory variables for fixed effects model parameters; $\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \dots \quad \beta_p]^T$ is the $(p{+}1)$-dimensional vector of fixed effects and b_i is the scalar random effect on the intercept of the linear predictor for observations in group i. We can fit and analyze models of this type using the `lme4` library [3] in the open-source statistical software `R` [5].

As SPEET main goal is to analyze the data trail of the student in real time in order to know as in advance as possible which profile the student belongs to, we decided to keep the set of covariates that are available at the time of the enrollment, and three more variables that could be recorded after the first semester of the first year of study. Thus, in the fixed-effect part of the model we consider the covariates that are described in table 1.

3 Results

We decide to randomly split the dataset into training and evaluation subsets, with a ratio of 80% for training and 20% for evaluation. Not all covariates turn out to be significant. We use stepwise backward elimination to obtain a reduced model: regressors `PreviousStudies` and `Nationality` are removed from the full model through this procedure. We can make the following considerations about fixed-effect parameters estimation:

- female students outperform their male counterpart: being a men penalize the log odds by 0.311. This information suggest that female students who decide to enroll in engineering degree are more resolute in their choice.
- The relationship between the linear predictor and the age at the time of the admission is negative. Generally, the aged students may have less time to devote to studies and this may affects their performance ($\hat{\beta} = -0.201$).

Variable	**Description**	**Type of variable**
`Sex`	sex	factor (2 levels: M, F)
`Nationality`	nationality	factor (2 levels: Italian, Foreigner)
`PreviousStudies`	high school studies	factor (3 levels: Liceo Scientifico, Istituto Tecnico, Other)
`AdmissionScore`	PoliMi admission test result	real number [0,100]
`AccessToStudiesAge`	age at the beginning of the BSc studies at PoliMi	natural number
`WeightedAvgEval1.1`	weighted average (by ECTS) of the evaluations during the first semester of the first year	real number [18,30]
`AvgAttempts1.1`	average number of attempts to be evaluated for both passed and not passed exams, during the first semester of the first year	real number [0,3]
`TotalCredits1.1`	number of ECTS credits obtained by the student during the first semester of the first year	natural number

Table 1: *Set of used covariates for the GLM model* (1)

- PoliMi admission test has a positive influence on the transformed response: a unit increase in the score improves the log odds by 0.008.
- The weighted average of the evaluations in the first semester shows strong positive effect on the transformed response: this result is reasonable and realistic. A unit increase improves the log odds by 0.060.
- `AverageExamAttempts1.1` has positive effect on the response: this information may suggest that if a student do not give up and tries to complete all his exams during the first semester of the first year, even after failing more than once, he is more likely to end successfully his programme ($\hat{\beta} = 0.243$).
- The effect of regressor `TotalCreditsObtained1.1` is positive as expected: a better performance during the first semester of the first year greatly improves the success probability in getting the degree ($\hat{\beta} = 0.196$).

Parameter	**Estimate**	**P-value**
`(Intercept)`	0.664	0.2354
`Sex(male)`	-0.311	0.0002
`AdmissionScore`	0.008	0.0124
`AccessToStudiesAge`	-0.201	4.87×10^{-11}
`WeightedAverageEvaluations1.1`	0.060	$<2 \times 10^{-16}$
`AverageExamAttempts1.1`	0.243	3.49×10^{-7}
`TotalCreditsObtained1.1`	0.196	$<2 \times 10^{-16}$

Table 2: *Fixed-effect coefficient estimates and P-values of model* (1)

As next step we underline the differences among the study programmes. Figure 1 shows the estimated random effects for all 19 groups in the dataset. Many of the 95% confidence intervals for $\hat{b}_i$ do not overlap the vertical line at zero, underlining substantial differences between the programmes. For example, level *Environmental and Land Planning Engineering* has the highest positive effect on the intercept: being a student from this programme improves the log odds by 1.486. On the contrary, studying *Civil Engineering* penalizes the log odds by 1.008.

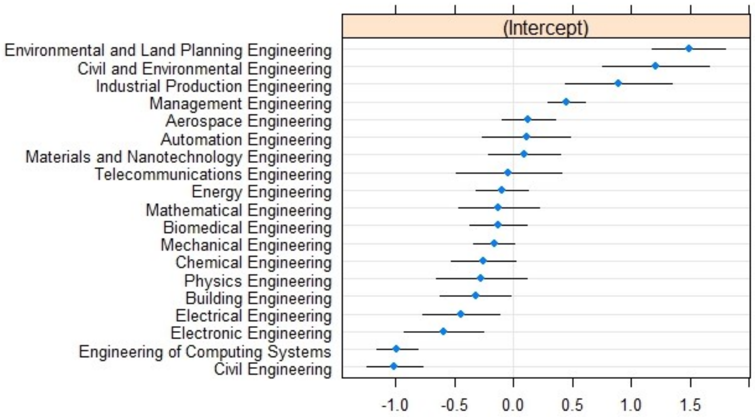


Fig. 1: *Random effect of the degree programme on the intercept in model* (1)

By using a multilevel model we can account for the interdependence of observations by partitioning the total variance into different components of variation due to the various cluster levels in the data. The intraclass correlation is equal to the percentage of variation that is found at the higher level and it is generally called the Variance Partition Coefficient [4]. Using the *latent variable approach* method the VPC is constant across all individuals and it is defined as $VPC = \frac{\sigma_b^2}{\sigma_b^2+\sigma_{lat}^2}$. Since the variance of the standard logistic distribution is $\sigma_{lat}^2 = \pi^2/3$ and the estimated variance is $\hat{\sigma}_b^2 = 0.4245$, the estimated VPC is equal to 0.1143. This means that 11.4% of variation in the response is attributed to the classification by degree type.

As last step, we use this model for classification and prediction of success probability. After fixing the optimal threshold value of $p_0 = 0.6$ using ROC curve analysis, we evaluate the predictive performance of the model by computing average classification indexes. We repeat 20 times the following procedure:

- randomly split the observations in training set (80% of the full dataset) and test set (20% of the full dataset)
- using the training set, fit the logit mixed-effect model and estimate its parameters
- estimate the success probability of observations in the test set and classify them using the optimal threshold value
- build the classification table and compute accuracy, sensitivity and specificity.

At the end of the 20 iterations we compute the average accuracy, sensitivity and specificity and their standard deviation, reported in table 3. High values of accuracy, sensitivity and specificity point to a good performance of the model. In addition, the model performance is notably robust, as highlighted by the low standard deviation of all performance indexes.

Index	Mean	Std deviation
Accuracy	0.899	0.0045
Sensitivity	0.925	0.0050
Specificity	0.850	0.0089

Table 3: *Performance indexes of a classifier based on the GLM* (1)

4 Conclusions

As far as the SPEET consortium knowledge, this is one of the first experiences of Learning Analytics at university level in Italy. Using predictive analytics we can give our educational institutes insights in future students outcomes: this predictions can be used to change particular programmes and deliver an optimal learning environment for the students.

Other further studies are being conducted within this project, proposing alternative nonparametric modeling in order to test the validity of the mixed-effect model and to analyse advantages and drawbacks of both methods. An immediate step further is extending the student profiling to other SPEET partners and compare the differences (if any) that arise at country level.

References

1. Agresti A., *Categorical Data Analysis*. 2nd ed. Wiley Series in Probability and Statistics. Wiley, 2007.
2. Barbu M., Vilanova R., Lopez Vicario J., Varanda M.J., Alves P., Podpora M., Prada M.A., Moran A., Torrebruno A., Marin S. and R. Tocu R., *SPEET Intellectual Output # 1, Data Mining Tool for Academic Data Exploitation, Literature review and first architecture proposal*, ERASMUS+ KA2 / KA203 (2017).
3. Bates D. *Lme4: Mixed-Effects Modeling With R*. 2010. `http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf`.
4. Goldstein H., Browne W., and Rasbash J.. *Partitioning Variation in Multilevel Models*. In: Understanding Statistics 1.4 (2002), pp. 223-231.
5. R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. `http://www.R-project.org/`
6. *SPEET, proposal for strategic partnerships (proposal narrative)* (2016) `https://www.speet-project.com/the-project`

dgLARS method for relative risk regression models

Il metodo dgLARS per i modelli di regressione del rischio relativo

Luigi Augugliaro and Angelo M. Mineo

Abstract With the introduction of high-throughput technologies in clinical and epidemiological studies, the need for inferential tools that are able to deal with fat data-structures, i.e., relatively small number of observations compared to the number of features, is becoming more prominent. To solve this problem, in this paper we propose an extension of the dgLARS method to relative risk regression model. The main idea of proposed method is to use the differential geometric structure of the partial likelihood function in order to select the optimal subset of covariates.
Abstract *L'introduzione di tecnologie di screening ad elevata capacità negli studi clinici ed epidemiologici ha reso preminente il problema dello sviluppo di metodologie inferenziali applicabili ai casi in cui la numerosità campionaria è inferiore al numero di parametri. In questo lavoro proponiamo un'estensione del metodo dgLARS ai modello di regressione del rischio relativo. L'idea di fondo del metodo proposto è quella di utilizzare la struttura geometrica della partial likelihood al fine di selezionare il sottoinsieme ottimo di variabili esplicative.*

Key words: dgLARS, relative risk regression models, sparsity, survival analysis.

1 Introduction

In the study of the dependence of survival time on covariates, the Cox proportional hazards model [3] has proved to be a major tool in many clinical and epidemiological applications. However, when the number of features is large, the simple Cox

Luigi Augugliaro
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: luigi.augugliaro@unipa.it

Angelo M. Mineo
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: angelo.mineo@unipa.it

proportional breaks down. Many variable selection techniques for linear regression models have been extended to the context of survival models. They include best-subset selection, stepwise selection, asymptotic procedures based on score tests, Wald tests and other approximate chi-squared testing procedures, bootstrap procedures and Bayesian variable selection. However, the theoretical properties of these methods are generally unknown. Recently a family of penalized partial likelihood methods, such as the Lasso [11] and the smoothly clipped absolute deviation method [5] were proposed for the Cox proportional hazards model. By shrinking some regression coefficients to zero, these methods select important variables and estimate the regression model simultaneously. Whereas the Lasso estimator does not possess oracle properties, the smoothly clipped absolute deviation estimator for linear models, has better theoretical properties. However, the non-convex form of the penalty term of the latter makes its optimization challenging in practice, and the solutions may suffer from numerical instability. In this paper we propose an alternative to the penalized inference methods. We extend the differential-geometric least angle regression method (dgLARS) [1] to the case of the Cox proportional hazards model.

2 The differential geometrical structure of a relative risk regression model

In analyzing survival data, one of the most important tools is the hazard function. Formally, let T be the absolutely continuous random variable associated with the survival time and let $f(t)$ be the corresponding probability density function. The hazard function is defined as $\lambda(t) = f(t)/\{1-\int_0^t f(s)ds\}$ and specifies the instantaneous rate at which failures occur for subjects that are surviving at time t. Suppose that the hazard function $\lambda(t)$ can depend on a p-dimensional vector of covariates which can depend on time and denoted by $\boldsymbol{x}(t) = (x_1(t),\ldots,x_p(t))^\top$. The relative risk regression models [10] are based on the assumption that the vector $\boldsymbol{x}(t)$ influence the hazard function $\lambda(t)$ by the following relation $\lambda(t;\boldsymbol{x}) = \lambda_0(t)\psi(\boldsymbol{x}(t);\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a p-dimensional vector of unknown fixed parameters, $\lambda_0(t)$ is the base hazard function at time t which is left unspecified and, finally, $\psi : \mathbb{R} \to \mathbb{R}$ is a fixed twice continuously differentiable function, called the relative risk function. The parameter space is such that $\psi(\boldsymbol{x}(t);\boldsymbol{\beta}) > 0$ for each $\boldsymbol{\beta}$; we also assume that the relative risk function is normalized, i.e., $\psi(\boldsymbol{0};\boldsymbol{\beta}) = 0$.

Suppose that n observations are available and let with t_i the ith observed failure time. Assume that we have k uncensored failure times and let us denoted by D the set of indices for which the corresponding failure time is observed. The remaining failure times are right censored. Under the assumption of independent censoring, the inference about $\boldsymbol{\beta}$ can be carried out by the following partial likelihood function

$$\mathscr{L}_p(\boldsymbol{\beta}) = \prod_{i\in D} \frac{\psi(\boldsymbol{x}_i(t_i);\boldsymbol{\beta})}{\sum_{j\in R(t_i)} \psi(\boldsymbol{x}_j(t_i);\boldsymbol{\beta})}, \tag{1}$$

where $R(t)$ denotes the risk set, i.e. the set of indices corresponding to the subjects how have not failed and are still under observation just prior to time t. In order to extend the dgLARS method to the relative risk regression model, it is useful to see the partial likelihood (1) as arising from a multinomial sampling scheme. Consider an index $i \in D$ and let $\boldsymbol{Y}_i = (Y_{ih})_{h \in R(t_i)}$ be a multinomial random variable with sample size equal to 1 and cell probabilities $\boldsymbol{\pi}_i = (\pi_{ih})_{h \in R(t_i)} \in \Pi_i$. Assuming that the random vectors $\boldsymbol{Y}_i$ are independent, the joint probability density function is an element of the set $S = \left\{\prod_{i \in D} \prod_{h \in R(t_i)} \pi_{ih}^{y_{ih}} : (\boldsymbol{\pi}_i)_{i \in D} \in \bigotimes_{i \in D} \Pi_i\right\}$. In the following of our differential geometric constructions, the set S will play the role of ambient space. Consider the following model for the conditional expected value of the random variable Y_{ih}: $E_{\boldsymbol{\beta}}(Y_{ih}) = \pi_{ih}(\boldsymbol{\beta}) = \psi(\boldsymbol{x}_h(t_i); \boldsymbol{\beta}) / \sum_{j \in R(t_i)} \psi(\boldsymbol{x}_j(t_i); \boldsymbol{\beta})$, then our model space is the set $M = \left\{\prod_{i \in D} \prod_{h \in R(t_i)} \pi_{ih}(\boldsymbol{\beta})^{y_{ih}} : (\boldsymbol{\pi}_i)_{i \in D} \in \bigotimes_{i \in D} \Pi_i\right\}$. The partial likelihood (1) is formally equivalent to the likelihood function associated with the model space M if we assume that for each $i \in D$, the observed y_{ih} is equal to one if h is equal to i and zero otherwise. Let $\ell(\boldsymbol{\beta})$ be the log-likelihood function associated to the model space M and let $\partial_m \ell(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \beta_m$. The tangent space $T_{\boldsymbol{\beta}} M$ of M at the model point $\prod_{i \in D} \prod_{h \in R(t_i)} \pi_{ih}(\boldsymbol{\beta})^{y_{ih}}$ is defined as that linear vector space spanned by the p elements of the score vector, formally $T_{\boldsymbol{\beta}} M = span\{\partial_1 \ell(\boldsymbol{\beta}), \ldots, \partial_p \ell(\boldsymbol{\beta})\}$. Under the standard regularity conditions, it is easy to see that $T_{\boldsymbol{\beta}} M$ is the linear vector space of the random variables $v(\boldsymbol{\beta}) = \sum_{m=1}^{p} v_m \partial_m \ell(\boldsymbol{\beta})$ with zero expected value and finite variance. As a simple consequence of the chain rule we have the following identity for any tangent vector belonging to the tangent space $T_{\boldsymbol{\beta}} M$, i.e.

$$v(\boldsymbol{\beta}) = \sum_{m=1}^{p} v_m \partial_m \ell(\boldsymbol{\beta}) = \sum_{i \in D} \sum_{h \in R(t_i)} \left(\sum_{m=1}^{p} v_m \frac{\partial \pi_{ih}(\boldsymbol{\beta})}{\partial \beta_m} \right) \frac{\partial \ell(\boldsymbol{\beta})}{\partial \pi_{ih}} = \sum_{i \in D} \sum_{h \in R(t_i)} w_{ih} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \pi_{ih}},$$

which shows that $T_{\boldsymbol{\beta}} M$ is a linear vector subspace of the tangent space $T_{\boldsymbol{\beta}} S$ spanned by the random variables $\partial_{ih} \ell(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \pi_{ih}$. To define the notion of angle between two given tangent vectors belonging to $T_{\boldsymbol{\beta}} M$, say $v(\boldsymbol{\beta})$ and $w(\boldsymbol{\beta})$, we shall use the information metric [9], i.e.

$$\langle v(\boldsymbol{\beta}); w(\boldsymbol{\beta}) \rangle_{\boldsymbol{\beta}} = E_{\boldsymbol{\beta}}(v(\boldsymbol{\beta}) w(\boldsymbol{\beta})) = \boldsymbol{v}^{\top} I(\boldsymbol{\beta}) \boldsymbol{w}, \tag{2}$$

where $\boldsymbol{v} = (v_1 \ldots, v_p)^{\top}$, $\boldsymbol{w} = (w_1 \ldots, w_p)^{\top}$ and $I(\boldsymbol{\beta})$ is the Fisher information matrix evaluated at $\boldsymbol{\beta}$. As observed in [6], the matrix $I(\boldsymbol{\beta})$ used in (2) is not exactly equal to the Fisher information matrix of the relative risk regression model, however it has the appropriate asymptotic properties for the inference [8].

3 dgLARS method for relative risk regression model

dgLARS method is a sequential method developed to estimate a sparse solution curve embedded in the parameter space based on a differential geometric characterization of the Rao score test statistic obtained considering the inner product between the bases of the tangent space $T_{\boldsymbol{\beta}}M$ and the tangent residual vector $r(\boldsymbol{\beta}) = \sum_{i\in D}\sum_{h\in R(t_i)} r_{ih}(\boldsymbol{\beta})\partial_{ih}\ell(\boldsymbol{\beta}) \in T_{\boldsymbol{\beta}}S$, where $r_{ih}(\boldsymbol{\beta}) = y_{ih} - \pi_{ih}(\boldsymbol{\beta})$. As observed in [1], the mth signed Rao score test statistic satisfies the following differential geometric characterization, i.e.

$$r^u_m(\boldsymbol{\beta}) = I_{mm}^{-1/2}(\boldsymbol{\beta})\partial_m\ell(\boldsymbol{\beta}) = \cos(\rho_m(\boldsymbol{\beta}))\|r(\boldsymbol{\beta})\|_{\boldsymbol{\beta}}, \tag{3}$$

where $I_{mm}(\boldsymbol{\beta})$ is the Fisher information for β_m, $\|r(\boldsymbol{\beta})\|^2_{\boldsymbol{\beta}} = E_{\boldsymbol{\beta}}(r(\boldsymbol{\beta})^2)$ and $\cos(\rho_m(\boldsymbol{\beta}))$ is a generalization of the Euclidean notion of angle between the mth column of the design matrix and the residual vector. Characterization (3) gives us a natural way to generalize the equiangularity condition [4]: two given predictors, say the mth and nth, satisfy the generalizes equiangularity condition at the point $\boldsymbol{\beta}$ when $|r^u_m(\boldsymbol{\beta})| = |r^u_n(\boldsymbol{\beta})|$. Inside the dgLARS theory, the generalized equiangularity condition is used to identify the predictors that are included in the model.

The nonzero estimates are formally defined as follows. For any data set there is a finite sequence of transition points, say $\gamma^{(1)} \geq \ldots \geq \gamma^{(K)} \geq 0$, such that for any fixed γ between $\gamma^{(k+1)}$ and $\gamma^{(k)}$ the sub vector of the non nonzero dgLARS estimates, denoted as $\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma) = (\hat{\beta}_m(\gamma))_{m\in\hat{\mathscr{A}}}$, satisfies the following conditions:

$$\begin{aligned} r^u_m\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma)\} &= s_m\gamma, \quad m \in \hat{\mathscr{A}} \\ |r^u_n\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma)\}| &< \gamma, \quad n \notin \hat{\mathscr{A}} \end{aligned}$$

where $s_m = \text{sign}\{\hat{\beta}_m(\gamma)\}$ and $\hat{\mathscr{A}} = \{m : \hat{\beta}_m(\gamma) \neq 0\}$, called active set, is the set of the indices of the predictors that are included in the current model, called active predictors. In any transition point, say for example $\gamma^{(k)}$, one of the following two conditions occurs:

1. there is a non active predictor, say the nth, satisfying the generalized equiangularity condition with any active predictor, i.e.,

$$|r^u_n\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma^{(k)})\}| = |r^u_m\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma^{(k)})\}| = \gamma^{(k)}, \tag{4}$$

   for any m in $\hat{\mathscr{A}}$, then it is included in the active set;
2. there is an active predictor, say the mth, such that

$$\text{sign}[r^u_m\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma^{(k)})\}] \neq \text{sign}\{\hat{\beta}_m(\gamma^{(k)})\}, \tag{5}$$

   then it is removed from the active set.

Given the previous definition, the path of solutions can be constructed in the following way. Since we are working with a class of regression models without intercept term, the starting point of the dgLARS curve is the zero vector this means that, at the starting point, the p predictors are ranked using $|r^u_m(\mathbf{0})|$. Suppose that $a_1 = \arg\max_m |r^u_m(\mathbf{0})|$, then $\hat{\mathscr{A}} = \{a_1\}$, $\gamma^{(1)}$ is set equal to $|r^u_{a_1}(\mathbf{0})|$ and the first segment of the dgLARS curve is implicitly defined by the nonlinear equation $r^u_{a_1}\{\hat{\beta}_{a_1}(\gamma)\} - s_{a_1}\gamma = 0$. The proposed method traces the first segment of the dgLARS curve reducing γ until we find the transition point $\gamma^{(2)}$ corresponding to the inclusion of a new index in the active set, in other words, there exists a predictor, say the a_2th, satisfying condition (4), then a_2 is included in $\hat{\mathscr{A}}$ and the new segment of the dgLARS curve is implicitly defined by the system with nonlinear equations:

$$r^u_{a_i}\{\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma)\} - s_{a_i}\gamma = 0, \quad a_i \in \hat{\mathscr{A}},$$

where $\hat{\boldsymbol{\beta}}_{\hat{\mathscr{A}}}(\gamma) = (\hat{\beta}_{a_1}(\gamma), \hat{\beta}_{a_2}(\gamma))^\top$. The second segment is computed reducing γ and solving the previous system until we find the transition point $\gamma^{(3)}$. At this point, if condition (4) occurs a new index is included in $\hat{\mathscr{A}}$ otherwise condition (5) occurs and an index is removed from $\hat{\mathscr{A}}$. In the first case the previous system is updated adding a new nonlinear equation while, in the second case, a nonlinear equation is removed. The curve is traced as previously described until parameter γ is equal to a fixed value that can be zero, if the the sample size is large enough, or a positive value if we are working in a high-dimensional setting, i.e., the number of predictors is larger than the sample size. In this way we can avoid the problems coming from the overfitting of the model. From a computational point of view, the entire dgLARS curve can be computed using the algorithms proposed in [2, 7].

4 Simulation study

In this section we compare the method introduced in Section 3 with three popular algorithms named CoxNet, CoxPath, and CoxPen. Given the fact that these methods have only been implemented only for Cox regression model, our comparison will focus on this kind of relative risk regression model. In the following of this section, dgLARS method applied to the Cox regression model is named dgCox model.

We simulated one hundred datasets from a Cox regression model where the survival times t_i $(i = 1, \dots, n)$ follow an exponential distributions with parameter $\lambda_i = \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)$, and $\boldsymbol{x}_i$ is sampled from a p-variate normal distribution $\mathrm{N}(\mathbf{0}, \Sigma)$; the entries of Σ are fixed to $\mathrm{corr}(X_m, X_n) = \rho^{|m-n|}$ with $\rho = 0.9$. The censorship is randomly assigned to the survival times with probability $\pi \in \{0.2, 0.4\}$. To emulate an high-dimensional setting, we fixed the sample size to 50, the number of predictors to 100 and $\beta_m = 0.5$ $(m = 1, \dots, 30)$; the remaining regression coefficients are zero in order to have a sparse vector. To remove the effects coming from the information measure used to select the optimal point of each paths of solutions, we evaluated the global behaviour of the paths by using the ROC curve and the corresponding Area Under the Curve (AUC). Figure 1 shows that dgCox model is clearly the superior approach for both levels of censorship. For the same false positive rate, the true positive rate of the dgCox method is around 10% higher than the rate obtained by CoxNet, CoxPath and CoxPen.

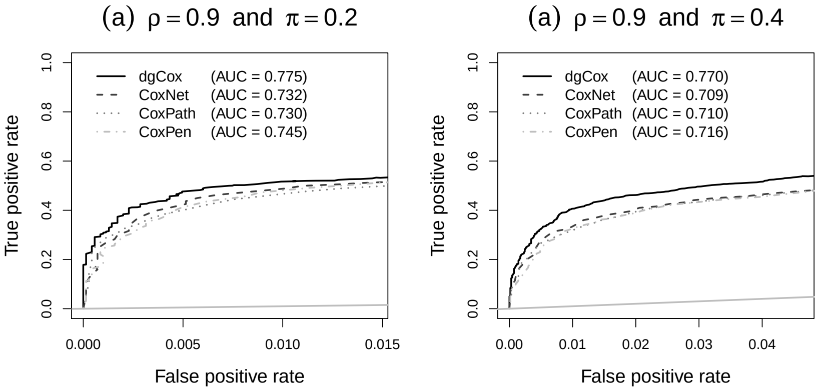


Fig. 1 Results from the simulation study; for each scenario we show the averaged ROC curve for dgCox, CoxNet, CoxPath and CoxPen algorithm. The average Area Under the Curve (AUC) is also reported. The 45-degree diagonal is also included in the plots.

References

1. Augugliaro L., Mineo, A. M., Wit, E.: Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. J. Roy. Statist. Soc. Ser. B. **75**(3), 471–498 (2013)
2. Augugliaro L., Mineo, A. M., Wit, E.: dglars: An R package to estimate sparse generalized linear models. J. Stat. Soft. **59**(8), 1–40 (2014)
3. Cox, D.: Regression models and life-tables. J. Roy. Statist. Soc. Ser. B. **34**(2), 187–220 (1972)
4. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Statist. **32**(2), 407–499 (2004)
5. Fan, J., Li, R.: Variable selection for Cox?s proportional hazards model and frailty model. Ann. Statist. **30**(1), 74–99 (2002)
6. Moolgavkar, S., Venzon, D. J.: Confidence regions in curved exponential families: application to matched case-control and survival studies with general relative risk function. Ann. Statist. **15**(1), 346–359 (1987)
7. Pazira, H., Augugliaro, L., Wit, E. C.: Extended differential geometric lars for high-dimensional glms with general dispersion parameter. Stat. Comput. **28**(4), 753–774 (2018)
8. Prentice, R., Self, S.: Asymptotic distribution theory for Cox-type regression models with general relative risk form. Ann. Statist. **11**(3), 804–813 (1983)
9. Rao, C. R.: On the distance between two populations. Sankhyā. **9**, 246–248 (1949)
10. Thomas, D.: General relative-risk models for survival time and matched case-control analysis. Biometrika. **37**(4), 673–686 (1981)
11. Tibshirani, R.: The lasso method for variable selection in the Cox model. Stat. Med. **16**, 385–395 (1997)

A Latent Class Conjoint Analysis for analysing graduates' profiles

Un modello Latent Class Conjoint per l'analisi dei profili dei laureati

Paolo Mariani, Andrea Marletta, Lucio Masserini and Mariangela Zenga

Abstract This paper aims to stabilize the relationship between universities and companies. Lombardy companies with at least 15 employees were asked them to manifest their preferences choosing among profiles of new graduates. A Latent Class Metric Conjoint Analysis is employed to evaluate the ideal graduate's profile for a job position and to detect the existence of subgroups of companies having homogeneous preferences about such features.
Abstract *Questo lavora mira ad analizzare la relazione tra aziende ed università. Alle imprese lombarde con almeno* 15 *dipendenti è stato chiesto di esprimere le loro preferenze fra alcuni profili dei neolaureati a loro sottoposti per una possibile nuova assunzione. Un modello Latent Class Conjoint è stato utilizzato per valutare il profilo ideale fra i candidati e individuare l'esistenza di sottogruppi di imprese.*

Key words: Labour Market, Latent Class Models, Conjoint Analysis, Electus

1 Introduction

During last years, the economic crisis exhibits effects about performances in business and particularly in the employment in all European countries. The impact of this crisis struck weaker segments of the labour market, in detail young person

Paolo Mariani
University of Milano-Bicocca, e-mail: paolo.mariani@unimib.it

Andrea Marletta
University of Milano-Bicocca, e-mail: andrea.marletta@unimib.it

Lucio Masserini
University of Pisa, e-mail: lucio.masserini@unipi.it

Mariangela Zenga
University of Milano-Bicocca, e-mail: mariangela.zenga@unimib.it

and people with less work experience. Most of the times, there is no possibility for turnover, so the younger people are not able to access the labour market. In Italy, according to Istat, from 2007 over 2014, the young unemployment rate (15-24 years) increased from 20.4% over to 44.7%.

In relation to the labour market, it reveals evident how the other side of the phenomena is represented by the companies and their expectations about the possibility for a new hiring. It appears useful to carry out information from the companies' point of view obtaining a deep analysis about what they are looking for.

In a perspective of synergy between education and labour market, a possible solution is represented by ELECTUS, an acronym standing for Education-for-Labour Elicitation from Companies' Attitudes towards University Studies, a research project involving several Italian universities [3]. The ELECTUS survey was conducted in 2015 using CAWI technique using a questionnaire containing two macro-sections. In the first part the entrepreneurs are asked to choose and rank four possible profiles of new graduates for five different job vacancies. In the second part the entrepreneurs are asked about their socio-demographic features.

The candidates' profile are characterized by six attributes: *Field of Study*, *Degree Mark*, *Degree Level*, *English Knowledge*, *Relevant Work Experience*, *Willingness to Travel on Business*.

The aim of this paper is to carry out a segmentation analysis of employers' preferences for graduates' profiles evaluated as candidates in a job position by using a Latent Class Metric Conjoint Analysis [1]. In fact, it has been highlighted that the tandem approach suffers from some problems: a) different clustering methods often yield different results, in terms of number of clusters and their composition; b) in presence of highly fractionated designs, as in our study, individual-level part-worth estimates are rather unstable and then may be untrustworthy when employed in successive clustering algorithms. On the other hand, the LCMCA, unlike cluster analysis, is a model-based approach in which model parameters and subgroups (segments or latent classes) are estimated simultaneously, and segments are composed of individuals whose part-worth utilities are similar.

The paper is organized as follows. Section 2 introduces Latent Class Metric Conjoint Analysis. Section 3 presents the results of the estimated model. Finally, Section 4 is reserved to discussion and final remarks.

2 Latent Class Conjoint Analysis

In this study, a Latent Class Metric Conjoint Analysis (LCMCA) [1] is employed is carried out to evaluate which characteristics of a graduate's profile employers prefer for a potential candidate in the job position of administrative clerk. In particular, in order to detect if there exist unobserved subgroups of employers having homogeneous preferences about graduates' characteristics for this position. Latent Class Metric Conjoint Analysis is a statistical modelling technique included in the more general class of Finite Mixture Models (FMMs) [4]. Following the FMMs approach, LCMCA relaxes the single homogeneous population assumption to allow for parameter differences across G latent classes and supposes that the marginal density function of the response variable y is given by a weighted sum over the G mixture components, with weights indicating the *a-priori* probability for an observation to come from a specific component [1]:

$$f(y_{ij}|\pi, x, z, \Sigma) = \sum_{g=1}^{G} \pi_{g|z} f_g(y_{ij}|x, z, \beta_g \Sigma_g) \tag{1}$$

where y_{ij} is the vector of response variable which refers to the rating expressed by employer i to conjoint profile j; $\pi = (\pi_1, \pi_2, \ldots, \pi_{G-1})$ are $G-1$ independent mixing proportions of the mixture, such that $0 \leq \pi_g \leq 1$; x is a matrix containing the M conjoint dummy variables which defines the profiles evaluated; β_g is the vector of the estimated conjoint part-worth coefficients for subgroup g, and $\Sigma = (\Sigma_1 \Sigma_2, \ldots, \Sigma_g)$ is $J \times J$ vector of covariance matrices of the error terms estimated for each subgroup. Moreover, given that the weights depend on a set of explanatory variables, also referred to as concomitant variables, z defines the vector of variables which characterise employers. Given the metric response variable, each of the conditional distributions, f_g, is conventionally specified as a conditional multivariate normal distribution. Instead, the prior probability of group membership varies as a multinomial logistic regression model, in function of the concomitant variables, as it follows:

$$\pi_{k|z} = \frac{exp(\gamma_{0g} + z\gamma_{1g})}{\sum_{g=1}^{G} exp(\gamma_{0g} + z\gamma_{1g})} \tag{2}$$

where γ_{0g} is the intercept while γ_{1g} contains the vector of regression coefficients, quantifying the effect of the concomitant variables on the prior probability for class g. For identification purpose, usually $\gamma_{01} = 0$ and $\gamma_{11} = 0$, and designate the first category as a reference class. Moreover, the following constraints hold: $\sum_{g=1}^{G} \pi_{g|z} = 1, \pi_{g|z} > 0$. Once the estimates of all the model parameters of the mixture of probability density are obtained, the posterior probability that an observation belongs to class g, denoted with p_{ig}, can be calculated by updating the previous according to the Bayes' theorem, as follows:

$$\hat{p}_{ig} = \frac{\hat{\pi}_{g|z} \hat{f}_{ig}(y_{ij}|x, z, \beta_g \Sigma_g)}{\sum_{g=1}^{G} \hat{\pi}_{g|z} \hat{f}_{ig}(y_{ij}|x, z, \beta_g \Sigma_g)} \tag{3}$$

where $\sum_{g=1}^{G} \hat{p}_{ig} = 1$ and $0 \leq \hat{p}_{ig} \leq 1$. The posterior probabilities provide a probabilistic allocation of observations to the latent classes and can be used to classify data by assigning each employer to the class with the maximum posterior probability. Parameter estimation was carried out via Maximum Likelihood (ML) by using the Expectation-Maximization (E-M) algorithm [2], in which conjoint part-worth coefficients and class membership are obtained simultaneously. The conventional Akaike's information criterion (AIC) was used for choosing the number of latent classes.

3 Application and Results

After estimating several LCMCAs, starting from the aggregate solution, with one class ($G = 1$), to the most complex one, with five latent classes ($G = 5$), the solution with $G = 3$ seems to represent the latent structure underlying the employers' ratings quite well.

Table 1 provides the estimated regression coefficients (or part-worths) for the $G = 3$ latent classes solution. Given that categorical attributes were preventively converted into the appropriate number of dummy variables, the intercept represents the average rating of the reference profile and refers to a graduate in foreign languages, with a Master's degree, a low final grades, with a knowledge of English language, no working experience, and not willing to business trips. At a first glance, it is clear the substantial difference between the aggregate part-worth coefficients and those in each sub-model of the three-class solution. Moreover, the aggregate model is also the one with the highest value of AIC and this confirms that a single set of regression coefficients estimated for all the employers may produce misleading results. The first class is the one with the lowest average rating corresponding to the reference profile (2.56) and identifies especially Economics as the most preferred degree, whereas Law, Statistics and Engineering are also appreciated but to a lesser extent. Among the employers included in this class, it seems preferable for a candidate to have some kind of work experience, excluding the internship. On the other hand, low final grades and willing to long-term business trips produce a lower preference. The second class is the one with the higher average rating (8.25) and this indicates that the reference profile is already highly appreciated. Given such a high average score, part-worth coefficients are almost all negative. In particular, employers within such class evaluate Economics, Engineering, Mathematics and computer sciences as less important degrees. On the contrary, a Bachelor's degree and a medium final grades increase employers' preference. The third class is the one with the intermediate average rating and its value is also very similar to that of the aggregate model (4.46). Mathematics and computer sciences and Economics are the most preferred degrees. Political science is also evaluated positively but to a lesser extent. For employers in this class, a previous work experience both as a stable experience and internship experience is relevant.

The Maximum likelihood estimates of the multinomial logistic regression model allows to determine which variables affect the latent classes' membership. Among the available variables which describe the characteristics and the context of the company, only the variables 'Recruitment of staff within one year', 'Company run by a manager' and 'Company committed also in the foreign market' seem to affect class membership when using the usual 0.05 as significance level. However, considering 0.10 as significance level also the variables 'Hired personnel over the past 3 years' and 'Education of the last administrative hired: graduated' contribute to explain class membership. In particular, the probability of being in Class 2 is higher for companies which plan to recruit staff within one year ($p = 0.007$) but lower for those that hired personnel over the past 3 years ($p = 0.070$) and those which have already taken a graduated as administrative ($p = 0.076$). On the other hand, the

Table 1 Maximum likelihood estimates of part-worth coefficients for each latent class

		Latent	class	
	1	2	3	Aggregate
Intercept	2.56**	8.25**	4.46**	4.79**
Philosophy and Literature	2.60**	-1.65**	-2.20	-0.88
Education sciences	1.54*	-0.02	-0.48	0.15
Political sciences	1.21	-1.23**	1.45**	0.97
Economics	6.52**	-5.21**	2.18**	1.92**
Law	4.29**	-3.75**	-0.47	-0.25
Statistics	3.99**	-2.77**	-0.63	0.42
Engineering	3.81**	-4.74**	-2.25**	-1.16*
Mathematics and computer sciences	2.68**	-5.41**	2.92**	0.45
Psychology	2.80**	-3.66**	-2.42**	-1.61*
Bachelor's degree	-0.23	0.59**	0.47	0.49
Low final grades	-1.62**	-1.13**	-1.01**	-1.25**
Medium final grades	-0.92*	1.70**	-0.94**	-0.50
No knowledge of English language	0.49	-0.75**	-1.57**	-0.87**
Internship experience	-0.56	0.84**	1.37**	0.46
Occasional working experience	2.10**	-0.25**	0.82	0.66
Stable working experience	1.04*	-2.28**	2.36**	1.07**
Willing to short-term business travels	-0.44	-1.46**	0.85**	-0.15
Willing to long-term business travels	-1.33*	-0.88**	0.53	-0.10

probability of being in Class 3 is higher for companies which plan to recruit staff within one year ($p < 0.0001$), for companies run by a manager ($p = 0.036$) and for companies committed also in the foreign market ($p = 0.033$) but lower for those that hired personnel over the past 3 years ($p = 0.072$). It seems adequate to identify the three latent groups by their peculiar features. In particular, the first group (26.4% of the sample) is characterized by companies lead by not a managerial view, working in a service sector in prevalence in domestic market, they neither will do recruitment new staff in the next year neither hired personnel over the past three years. This group could be named *Domestic Consolidated Companies*. The best profile for the AC required by these companies results to be related to a classical view of the position: a well graduated in Economics with some kind of working experience. The second group (17.1% of the sample) is composed by big sized companies, they will recruit staff in the next year, but they did not hire personnel over the past three years. This group could be called as *Static Companies*: respect to the expected profile of new graduates they seem to prefer new graduates with a not suitable major for AC (Language) with Bachelor's degree and a medium final grade, the English knowledge and working experience are not required. In some way, they prefer new graduates who must be fully trained. The third group (56.5% of the sample) is represented by small or medium enterprises, guided by a manager and committed also in the foreign market with a willingness to recruit new staff in the next year. They can be named as *Dynamic Companies*. The new graduate profile fits the description of these firms: it has to be a student in Economics or Political Sciences major, with a higher final grade and an English knowledge suitable to communicate with foreign people. Important requirements are also working experience and willing to business trips.

4 Discussion and final remarks

Using the survey ELECTUS, a segmentation of employers' preferences for graduates' profiles for administrative clerk is carried out by using a Latent Class Metric Conjoint Analysis. In general the features of the profiles for the new graduates' job are very different for every sub-groups but all respondents agree that a low final grades at graduation is not a preferable. Certainly the characteristics of the companies could influence the preferences about graduates' characteristics: the membership of the latent groups seems in fact to be effected by peculiar factors. In fact, the *Domestic Consolidated Companies*, run by not a managerial view, working in a service sector in prevalence in domestic market, without willingness to recruit in the past and in the future, require a well graduated students in Economics with some kind of working experience for AC position. The *Static Companies*, composed by big sized companies with the willingness to recruit in the future but not in the past, prefer new graduates who must be fully trained. The *Dynamic Companies*, represented by SME lead by a manager and committed also in the foreign market with a willingness to recruit new staff in the next year, look for a new graduate in Economics or Political Sciences major, with a higher final grade and an English knowledge suitable to communicate with foreign people with working experience and willing to business trips.

References

1. DeSarbo, W.S, Wedel, M.A., Vriens, M., Ramaswamy, V.: Latent Class Metric Conjoint Analysis. Marketing Letters, 3(2), 137-288 (1992).
2. Dempster, A., Laird, N., Rubin, D.: Maximum Likelihood from Incomplete Data via the EM-Alogrithm. Journal of the Royal Statistical Society: Series B, 39(1), 1-38 (1977).
3. Fabbris, L. and Scioni, M. Dimensionality of scores obtained with a paired-comparison tournament system of questionnaire item. In A. Meerman and T. Kliewe, eds., Academic Proceedings of the 2015 University-Industry Interaction Conference (2015).
4. McLachlan, G. J., Peel, D.: Finite Mixture Models. New York: Wiley (2000).
5. Wedel, M.A.: Concomitant variables in finite mixture models. Statistica Neerlandica, 56(3), 362-375 (2002).

A longitudinal analysis of the degree of accomplishment of anti-corruption measures by Italian municipalities: a latent Markov approach

Analisi longitudinale del grado di attuazione di misure anti-corruzione nei comuni italiani: un approccio latent Markov

Simone Del Sarto, Michela Gnaldi, Francesco Bartolucci

Abstract The recent Italian anti-corruption law has introduced a new figure, the supervisor for corruption prevention, who has to fill in an annual report about the accomplishment of anti-corruption measures within the institution he/she represents. Using data coming from such annual reports referred to a sample of Italian municipalities, a latent Markov model is fitted to investigate the evolution over time of the degree of accomplishment of anti-corruption measures. First results evidence three latent states of increasing virtuosity. Moreover, at the beginning of the study, the most of the sample belongs to the low and intermediate states of virtuosity, even if there is evidence of high probabilities to move to upper states over time.

Abstract *La recente legge anti-corruzione italiana ha introdotto una nuova figura, il responsabile per la prevenzione della corruzione, che deve predisporre una relazione annuale sull'attuazione di misure anti-corruzione in capo all'istituzione che rappresenta. Per analizzare i dati provenienti da tali relazioni annuali riferiti a un campione di comuni italiani, è stato stimato un modello latent Markov, che permette di investigare l'evoluzione temporale del grado di attuazione di misure anti-corruzione. I primi risultati evidenziano l'esistenza di tre stati latenti corrispondenti a gradi crescenti di virtuosità. Inoltre, agli inizi dello studio, la maggior parte del campione appartiene ai primi due stati (bassa e intermedia virtuosità), sebbene vi siano alte probabilità di spostarsi nel tempo in stati corrispondenti a gradi più alti di virtuosità.*

Key words: Corruption prevention, latent Markov, supervisor for corruption prevention

Simone Del Sarto
Italian National Institute for the Evaluation of the Education System (INVALSI), e-mail: simone.delsarto@email.com

Michela Gnaldi
Department of Political Science, University of Perugia e-mail: michela.gnaldi@unipg.it

Francesco Bartolucci
Department of Economics, University of Perugia, e-mail: francesco.bartolucci@unipg.it

1 Introduction

The recent approaches to combat corruption have seen a shift from a penal and repressive focus to a broader preventive approach. Attention has to be paid not only on all illegal and criminal conducts, but also on any social behaviour and malpractice, even if it is not framed in a specific type of penal offence against public administration [2].

The Italian legislation has recently adopted such preventive perspective with law n. 190 of 2012, named “Provisions for the prevention and repression of corruption and lawlessness in the public administration”. It introduces two main tools to be implemented within all Italian public administrations, in order to reduce the risk of occurrence of corruptive events. The first is a three-year plan for corruption prevention (PTPC), by which each administration evaluates its own internal situation in terms of exposure levels to the risk of corruption and specifies potential organisational changes to reduce such a risk. The second is the introduction of a new figure, the supervisor for corruption prevention (RPC) that, within the public administrative unit (i.e., regions, municipalities, etc.) he/she represents, fills in an annual report about the degree of accomplishment of anti-corruption measures. This report is based on a questionnaire made available by the Italian National Anticorruption authority (ANAC). Previous works [3, 4, 5] exploited the data contained in the RPC forms for investigating the degree of accomplishment of anti-corruption measures by Italian municipalities, with the aim of ascertaining clusters of units characterised by distinctive anti-corruption behaviours, their geographical distribution, and the association between anti-corruption behaviours and some relevant covariates (e.g., the municipality size).

The objective of the present work is to extend previous findings by deepening our understanding as regards the evolution over time of anti-corruption behaviours of Italian municipalities. A statistical model suitable for the aim at issue is the Latent Markov (LM) model, whose original formulation is due to Wiggins [7]; for an up to date review, see Bartolucci et al. [1]. The LM model is tailored to the analysis of longitudinal dataset, specifically made of several response variables observed at different time occasions for each unit. It allows us to cluster the sample units in latent states as regards the underlying latent variable (degree of accomplishment of anti-corruption measures in our case) and to describe the evolution over time of the transitions among latent states.

The reminder of this work is organised as follows. Section 2 briefly describes the data used for this analysis, while the LM model is introduced in Section 3. Section 4 shows the results, while some concluding remarks are provided in Section 5.

2 The data

The data we consider are collected through the RPC forms for the years from 2014 – the year in which the anti-corruption law entered into force – to 2016. The sample is made of 213 municipalities, comprising all Italian province municipalities, all the other municipalities with at least 40,000 inhabitants and particular "advised" municipalities, as stated by ANAC act n. 71 of 2013.

The objective of the present work is to sketch the evolution over time of the Italian municipalities attitude as regard the degree of implementation of anti-corruption measures. For this reason, we focus on ten questions of the RPC form, requiring the administrations to state whether a specific anti-corruption measure has been accomplished in the reference year. The selected ten questions concern the following subjects:

1. monitoring the sustainability of all measures, general and specific, identified in the PTPC;
2. specific measures, in addition to mandatory ones;
3. computerising the flaw to fuel data publication in the "transparent administration" website section;
4. monitoring data publication processes;
5. training of employees, specifically dedicated to prevention of corruption;
6. staff turnover as a risk prevention measure;
7. checking the truthfulness of statements made by parties concerned with unfitness for office causes;
8. measures to verify the existence of incompatibility conditions;
9. prearranged procedures for issuing permits for assignments performance;
10. whistleblowing, which is a procedure for reporting the collection of misconduct by public administration employees.

Three possible answers can be provided to these items, ordered according to their virtuosity: (A) "Yes, the anti-corruption measure has been accomplished" (the most virtuous behaviour); (B) "No, the anti-corruption measure has not been accomplished because it was not expected by the PTPC" (intermediate level of virtuosity); (C) "No, the anti-corruption measure has not been accomplished but it was expected by the PTPC" (the least virtuous conduct).

3 The latent Markov model

Suppose that for every unit we observe the vector of r response variables at occasion $t = 1,\dots,T$, denoted by $\boldsymbol{Y}^{(t)} = [Y_1^{(t)},\dots,Y_r^{(t)}]^\top$. Each variable $Y_j^{(t)}$ is categorical with l_j modalities, coded from 0 to $l_j - 1$, $j = 1,\dots,r$. We assume that a latent process $\boldsymbol{U} = [U^{(1)},\dots,U^{(T)}]^\top$ affects the response variables: such a process follows a first-order Markov chain with state space $\{1,\ldots,k\}$. Local independence is assumed for the variables in each $\boldsymbol{Y}^{(t)}$, so its r components are conditionally independent given $U^{(t)}$.

The model at issue has the following parameters:

- $\phi_{jy|u}$: conditional response probability that component $Y_j^{(t)}$ assumes modality y, given latent space u, with $j=1,\ldots,r$, $t=1,\ldots,T$, $y=0,\ldots,l_j-1$ and $u=1,\ldots,k$;
- π_u: initial probabilities, with $u=1,\ldots,k$;
- $\pi_{u|\bar{u}}^{(t)}$: transition probabilities from state $\bar{u}$ to state u at time t, with $u,\bar{u}=1,\ldots,k$ and $t=2,\ldots,T$.

The model assumptions imply that

$$f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u})=\prod_{t=1}^{T}\phi_{\boldsymbol{y}^{(t)}|u^{(t)}},$$

where $\boldsymbol{Y}$ is a vector obtained by the union of the vectors $\boldsymbol{Y}^{(t)}$, $t=1,\ldots,T$ and in general $\phi_{\boldsymbol{y}|u}$ is the probability that $\boldsymbol{Y}^{(t)}$ assumes value $\boldsymbol{y}$, given latent state u. Moreover, given the local independence assumption, $\phi_{\boldsymbol{y}|u}$ can be obtained as the product of the single conditional probabilities over the r components, as follows:

$$\phi_{\boldsymbol{y}|u}=\prod_{j=1}^{r}\phi_{jy_j|u}.$$

The model parameters are estimated by maximum likelihood through an Expectation-Maximisation algorithm: see [1] for details.

4 Results

According to the Bayesian Information Criterion (BIC) [6], we find evidence of $k=3$ latent states. By inspecting the conditional response probabilities (see Table 1), it is possible to notice that such states correspond to increasing levels of corruption prevention fulfilment. The estimates of the initial probabilities are reported in Table 2(a). Hence, in 2014 (beginning of the study), the state of the most virtuous municipalities (i.e., those grouped in the third latent state) collects less than 15% of administrations, while the other two states approximately equally split the remaining units.

Regarding the transition probabilities – see Table 2(b) and 2(c) – we note that the probabilities to remain in the current state are higher than those to move to other states and such probabilities are increasing over time. However, from 2014 to 2015 units belonging to the first two states (characterised by low and intermediate virtuosity) have large probability to move to higher state, hence to increase their compliance of anti-corruption measures. In 2016 the situation seems to stabilise, since almost all the off-diagonal elements of the relative transition matrix are close to 0, expect for a share of probability (around 20-30%) to move to the third state for units belonging to the first two states.

Table 1 Estimates of the conditional response probabilities

Item 1				Item 6			
		State				State	
Response	1	2	3	Response	1	2	3
A	0.526	0.866	0.929	A	0.193	0.471	0.647
B	0.126	0.022	0.022	B	0.253	0.156	0.131
C	0.348	0.111	0.049	C	0.554	0.373	0.222
Item 2				Item 7			
		State				State	
Response	1	2	3	Response	1	2	3
A	0.332	0.792	0.830	A	0.096	0.145	0.839
B	0.112	0.000	0.005	B	0.136	0.018	0.148
C	0.556	0.208	0.165	C	0.769	0.837	0.013
Item 3				Item 8			
		State				State	
Response	1	2	3	Response	1	2	3
A	0.511	0.729	0.819	A	0.086	0.201	0.833
B	0.048	0.025	0.072	B	0.125	0.030	0.137
C	0.441	0.246	0.110	C	0.789	0.769	0.029
Item 4				Item 9			
		State				State	
Response	1	2	3	Response	1	2	3
A	0.744	0.958	0.946	A	0.622	0.953	0.923
B	0.070	0.000	0.009	B	0.072	0.012	0.026
C	0.185	0.042	0.045	C	0.306	0.035	0.051
Item 5				Item 10			
		State				State	
Response	1	2	3	Response	1	2	3
A	0.729	0.945	0.937	A	0.276	0.899	0.841
B	0.202	0.036	0.063	B	0.230	0.016	0.120
C	0.069	0.019	0.000	C	0.493	0.085	0.038

A: anti-corruption measure accomplished; B: anti-corruption measure not accomplished because it was not expected by the PTPC; C: anti-corruption measure not accomplished even if it was expected by the PTPC.

Table 2 Estimates of the initial probabilities (a) and of the transition probalities from 2014 to 2015 (b) and from 2015 to 2016 (c)

(a)

State	Initial probability
1	0.412
2	0.450
3	0.139

(b)

State	1	2	3
1	0.445	0.339	0.216
2	0.013	0.553	0.434
3	0.045	0.228	0.727

(c)

State	1	2	3
1	0.798	0.000	0.202
2	0.000	0.717	0.283
3	0.014	0.085	0.901

5 Conclusions

In this work, a longitudinal analysis of the Italian municipalities compliance as regards anti-corruption measures is presented. The information contained in the annual reports of the supervisor for corruption prevention (RPC forms) are exploited with reference to a set of items about the accomplishment of anti-corruption measures. Data about a sample of more than 200 municipalities are used for the years 2014, 2015 and 2016. A latent Markov model is fitted on the data at issue, finding evidence of three states of increasing virtuosity in terms of anti-corruption behaviour. Results show that in 2014 – year of introduction of the new Italian anti-corruption law – less than 15% of the sample belongs to the group of the most virtuous municipalities, then only few administrations have accomplished measures for contrasting corruptive events. The analysis of the transition matrices reveals high probabilities to move to higher virtuosity states, even though a clear tendency to stay in the current state is observable, especially in the last year.

References

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. CRC Press, Boca Raton, FL (2013)
2. Carloni, E.: Italian anti-corruption and transparency policy. In: Grasse, A., Grimm, M., Labitzke, J. (eds) Italien zwischen Krise und Aufbruch. Auf dem Weg zur dritten Republik?, p.25. Wiesbaden, Springer Fachmedien (in press)
3. Gnaldi, M., Del Sarto, S.: Characterising Italian municipalities according to the annual report of the prevention-of-corruption supervisor: a Latent Class approach, In: Petrucci, A., Verde, R. (eds.) Data Science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society, pp. 513-518. FUP, Firenze (2017)
4. Del Sarto, S., Gnaldi, M.: Considerazioni sull'efficacia delle relazioni del Responsabile della Prevenzione della Corruzione come strumento di prevenzione. In: Gnaldi, M., Ponti, B. (eds.) Misurare la corruzione oggi: obiettivi, metodi, esperienze. Franco Angeli, Roma (in press)
5. Gnaldi, M., Del Sarto, S.: Corruption prevention: first evidences from the 2015 RPC survey for Italian municipalities (in press)
6. Schwarz, G.: Estimating the dimension of a model. Ann Stat, **6**, 461—464 (1978)
7. Wiggins, L.M.: Panel Analysis: Latent Probability Models for Attitude and Behaviour Processes. Elsevier, Amsterdam (1973)

Modelling the effects of covariates for unbiased estimates in ecological inference methods

Analisi degli effetti delle covariate per stime non distorte nei metodi di inferenza ecologica

Venera Tomaselli, Antonio Forcina and Michela Gnaldi

Abstract After showing that the estimates provided by three main ecological inference methods are heavily biased when compared to multilevel logistic models applied to a set of real individual data, the paper argues that ecological bias can be corrected only by accounting for relevant covariates. In addition, a data generating mechanism where bias cannot even be corrected by using covariates is described.

Abstract *Dopo aver dimostrato che le stime ottenute mediante i tre principali metodi di inferenza ecologica sono fortemente distorte rispetto a quelle ottenute applicando modelli logistici multilivello, lo studio conclude che le distorsioni possono essere corrette soltanto tenendo conto di covariate pertinenti. E', inoltre, descritto un sistema di generazione di dati dove le distorsioni non possono essere corrette neppure usando covariate.*

Key words: ecological fallacy, biased estimates, covariates effects, multilevel logistic regression.

1 Ecological Fallacy and ecological inference models

Since Robinson's seminal paper [16], it is well known that the association between two variables estimated from data aggregated within territorial units, like polling stations, may be substantially biased compared to the association that would emerge

Venera Tomaselli (*corresponding author*)
Department of Political and Social Sciences, University of Catania, e-mail: tomavene@unict.it
Antonio Forcina
Department of Economics, University of Perugia, e-mail: forcinarosara@gmail.com
Michela Gnaldi
Department of Political Sciences, University of Perugia, e-mail: michela.gnaldi@unipg.it

if data recorded at the individual level were available. This phenomenon became known as the *ecological fallacy*.

Subramanian et al. [19] pointed out that, in certain contexts, the degree of association at the individual level may depend on modelling assumptions and thus may not be such an objective quantity as Robinson seemed to believe. An important implication of this result is that, when the estimates from ecological and individual level studies do not agree, additional investigation may be necessary before concluding that the ecological estimates are inappropriate.

Since the introduction of ecological regression by Goodman [8], several methods of ecological inference have been developed by King [11], [12] and coworkers [17]; their merits are debated [5], [2], [14]. Though less popular, the methods proposed by Brown and Payne [1] and Greiner and Quinn [9] may be considered as valid alternatives.

In a recent paper [6] the authors, by elaborating on the work of Wakefield [20] and Firebaugh [3], argue that, when certain assumptions are violated, the estimates produced by any method of ecological inference are going to be biased. Though bias might be corrected by modelling the effects of relevant covariates, even this may fail under certain data generating mechanisms.

2 Ecological Estimates by Multilevel Logistic Models

This paper is based on the analysis of an extensive set of individual data on voting behaviour in the Democratic Party Primary election for the candidate Mayor in the city of Palermo, Italy, in 2012. For each polling station, the data provide the joint distribution of voters classified by their decision (vote or not at the primary election) on one side and their age and sex on the other.

The estimates of ecological inference methods rely on the marginal distribution of voting decisions and that of sex, age; when individual data are available, these estimates can be compared with the actual proportions of voters within each age by sex group. In addition, by applying multilevel logistic models one can check if the propensity to vote depends on the relative size of the sex by age groups together with other covariates: when this happens it can be shown that ecological estimates are going to be biased.

The estimates of voting probabilities provided by the Goodman [8] regression model, the King [17] multinomial-Dirichlet and the modified Brown and Payne model [4] without covariates (Table 1) are substantially different from those based on individual data: for certain age groups, estimated probabilities are close to 0 while, for other age groups, they are much higher than the observed proportions.

Next we apply multilevel models [7],[10],[18] to the individual dataset with three objectives: (i) to verify whether the estimates provided by the raw proportions (used by Robinson) and those obtained from multilevel models are substantially different as in [19], (ii) to obtain an estimate of the different variance components and (iii) to detect the appropriate covariates to be used in the ecological inference models. For each polling station, voters are cross-classified by sex (M, F) and 6 age groups and, for each of these 12 categories, according to decision to vote or not. These data may be seen as 12 binomial variables nested within a polling station, with the polling stations grouped into 31 seats. For these data there are three sources of random variation:

(i) binomial within the polling stations;
(ii) among polling stations within seats with an estimated standard deviation of 0.2311;
(iii) among seats with an estimated standard deviation of 0.2547.

To investigate how the propensity to vote depends on available covariates, several different models were explored and the following highly significant covariates were selected:

- *pd*, the proportion of voters for the Democratic Party at the municipal election held a month later (3.8% on average for all the polling stations within total eligible voters);
- *idv*, the proportion of voters for the *Italia dei Valori* Party at the same municipal election (5.1% on average for all the polling stations within total eligible voters);
- *mol*, the proportion of males aged between 45 and 74 (45.0% within male eligible voters);
- *fol*, the proportion of females aged between 45 and 74 (45.4% within female eligible voters).

By the multilevel logistic models fitted separately to each age group, where the observations at the lowest level are the number of voters (classified by sex and age group) nested within the polling stations which, in turn, are nested within the seats, as potentially relevant covariates, we considered the proportions of eligible voters belonging to each age group separately for males and females, in addition to the *pd* and *idv* covariates described above. The parameter estimates of the propensity to vote based on the effects of the relevant covariates are displayed in Table 2.

Though the proportions of voters aged 45-65 and 65-75 were significant most of the times, when *pd* and *idv* were also used, some of the previous covariates appeared to no longer have a significant effect. This could be due to the fact that *pd* and *idv* are closely related to the age distribution within each polling station: when either *pd*

Table 1 Ecological inference estimates of the probability of voting by sex and age groups, without covariates.

	Sex	Age groups					
Method		18-25	25-30	30-45	45-65	65-75	over 75
Goodman	M	0.000	0.000	0.000	0.123	0.147	0.000
	F	0.000	0.000	0.000	0.124	0.000	0.028
Brown-Payne	M	0.000	0.000	0.000	0.000	0.278	0.000
(revised)	F	0.000	0.000	0.000	0.148	0.000	0.152
King OLS	M	0.001	0.001	0.000	0.002	0.264	0.007
	F	0.000	0.001	0.000	0.144	0.004	0.161

Table 2 Estimated parameters of the propensity to vote based on the effects of the relevant covariates by multilevel logistic models.

Parameters	Age groups					
	18-25	25-30	30-45	45-65	65-75	over 75
F	-4.9207$^\bullet$	-4.5096$^\bullet$	-4.4335$^\bullet$	-3.6369$^\bullet$	-4.7301$^\bullet$	-5.0218$^\bullet$
pd	13.1413$^\bullet$	14.8040$^\bullet$	8.7711$^\bullet$	12.4372$^\bullet$	7.9891$^\bullet$	7.1119$^\bullet$
idv	4.5733*	0.0000°	4.2559*	5.2272$^\bullet$	4.0019*	10.3799$^\bullet$
$P(45-65)$	2.3815*	2.6580*	1.6830*	0.0000°	2.0258*	0.0000°
$P(65-75)$	2.3888*	0.0000°	1.9564*	1.3247*	2.9426$^\bullet$	0.0000°
$M-F$	0.0256°	-0.0570°	0.0853$^\bullet$	0.0898$^\bullet$	0.4785$^\bullet$	0.8171$^\bullet$

F is the intercept within females; $M-F$ is the difference in intercept between males and females; $\circ$ = not significant, $\star$ = 5% significant, $*$ = 1% significant, $\bullet$ = p-value smaller than 0.001.

or *idv* increases, the proportion of eligible voters in the 18-45 age group decreases while the proportion in the age range from 45 to 75 and over increases.

The fact that in the multilevel logistic models applied to the individual data the propensity to vote depends significantly on covariates measured at the level of polling station provides an explanation for the bias present in ecological inference estimates.

However, both the King and the Brown-Payne methods allow modelling the effects of covariates on the logit of the propensity to vote. A rather disappointing (but at the same time rather intriguing) feature of the Palermo data is that both the King and the Brown and Payne model continues to provide biased estimates even if we allow the logit of the propensity to vote to depend on the same covariates which were detected as significant in the multilevel logistic models applied to the individual data.

An important result of this paper is that the failure of covariates to correct the ecological bias is not a feature specific to the Palermo data set. To support this claim we describe a plausible data generating mechanism which may have been working in the Palermo Primary election and explain why, under these conditions, modelling the effects of covariates may not correct the bias.

Let q_{as} be the probability of voting at the Primary election for an eligible voter with sex s and affiliation to centre-left parties $a = 0,\,1$. Let also v_{us} denote the proportion of eligible voters of sex s who are affiliated to the same parties in the polling station u. Then, it is easily shown that:

$$\pi_{us1} = q_{0s}(1 - v_{us}) + q_{1s}v_{us}. \tag{1}$$

There are two important features in this equation: (i) it depends on the proportion of voters affiliated to centre-left parties (which cannot be observed), rather than on the proportion of females; (ii) the dependence is linear rather than logistic. Artificial electoral data based on equation (1) were generated at random as follows:

1. for each polling station, splitting eligible voters at random between affiliated and not and then among females and males in such a way that sex and affiliation are correlated;

2. assigning plausible values to the q_{as} probabilities;
3. generating random electoral data as in a revised Brown and Payne model.

Table 3 shows the estimated proportions of females and males voters using individual data and the Brown and Payne and King OLS models. These estimates are computed on an artificial data set generated according to the procedure described above. Because it contains 16,000 polling stations with 1,000 voters each, standard errors of the raw proportions in the individual data are very small (less than 0.0005). As a consequence, any new replication should produce, essentially, the same estimates. Since the differences between estimates from ecological and individual data are substantially large, relative to the very large sample size, they are certainly due to bias and not to random variations.

Table 3 Estimated proportions of voters in the artificial data; Ind=Individual data, BP=Brown and Payne and King=King OLS.

Females			Males		
Ind	BP	King	Ind	BP	King
0.0464	0.1012	0.1015	0.0548	0.0000	0.000

3 Conclusions

The findings in this paper indicate that the only possibility to correct ecological bias is to model the effects of covariates; also Liu [14] noted that the estimates from the King's model improved substantially by including certain covariates. However, while Liu was searching among all possible covariates, the results in this paper show that only covariates strongly correlated with the marginal proportions in the explanatory variables (sex and age in our context) are relevant.

An interesting result here is that, while the extended version of an ecologic inference model with covariates does provide a very accurate fit of the total number of voters in the Primary election in each polling station, the estimated number of the same voters by sex and age groups are not much better than those obtained by the same model without covariates.

When voter's choices depend on covariates measured at the level of polling station, any method of ecological inference that does not account for this is going to provide biased estimates. But even this may fail, as in the Palermo data and in the artificial data set generated at random according to a model which assumes that voting decisions depend on party affiliation rather than sex and age.

References

1. Brown, P. J. & Payne, C. D.: Aggregate data, ecological regression, and voting transitions. J. Am. Statist. Assoc. **81**, 452–460 (1986)
2. Cho, W. K. T.: If the assumption fits: A comment on the King ecological inference solution. Polit. Anal. **7**, 143–163 (1998)
3. Firebaugh, G.: A rule for inferring individual-level relationships from aggregate data. Am. Sociol. Rev. **43**, 557–572 (1978)
4. Forcina, A., Gnaldi, M., & Bracalente, B: A revised Brown and Payne model of voting behaviour applied to the 2009 elections in Italy. Stat. Methods Appl. **21**, 109–119 (2012)
5. Freedman, D.A., Klein, S.P., Ostland, M., & Roberts, M.R.: On solutions to the ecological inference problem. J. Amer. Statist. Assoc. **93**, 1518–1522 (1998)
6. Gnaldi, M., Tomaselli, V., & Forcina, A.: Ecological Fallacy and Covariates: New Insights based on Multilevel Modelling of Individual Data. I. Statist. Rev. **86**, (2018) doi:10.1111/insr.12244
7. Goldstein, H.: Multilevel Statistical Models. 4th ed. John Wiley & Sons, Chichester, UK (2011)
8. Goodman, L. A.: Ecological regressions and behavior of individuals. Am. Sociol. Rev., **18**, 351–367 (1953)
9. Greiner, J. D. & Quinn, K. M.: R×C ecological inference: Bounds, correlations, flexibility and transparency of assumptions. J. Roy. Statist. Soc. Ser. A **172**, 67–81 (2009)
10. Hox, J. J., Moerbeek, M., & van de Schoot, R.: Multilevel Analysis: Techniques and Applications. 2nd ed. Routledge, New York, NJ (2010)
11. King, G.: A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton University Press, Princeton NJ (1997)
12. King, G.: The future of ecological inference research: A comment on Freedman et al.. J. Amer. Statist. Assoc. **94**, 352–355 (1999)
13. King, G., Rosen, O., & Tanner, M. A.: Binomial-Beta hierarchical models for ecological inference. Sociol. Methods Res. **28**, 61–90 (1999)
14. Liu, B.: EI extended model and the fear of ecological fallacy. Sociol. Methods Res. **20**, 1–23 (2007)
15. Ng, K. W., Tian, G. L., & Tang, M. L.: Dirichlet and Related Distributions: Theory, Methods and Applications. John Wiley & Sons, Chichester UK (2011)
16. Robinson, W. S.: Ecological correlations and the behavior of individuals. . Am. Sociol. Rev. **15**, 351–357 (1950)
17. Rosen, O., Jiang, W., King, G., & Tanner, M. A.: Bayesian and frequentist inference for ecological inference: The R×C case. Stat. Neerl. **55**, 134–156 (2001)
18. Snijders, T. A. B. & Bosker, R. J.: Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage, London UK (2012)
19. Subramanian, S. V., Jones, K., Kaddour, A., & Krieger, N.:Revisiting Robinson: the perils of individualistic and ecologic fallacy. Int. J. Epidem. **38**, 342–360 (2009)
20. Wakefield, J.: Ecological inference for 2×2 tables (with discussion).J. Roy. Statist. Soc. Ser. A **167**, 1–42 (2004)

Advances in Time Series

Filtering outliers in time series of electricity prices

Ilaria Lucrezia Amerise

Abstract In this paper, we describe a tool for detecting outliers in electricity prices. The tool primarily consists of a filtering procedure (Whittaker smoother) that removes unpredictable effects and captures long-term smooth variations. So as to identify outliers, we compare observed prices with smoothed ones. If the difference between the two exceeds a predetermined limit, the corresponding prices are considered anomalous and candidates for an appropriate statistical treatment. The new tool is compared with another method based on local regression.

Key words: Whittacker smoother, penalized least squares, price spikes.

1 Introduction

The frequent occurrence of irregular, abrupt, extreme movements or spikes in spot prices is one of the most salient characteristics of electricity markets. These spikes are of very short duration so that only a few traces of their influence is revealed at adjacent time points. It should be recognized that price peaks and/or troughs are natural characteristics of the electricity market and are very important for energy market participants. On the other hand, we do not expect that the choice of forecasting method to be conditioned by any mere irregularity in prices. The main purpose of the present study is to identify outlying spikes by using a non-parametric filter that returns a smoothed time series, which acts as benchmark against which moderate and extreme spike prices can be identified. The same pre-processing can be applied to other phenomena with fast dynamics, such as meteorological time series, stock market indices and exchange rates.

Ilaria Lucrezia Amerise
Università della Calabria, Dipartimento di Economia, Statistica e Finanza - Cosenza, Italy e-mail: ilaria.amerise@unical.it

Let us suppose that a time series of electricity prices consists of a, not-better-specified, underlying long-term pattern and of a non-observable random error

$$p_t = \widehat{p}_t + a_t, \qquad t = 1, 2, \cdots, n \tag{1}$$

where p_t is the price in €/MWh (if necessary, log-transformed) on day t, n is the length of the time series, $\widehat{p}_t$ is the baseline and the errors a_t have zero mean. The baseline $\widehat{p}_t$ aims to capture any predictable variation in electricity price behavior arising from regularities over time and it is used to interpolate or filter the observed prices. The stochastic term a_t can be interpreted as what remains of the price after the baseline has been removed. Deviations between observed and baseline prices falling outside a prefixed range will indicate aberrant prices.

To keep (1) simple and general, the description of the baseline has to be smooth, with no preconceptions about its form. Perhaps the simplest smoothing function is a straight line parallel to the time axis. The smoothness, however, is just one side of the coin. The filtered prices must also be close to the observed prices. For example, a polynomial with the number of parameters equal to the number of data points will give the exact fit. Since the two requirements of adequacy to p_t and roughness of $\widehat{p}_t$ may conflict, the choice of $\widehat{p}$ is inevitably a compromise solution.

In this paper, we discuss the penalized least squares method of smoothing a time series. The basic idea dates back at least as far as [4] and the method has been highly studied since. The paper is organized as follows. In Section 2, we present our basic method and show how a valid time series smoother can be obtained by balancing goodness of fit and lack of smoothness. Section 3 discusses the detection of outliers in hourly time series of electricity prices. The alternative method proposed by [3] is briefly outlined. The case studies in Section 3 serve to exemplify how the two methods can be used as a data preparation tool. The final section discusses our results and points out some further applications and improvements.

2 Whittacker smoother

The scope of this section is to model p_t by applying a baseline time series which best matches the original whilst being as smooth as possible. By assigning the proportion in which an increase in the goodness of fit is to be taken as counterbalancing a decrease in the smoothness, we can determine the series which best harmonizes the two requirements. The Whittacker smoother (WS) requires

$$\min_{\widehat{p}_t} Q_t\left(\widehat{p}_t\right) = L_t\left(\widehat{p}_t\right) + \lambda R_t\left(\widehat{p}_t\right) \tag{2}$$

$$L_t\left(\widehat{p}_t\right) = \sum_{t=1}^{n} \left(\widehat{p}_t - p_t\right)^2, \qquad R_t\left(\widehat{p}_t\right) = \sum_{t=m}^{n} \left[\nabla^m \widehat{p}_t\right]^2, \qquad t = 1, 2, \cdots, n \tag{3}$$

Here m is a positive integer, $\lambda > 0$ is a scalar and $\nabla = (1-B)$ is the difference operator: $\nabla\widehat{p}_t = \widehat{p}_t - \widehat{p}_{t-1}$. Note that $R_t\left(\widehat{p}_t\right)$ is null if $\widehat{p}_t$ is a polynomial of degree m.

The component L_t reflects the lack of fit, which is measured as the usual sum of squares of differences. The component R_t expresses the lack of smoothness in terms of the m-th differences of $\widehat{p}_t$. The positive constant λ controls the solution trade-off between smoothness and fidelity to p_t. If $\lambda \to 0$, then the dominant component of (2) will be the residual sum of squares and $\widehat{p}_t$ will resemble the original data points increasingly closely, no matter how irregular it may be. As the value of λ approaches ∞, the resulting baseline will approach a polynomial of degree m. Thus, an optimal relationship between the two components could and should be provided. It may be easily shown that

$$\nabla^m \widehat{p}_j = \sum_{j=i}^{n} D_{ij}\widehat{p}_j \tag{4}$$

$$D_{ij} = \begin{cases} (-1)^{m+j-i}\binom{m}{j-i} & i=1,2,\cdots,m; j=i,i+1,\cdots,i+m \\ 0 & \text{otherwise} \end{cases} \quad \text{for } 1 \le i \le n-m\,. \tag{5}$$

Thanks to this relationship, model (2) can be expressed in matrix notation

$$\mathbf{Q}(\widehat{\mathbf{p}}) = (\widehat{\mathbf{p}} - \mathbf{p})^t(\widehat{\mathbf{p}} - \mathbf{p}) + \lambda(\mathbf{D}\widehat{\mathbf{p}})^t(\mathbf{D}\widehat{\mathbf{p}})\,. \tag{6}$$

where $\mathbf{D}$ is a rectangular matrix with $(n-m)$ rows and n columns. Deriving $\mathbf{Q}(\widehat{\mathbf{p}})$ with respect to $\widehat{\mathbf{p}}$ and equating the partial derivatives to zero gives the linear system

$$\mathbf{A}\widehat{\mathbf{p}} = \mathbf{p} \quad \text{wih } \mathbf{A} = (\mathbf{I}_n + \lambda\mathbf{B}), \quad \mathbf{B} = \mathbf{D}^t\mathbf{D} \tag{7}$$

Here $\mathbf{I}_n$ is the $(n \times n)$ identity matrix and $\mathbf{A}$ is a symmetrical, positive definite matrix of order $(n \times n)$. It follows that the minimum occurs when $\widehat{\mathbf{p}} = \mathbf{A}^{-1}\mathbf{p}$.

Taking into account the considerable length of time series collected in the electricity market, the solution of (7) appears problematic. However, the difficulties are fewer than at first appear. Several authors, in fact, have devised very efficient computing software by exploiting the characteristics of the matrices involved. See [2]. It should be pointed out that any single value of λ could capture the smooth components of time series such as those encountered in the electricity market. A virtually flat baseline using a large value of λ will generate more false rejections of valid prices and a peak-rich one using a small value of λ will lead to false acceptances. According to [1], an objective value of λ can be obtained by minimizing the generalized cross-validation (GCV) score

$$\min_{\lambda} gcv(\lambda) = \frac{n^{-1}\sum_{t=1}^{n}\left[p_t - \widehat{p}_t(\lambda)\right]^2}{\left[1 - n^{-1}Tr(\mathbf{A}^{-1})\right]^2} \tag{8}$$

where Tr denotes the matrix trace. Criterion (8) is a measure of the leave-one-out mean squared prediction error. The effect of applying gcv to hourly time series is shown in Fig. 1 with different degrees of the polynomial in the smoothness component.

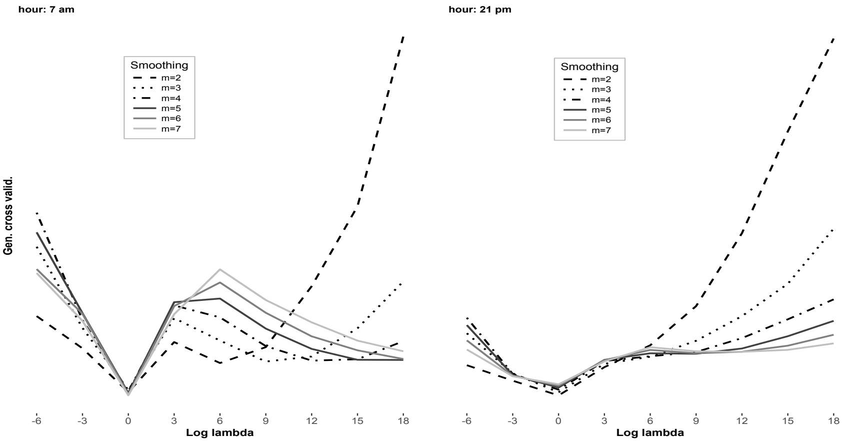


Fig. 1 Zonal electricity price in Northern Italy. Search for the best λ.

In practice, values of gcv are computed for many trial values of λ. Since no analytical solution exists, we need to resort to a numerical method to solve (8). We perform the minimization of (8) by performing a simple grid search for λ in a set of values for $log(\lambda) : 0, 10^j, j = 0, 1, \cdots, 9$. The $\hat{\lambda}$ value corresponding to the minimum of the minima will indicate the apparently most suitable Whittaker smoother for the time series under investigation. In so doing, the procedure for finding a baseline is fully automated. However, the simple fact that a procedure can be executed using a computer does not, of course, make it objective. It is not by coincidence that we said apparently because there is no guarantee that the λ corresponding to the global minimum will produce a valid smoother. Unfortunately, some heuristic calculations show that the minimizer of (8) will give a $\hat{\lambda}$ that leads to substantial under-smoothing and the resulting long-term pattern tends to have too many oscillations that often show up in the wrong places. These cast many doubts on the validity of (8).

As a final consideration, we note that the order of difference m is, to a limited extent, another parameter, other than λ, influencing the result of filtering a time series. However, in practice, the choice is confined to $m = 2$ or $m = 3$ because the curves for $m > 3$ show a similar pattern.

3 Detection of outliers

As stated in the introduction, we detect outliers on the basis of the difference between original and interpolated prices $\widehat{a}_t = p_t - \widehat{p}_t$ $t = 1,2,\cdots,n$. In this regard, it is necessary to define a lower and upper bound for the residuals $\widehat{a}_t$

$$\widehat{a}_t < Q_\beta - K\left(Q_{1-\beta} - Q_\beta\right) \quad \text{or} \quad \widehat{a}_t > Q_{1-\beta} + K\left(Q_{1-\beta} - Q_\beta\right) \quad t = 1,2,\cdots,n \quad (9)$$

where Q_β is the β-quantile and K is a positive multiplicative factor. If a residual surpasses the fences, then the corresponding price is considered an outlier. Obviously, different definitions of K and β may lead to quite different results and identification of price spikes. As an illustration of the potential merit/demerit balance of the proposed method, we compare the results of WS with those obtained by using the *tsoutliers* routine of *forecast* package in *R* software. Put briefly, *tsoutliers* is a recursive filter that decomposes the time series into seasonal, trend and irregular components, which are used to construct a reference time series based on locally weighted polynomial regression (loess). The *tsoutliers* function has the default values $\beta = 0.1$, $K = 2$. We have applied these bounds to the time series of hourly zonal electricity prices in northern Italy for the years 2013-2017. The method indicates 47 suspect spikes. The grid search proposes $\hat{\lambda} = 282'613$ and $m = 3$. To obtain the same number of outliers with the WS, we used $\beta = 0.25$ and $K = 2.714$. However, there are only 26 values in common with $tsoutliers$.

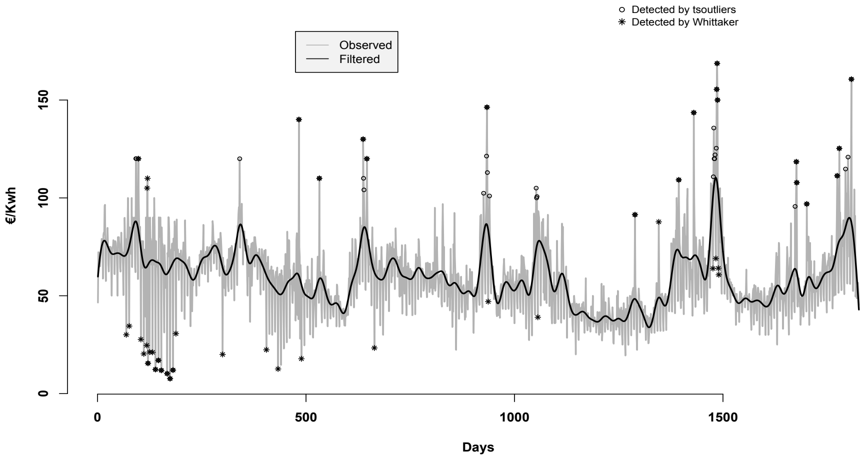


Fig. 2 Outliers detection in hourly prices, h: 10am, Zone: North, period: 2013-2017.

Once an extreme outlier has been detected and a complete time-series is required, the aberrant value may be replaced by a less unusual value. This question, however, is not discussed in the present paper.

4 Results and conclusions

In Table 1, we have collected the results obtained from the examination of $144 = 6 * 24$ time series (one of these has been used in Fig. 2). The data employed are hourly spot prices from the Italian day-ahead zonal market. The time series runs from 1am on Tuesday, 1st January 2013 to 24pm on Sunday, 31st December 2017, yielding data for each hour of the day and for each one of the six zones of the Italian market 1826 days. Table 1 shows the frequency of outliers and the three quartiles of the relative magnitude of outliers: $(p_t - \hat{p}_t)/p_t, p_t > 0$. The percentage of peaks and valleys that were marked as outliers by both methods is about 24‰ (in practice, 18 hourly outliers each month), which is less than a half of that detected by at least one of the filters. The findings in the table reveal that WS tends to be more aggressive than $tsoutliers$ because it not only locates a greater number of outliers than $tsoutliers$, but also because their relative size is smaller.

Table 1 Comparison of the Whittaker filter with the Hyndman filter.

	N	Q_1	Q_2	Q_3		N	Q_1	Q_2	Q_3
$H:$	46‰	0.110	0.318	0.496	$W:$	35‰	0.146	0.240	0.389
$H \cup W:$	56‰	0.112	0.301	0.521	$H \cap W:$	24‰	0.175	0.404	1.148

The two methods agree fairly well with each other in the case of extreme spikes, asis evident from the quartiles of $H \cap W$. Since we have little information about the real position and magnitude of the possible outliers, it is unclear which filter is more appropriate for our type of data.

In conclusion we can say that, although it is not possible to identify a single best method for outlier detection in day-ahead Italian zonal hourly electricity prices, the two procedures that we have applied provide sufficiently distinct results to allow them to be used in combination. Therefore, it seems logical to suggest that aberrant prices should be searched for among those that lie outside the fences of both methods.

References

1. Craven, P., Wahba, G.: Smoothing noisy data with spline functions estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math. **31**, 377–403 (1979)
2. Garcia, D.: Robust smoothing of gridded data in one and higher dimensions with missing values. Comput. Stat. Data Anal. **54**, 1167–1178 (2010)
3. Hyndman, R.J. and Khandakar, Y . Automatic time series forecasting: the forecast package for R. J. Stat. Softw. **26**,1–22 (2008)
4. Whittaker, E.T.: On a new method of graduation. Proc. R. Soc. Edinburgh. **44**, 77–83 (1923)

Time-varying long-memory processes

Processi a memoria lunga con parametro frazionario variabile nel tempo

Luisa Bisaglia and Matteo Grigoletto

Abstract In this work we propose a new class of long-memory models with time-varying fractional parameter. In particular, the dynamics of the long-memory coefficient, d, is specified through a stochastic recurrence equation driven by the score of the predictive likelihood, as suggested by Creal *et al.* (2013) and Harvey (2013).

Key words: long-memory, GAS model, time-varying parameter

1 Introduction

Long-memory processes have proved to be useful tools in the analysis of many empirical time series. These series present the property that the autocorrelation function at large lags decreases to zero like a power function rather than exponentially, so that the correlations are not summable.

One of the most popular processes that takes into account this particular behavior of the autocorrelation function is the AutoRegressive Fractionally Integrated Moving Average process (ARFIMA(p,d,q)), independently introduced by Granger and Joyeux (1980) and Hosking (1981). This process generalizes the ARIMA(p,d,q) process by relaxing the assumption that d is an integer.

The ARFIMA(p,d,q) process, Y_t, is defined by the difference equation

$$\Phi(B)(1-B)^d(Y_t-\mu)=\Theta(B)\varepsilon_t,$$

where $\varepsilon_t \sim WN(0,\sigma^2)$, and $\Phi(\cdot)$ and $\Theta(\cdot)$ are polynomials in the backward shift operator B of degrees p and q, respectively. Furthermore, $(1-B)^d = \sum_{j=0}^{\infty}\pi_j B^j$,

Luisa Bisaglia
Dept. of Statistical Science, University of Padova, e-mail: luisa.bisaglia@unipd.it

Matteo Grigoletto
Dept. of Statistical Science, University of Padova, e-mail: matteo.grigoletto@unipd.it

with $\pi_j = \Gamma(j-d)/[\Gamma(j+1)\Gamma(-d)]$, where $\Gamma(\cdot)$ denotes the gamma function. When the roots of $\Phi(B)=0$ and $\Theta(B)=0$ lie outside the unit circle and $\mid d \mid < 0.5$, the process is stationary, causal and invertible. We will assume these conditions to be satisfied.

When $d \in (0, 0.5)$ the autocorrelation function of the process decays to zero hyperbolically at a rate $O(k^{2d-1})$, where k denotes the lag. In this case we say that the process has a long-memory behavior. When $d \in (-0.5, 0)$ the process is said to have intermediate memory.

If $p = q = 0$, the process $\{Y_t,\ t = 0, \pm 1, \ldots\}$ is called Fractionally Integrated Noise, $FI(d)$. In the following we will concentrate on $FI(d)$ processes with $d \in (-0.5, 0.5)$.

Several papers have addressed the detection of breaks in the order of fractional integration. Some of these works allowed for just one unknown breakpoint (see, for instance, Berand and Terrin, 1996; Yamaghuchi, 2011). Others treated the number of breaks as well as their timing as unknown (Ray and Tsay, 2002; Hassler and Meller, 2014). Boutahar *et al.* (2008) generalize the standard long memory modeling by assuming that the long memory parameter d is stochastic and time-varying. The authors introduce a STAR process, characterized by a logistic function, on this parameter and propose an estimation method for the model. Finally, Roueff and von Sachs (2011) take into account the time-varying feature of long-memory parameter d using the wavelets approach.

Our approach is completely different because we allow the long memory parameter d to vary at each time t. Moreover, our approach is based on the theory of Generalized Autoregressive Score (GAS) models. In particular, the peculiarity of our approach is that the dynamics of the long-memory parameter is specified through a stochastic recurrence equation driven by the score of the predictive likelihood. In this way we are able to take into account also smooth changes of the long-memory parameter.

2 GAS model

To allow for time-varying parameters, Creal *et al.* (2013) and Harvey (2013) proposed an updating equation where the innovation is given by the score of the conditional distribution of the observations (GAS models). The basic framework is the following. Consider a time series $\{y_1, \cdots, y_n\}$ with time-t observation density $p(y_t \mid \psi_t)$, where $\psi_t = (f_t, \theta)$ is the parameter vector, with f_t representing the time-varying parameter(s) and θ the remaining fixed coefficients.

In time series the likelihood function can be written via prediction errors as:

$$\mathscr{L}(y, \psi) = p(y_1; \psi_1) \prod_{t=2}^{n} p(y_t \mid y_1, \cdots, y_{t-1}; \psi_1, \cdots, \psi_t) \; .$$

Thus, the t-th contribution to the log-likelihood is:

$$l_t = \log p(y_t \mid y_1, \cdots, y_{t-1}; f_1, \cdots, f_t; \theta) = \log p(y_t \mid y_1, \cdots, y_{t-1}; f_t; \theta) \ ,$$

where we assume that $f_1, \cdots, f_t$ are known (because they are realized).

The parameter value for the next period, f_{t+1}, is determined by an autoregressive updating function that has an innovation equal to the score of l_t with respect to f_t. In particular, we can assume that:

$$f_{t+1} = \omega + \beta f_t + \alpha s_t \ ,$$

where the innovation s_t is given by

$$s_t = S_t \cdot \nabla_t \ ,$$

with

$$\nabla_t = \frac{\partial \log p(y_t \mid y_1, \cdots, y_{t-1}; f_t, \theta)}{\partial f_t} \tag{1}$$

and

$$S_t = \mathscr{I}_{t-1}^{-1} = -E_{t-1}\left[\frac{\partial^2 \log p(y_t \mid y_1, \cdots, y_{t-1}; f_t, \theta)}{\partial f_t \partial f_t'}\right]^{-1} \ . \tag{2}$$

By determining f_{t+1} in this way, we obtain a recursive algorithm for the estimation of time-varying parameters.

3 TV-FI(d) model

In this section, we extend the class of $FI(d)$ models, by allowing the long-memory parameter d to change over time. The dynamics of the time-varying coefficient d_t is specified in the GAS framework outlined above.

The $TV-FI(d)$ model is described by the following equations:

$$(1-B)^{d_t}\, y_t = \varepsilon_t \ ,$$

$$d_{t+1} = \omega + \beta d_t + \alpha s_t \ , \tag{3}$$

where $\varepsilon_t \sim iid\mathscr{N}(0, \sigma^2)$, and $s_t = S_t \nabla_t$ with S_t and ∇_t defined below.

To calculate the score of the log-likelihood it is preferable to consider the use of autoregressive representation (see, for instance, Palma, 2007):

$$(1-B)^{d_t} y_t = y_t + \sum_{j=1}^{\infty} \pi_j(d_t) y_{t-j} = \varepsilon_t \ ,$$

where

$$\pi_j(d_t) = \prod_{k=1}^{j} \frac{k-1-d_t}{k} = -\frac{d_t \Gamma(j-d_t)}{\Gamma(1-d_t)\Gamma(j+1)} = \frac{\Gamma(j-d_t)}{\Gamma(-d_t)\Gamma(j+1)} \ .$$

In practice, only a finite number n of observations is available. Therefore, we use the approximation

$$y_t = -\pi_1(d_t)y_{t-1} - \pi_2(d_t)y_{t-2} - \cdots - \pi_m(d_t)y_{t-m} + \varepsilon_t \ ,$$

with $m < n$. Then, the t-th contribution, $t = 1, \ldots, n$, to the log-likelihood is:

$$l_t(d_t, \sigma^2) = c - \log(\sigma^2) - \frac{1}{\sigma^2}\left(y_t + \sum_{j=1}^{t-1} \pi_j(d_t) y_{t-j}\right)^2$$

where c is a constant and the corresponding score of the predictive likelihood, see equation (1), becomes

$$\nabla_t = -\frac{1}{\sigma^2}\left(y_t + \sum_{j=1}^{t-1} \pi_j(d_t)y_{t-j}\right)\left(\sum_{j=1}^{t-1} \nu_j(d_t)y_{t-j}\right) \ , \tag{4}$$

where

$$\nu_j(d_t) = \frac{\partial \pi_j(d_t)}{\partial d_t} = \pi_j(d_t)\left(-\Psi(j-d_t) + \Psi(1-d_t) + \frac{1}{d_t}\right) \ , \tag{5}$$

with $\Psi(\cdot)$ representing the digamma function. Finally, we find that S_t in equation (2) is

$$S_t = \sigma^2 \cdot \left(\sum_{j=1}^{t-1} \nu_j(d_t)y_{t-j}\right)^{-2} \ .$$

4 Some Monte Carlo results

We simulated $y_1, \ldots, y_n$ from a process

$$(1-B)^{d_t} y_t = \varepsilon_t \ , \tag{6}$$

where $\varepsilon_t \sim iid\mathcal{N}(0, \sigma^2)$, and d_t is defined by

$$d_t = 0.1 + 0.3 \ \frac{t}{n} \tag{7}$$

or

$$d_t = 0.1 + 0.3 \ \Phi\left(\frac{t-n/2}{3\sqrt{n}}\right) \ , \tag{8}$$

with $\Phi(\cdot)$ indicating the standard Gaussian distribution function.

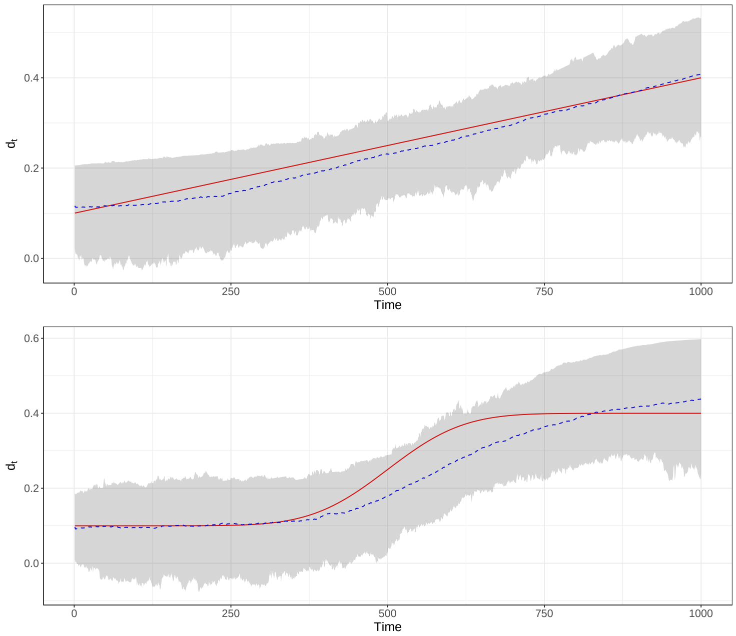


Fig. 1 Result of 200 Monte Carlo simulations, where a time variable fractional parameter (solid line) is estimated with a TV-I(d) model. The dashed line represent the average estimates, while the gray band shows the empirical 95% intervals.

The evolution of d_t was then estimated using the TV-FI(d) model introduced above. It should be noted that in GAS models the scaling defined by (2) is often replaced by S_t^γ, for some suitable γ. We found results (Creal *et al.*, 2013) to be more stable with $\gamma = 0.5$. Also, GAS models can easily be accommodated in order to include a link function $\Lambda(\cdot)$, typically with the objective to constrain the parameter of interest to vary in some region. We used

$$d_t = \Lambda(g_t) = a + (b-a)\,\frac{e^{g_t}}{1+e^{g_t}}\ ,$$

so that $d_t \in (a,b)$, while $g_t \in \mathbb{R}$. Recursion (3) is then defined in terms of g_t, with (4) and (5) easily adjusted for the reparametrization.

It should be remarked that d_0, the value of the fractional at time 0, is necessary to define the likelihood. In the following, we treat d_0 as a parameter to be estimated along with the others.

We obtained 200 Monte Carlo replications from the process defined by (6), and (7) or (8), setting $n = 1000$ and $\sigma = 2$.

For each replication, the TV-FI(d) model was estimated by maximum likelihood, setting $(a,b) = (-0.4, 0.6)$ and $\omega = 0$, while estimating $(d_0, \alpha, \beta, \sigma)$.

Simulation results are shown in Figure 1. The solid line shows the true evolution of d_t, while the dashed line is its estimate, averaged over the Monte Carlo replications. The gray band represents the empirical 95% intervals.

References

1. Beran J. and Terrin N.: Testing for a change of the long-memory parameter. Biometrika, **83**, 627–638 (1996).
2. Boutahar M., Dufrénot G. and Péguin-Feissolle A.: A Simple Fractionally Integrated Model with a Time-varying Long Memory Parameter d_t. Computational Economics, **31**, 225–241 (2008).
3. Creal, Drew D., Koopman S.I. and Lucas A.: Generalized Autoregressive Score Models with Applications. Journal of Applied Econometrics, **28**, 777–795 (2013).
4. Granger, C.W.J. and Joyeux, R.: An introduction to long-range time series models and fractional differencing. Journal of Time Series Analysis, **1**, 15–30 (1980).
5. Harvey, A.C.: Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series. Econometric Series Monographs. Cambridge University Press (2013).
6. Hassler, U. and Meller, B.: Detecting multiple breaks in long memory the case of U.S. inflation. Empirical Economics, **46**, 653–680 (2014).
7. Hosking, J.R.M.: Fractional differencing. Biometrika, **68**, 165–176 (1981).
8. Palma, W.: Long-memory Time Series. Wiley, New Jersey (2007).
9. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (2015). URL https://www.R-project.org/.
10. Ray, B.K. and Tsay R.S.: Bayesian methods for change-point detection in long-range dependent processes. Journal of Time Series Analysis, **23**, 687–705 (2002).
11. Roueff, F. and von Sachs, R.: Locally stationary long memory estimation. Stochastic Processes and their Applications, **121**, 813–844 (2011).
12. Yamaguchi, K.: Estimating a change point in the long memory parameter. Journal of Time Series Analysis, **32**, 304–314 (2011).

Chapter 1
Statistical Analysis of Markov Switching DSGE Models

Maddalena Cavicchioli

Abstract We investigate statistical properties of Markov switching Dynamic Stochastic General Equilibrium (MS DSGE) models: L^2 structure, stationarity, autocovariance function, and spectral density.

Abstract In questo lavoro si studiano le proprietà statistiche dei modelli Markov switching Dynamic Stochastic General Equilibrium (MS DSGE): L^2 struttura, stazionarietà e le funzioni di autocovarianza e di densità spettrale.

Key words: Multivariate DSGE, State-Space models, Markov chains, changes in regime, autocovariance structure, spectral density function

1.1 Introduction

Dynamic Stochastic General Equilibrium (DSGE) models have recently gained a central role for analyzing the mechanisms of propagation of economic shocks. See, for example, Fernández–Villaverde *et al.* (2007) and Giacomini (2013). However, empirical work has shown that DSGE models have failed to fit the data very well over a long period of time. In fact, the changes of parameters require the economists to re–estimate them. This observation leads to Markov switching (MS) DSGE models, that is, DSGE models in which the coefficient parameters are assumed to depend on the state of an unobserved Markov chain.

Since the influential work of Hamilton (1989, 1990), Markov switching models are widely used to capture the business cycle regime shifts typically observed in

Maddalena Cavicchioli
University of Verona, Dipartimento di Scienze Economiche, Via Cantarane 24, 37129 Verona (Italy), e-mail: maddalena.cavicchioli@univr.it

economic data. See also Hamilton (1994, §22), and Hamilton (2005) for more recent references. Markov switching VARMA models have been studied by many authors. For information concerning the stationarity, estimation, consistency, asymptotic variance and model selection of MS VARMA models see Hamilton (cited above), Krolzig (1997), Francq and Zakoïan (2001), Yang (2000), Zhang and Stine (2001), and Cavicchioli (2014a). See also Cavicchioli (2014b), where explicit matrix expressions for the maximum likelihood estimator of the parameters in MS VAR(CH) models have been derived. Higher-order moments and asymptotic Fisher information matrix of MS VARMA models are provided in Cavicchioli (2017a) and (2017b), respectively.

The first purpose of the present paper is to investigate the L^2–structure of the MS DSGE models. We derive stationarity conditions, compute explicitly the autocovariance function for such models, and give stable VARMA representation of them. A second goal of the paper is to propose a tractable method to derive the spectral density in a matrix closed-form of MS DSGE models. Then we illustrate some statistical properties of their spectral representation.

1.2 Stationarity of MS DSGE models

Let us consider the following Markov switching DSGE (in short, MS DSGE) model

$$\mathbf{x}_t = \mathbf{A}_{s_t}\mathbf{x}_{t-1} + \mathbf{B}_{s_t}\mathbf{w}_t \tag{1.1}$$

$$\mathbf{y}_t = \mathbf{C}_{s_t}\mathbf{x}_{t-1} + \mathbf{D}_{s_t}\mathbf{w}_t \tag{1.2}$$

where $\mathbf{x}_t$ is an $n \times 1$ vector of possibly unobserved state variables, $\mathbf{y}_t$ is the $k \times 1$ vector of observable variables, and $\mathbf{w}_t \sim \mathrm{NID}(0, \mathbf{I}_m)$ is an $m \times 1$ vector of economic shocks. The matrices $\mathbf{A}_{s_t} \in \mathbb{R}^{n\times n}$, $\mathbf{B}_{s_t} \in \mathbb{R}^{n\times m}$, $\mathbf{C}_{s_t} \in \mathbb{R}^{k\times n}$ and $\mathbf{D}_{s_t} \in \mathbb{R}^{k\times m}$ are real random matrices.

The process (s_t) is a homogeneous stationary irreducible and aperiodic Markov chain with finite state-space $\Xi = \{1, \ldots, d\}$. Let $\pi(i) = \Pr(s_t = i)$ denote the ergodic probabilities, which are positive. Let $p(i, j) = \Pr(s_t = j | s_{t-1} = i)$ be the stationary transition probabilities. Then the $d \times d$ matrix $\mathbf{P} = (p(i, j))$ is called the *transition probability matrix*. The process (s_t) is independent of $(\mathbf{w}_t)$.

In order to investigate the stationarity properties of the process $(\mathbf{y}_t)$, we use the following vectorial representation of the MS DSGE model:

$$\mathbf{z}_t = \boldsymbol{\Phi}_{s_t}\mathbf{z}_{t-1} + \boldsymbol{\Psi}_{s_t}\mathbf{w}_t \tag{1.3}$$

where

$$\mathbf{z}_t = \begin{pmatrix}\mathbf{x}_t\\ \mathbf{y}_t\end{pmatrix} \qquad \boldsymbol{\Phi}_{s_t} = \begin{pmatrix}\mathbf{A}_{s_t} & 0\\ \mathbf{C}_{s_t} & 0\end{pmatrix} \qquad \boldsymbol{\Psi}_{s_t} = \begin{pmatrix}\mathbf{B}_{s_t}\\ \mathbf{D}_{s_t}\end{pmatrix}.$$

Then $\mathbf{z}_t$ is $p \times 1$, Φ_{s_t} is $p \times p$ and Ψ_{s_t} is $p \times m$, where $p = n + k$. Let Φ_i and Ψ_i be the matrices obtained by replacing s_t by i in Φ_{s_t} and Ψ_{s_t}, for $i = 1, \dots, d$. Let us consider the following matrices:

$$\mathbb{P}_{\Phi\otimes\Phi} = \begin{pmatrix} p(1,1)\{\Phi_1 \otimes \Phi_1\} & p(2,1)\{\Phi_1 \otimes \Phi_1\} & \cdots & p(d,1)\{\Phi_1 \otimes \Phi_1\} \\ p(1,2)\{\Phi_2 \otimes \Phi_2\} & p(2,2)\{\Phi_2 \otimes \Phi_2\} & \cdots & p(d,2)\{\Phi_2 \otimes \Phi_2\} \\ \vdots & \vdots & & \vdots \\ p(1,d)\{\Phi_d \otimes \Phi_d\} & p(2,d)\{\Phi_d \otimes \Phi_d\} & \cdots & p(d,d)\{\Phi_d \otimes \Phi_d\} \end{pmatrix}$$

and

$$\Pi_{\Psi\otimes\Psi} := \begin{pmatrix} \pi(1)\{\Psi_1 \otimes \Psi_1\} \\ \pi(2)\{\Psi_2 \otimes \Psi_2\} \\ \vdots \\ \pi(d)\{\Psi_d \otimes \Psi_d\} \end{pmatrix}.$$

Then $\mathbb{P}_{\Phi\otimes\Phi}$ is $(d\,p^2) \times (d\,p^2)$ and $\Pi_{\Psi\otimes\Psi}$ is $(d\,p^2) \times m^2$. Consider the top Lyapunov exponent

$$\gamma_\Phi = \inf_{t\in\mathbb{N}}\{E\,\frac{1}{t}\,\log_e ||\,\Phi_{s_t}\,\Phi_{s_{t-1}}\cdots\Phi_{s_1}\,||\,\}. \tag{1.4}$$

Theorem 1. *If* $\gamma_\Phi < 0$*, then the process* $(\mathbf{y}_t)$ *is the unique strictly stationary solution of the* MS DSGE *model in* (1.1) *and* (1.2).

Theorem 2. *If* $\rho(\mathbb{P}_{\Phi\otimes\Phi}) < 1$*, where* $\rho(\cdot)$ *denotes the spectral radius, then the process* $(\mathbf{y}_t)$ *is the unique nonanticipative second–order stationary solution of the* MS DSGE *model in* (1.1) *and* (1.2).

1.3 Autocovariance structure and spectral analysis

Let $(\mathbf{z}_t)$ (and $(\mathbf{y}_t)$) be second-order stationary. For every h, define

$$W(h) = \Pi_{E(\mathbf{z}_t\mathbf{z}'_{t-h})} := \begin{pmatrix} \pi(1)\,E(\mathbf{z}_t\,\mathbf{z}'_{t-h}|\,s_t = 1) \\ \pi(2)\,E(\mathbf{z}_t\,\mathbf{z}'_{t-h}|\,s_t = 2) \\ \vdots \\ \pi(d)\,E(\mathbf{z}_t\,\mathbf{z}'_{t-h}|\,s_t = d) \end{pmatrix} \in \mathbb{R}^{(d\,p)\times p}.$$

For $h = 0$, we prove that

$$\operatorname{vec} W(0) = \mathbb{P}_{\Phi\otimes\Phi}\,\operatorname{vec} W(0) + \Pi_{\Psi\otimes\Psi}\,\operatorname{vec}(\mathbf{I}_m) \tag{1.5}$$

where $\mathbf{I}_m$ is the identity matrix of order m. Let $\mathbb{P}_\Phi$ be the $(d\,p) \times (d\,p)$ matrix obtained by replacing $\Phi_i \otimes \Phi_i$ by Φ_i, for $i = 1, \dots, d$, in the definition of $\mathbb{P}_{\Phi\otimes\Phi}$. For $h \geq 0$, we prove that

$$W(h) = \mathbb{P}^h_\Phi\,W(0). \tag{1.6}$$

Theorem 3. *Suppose that* $\rho(\mathbb{P}_{\Phi\otimes\Phi}) < 1$. *Then the autocovariance function of the process* $(\mathbf{y}_t)$ *defined by the* MS DSGE *model in* (1.1) *and* (1.2) *is given by*

$$\Gamma_{\mathbf{y}}(h) = \mathbf{f}(\mathbf{e}' \otimes \mathbf{I}_p)\mathbb{P}^h_{\Phi} W(0)\,\mathbf{f}'$$

for every $h \geq 0$, *where*

$$\operatorname{vec} W(0) = \left(\mathbf{I}_{dp^2} - \mathbb{P}_{\Phi\otimes\Phi}\right)^{-1} \Pi_{\Psi\otimes\Psi} \operatorname{vec}\mathbf{I}_m$$

$$\mathbf{f} = (0_{k\times n} \qquad \mathbf{I}_k) \in \mathbb{R}^{k\times p} \qquad\qquad \mathbf{e} = (1\ \cdots\ 1)' \in \mathbb{R}^d.$$

Theorem 3 allows to obtain a stable VARMA representation for any second-order stationary process $(\mathbf{y}_t)$ driven by an MS DSGE model.

Theorem 4. *Suppose that* $\rho(\mathbb{P}_{\Phi\otimes\Phi}) < 1$. *The spectral density matrix of the process* $(\mathbf{y}_t)$ *driven by the* MS DSGE *in* (1.1) *and* (1.2) *is given by*

$$\mathbf{F}_{\mathbf{y}}(\omega) = \mathbf{f}(\mathbf{e}' \otimes \mathbf{I}_p)[-\mathbf{I}_{dp} + 2Re\,Y(\omega)]W(0)\mathbf{f}'$$

where

$$Y(\omega) = (\mathbf{I}_{dp} - \mathbb{P}_{\Phi} e^{-i\omega})^{-1}$$

is a $(dp)\times(dp)$ *complex matrix and* $ReY(\omega)$ *denotes the real part of* $Y(\omega)$.

Theorems 1–4 are the main results proved in Cavicchioli (2018). Examples and numerical applications complete the mentioned paper.

1.4 Reference

[1] Cavicchioli, M. (2014a) *Determining the number of regimes in Markov–switching* VAR *and* VMA *models*, Journal of Time Series Analysis **35** (2), 173–186.

[2] Cavicchioli, M. (2014b) *Analysis of the likelihood function for Markov–switching* VAR(CH) *models*, Journal of Time Series Analysis **35** (6), 624–639.

[3] Cavicchioli, M. (2017a) *Higher order moments of Markov switching* VARMA *models*, Econometric Theory **33** (6), 1502–1515.

[4] Cavicchioli, M. (2017b) *Asymptotic Fisher information matrix of Markov switching* VARMA *models*, Journal of Multivariate Analysis **157**, 124–135.

[5] Cavicchioli, M. (2018) *Spectral Representation and Autocovariance Structure of Markov switching DSGE models*, mimeo.

[6] Fernández-Villaverde, J., Rubio-Ramirez, J., Sargent, T., and Watson, M. (2007) ABCs (*and* Ds) *of understanding* VARs, American Economic Review **97**, 1021–1026.

[7] Francq, C., and Zakoïan, J.M. (2001) *Stationarity of Multivariate Markov-Switching* ARMA *Models*, Journal of Econometrics **102**, 339–364.

[8] Giacomini, R. (2013) *The relationship between* DSGE *and* VAR *models*, in: VAR Models in Macroeconomics-New Developments and Applications: Essays in Honor of C.A. Sims (Fomby, T.B., , Lutz, K., and Murphy, A., eds.), Advances in Econometrics Vol. 32, Emerald Group Publ. Limited, pp. 1–25.

[9] Hamilton, J.D. (1989) *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica **57** (2), 357–384.

[10] Hamilton, J.D. (1990) *Analysis of time series subject to changes in regime*, Journal of Econometrics **45**, 39–70.

[11 Hamilton, J.D. (1994) *Time Series Analysis*, Princeton Univ. Press, Princeton, N.J.

[12] Hamilton, J.D. (2005) *Regime Switching Models*, in New Palgrave Dictionary of Economics. BROOKS, Chapter 9.

[13] Krolzig, H.M. (1997) *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference and Application to Business Cycle Analysis*, Springer Verlag, Berlin-Heidelberg-New York.

[14] Yang, M. (2000) *Some properties of vector autoregressive processes with Markov–switching coefficients*, Econometric Theory **16**, 23–43.

[15] Zhang, J., and Stine, R.A. (2001) *Autocovariance structure of Markov regime switching models and model selection*, Journal of Time Series Analysis **22** (1), 107–124.

Forecasting energy price volatilities and comovements with fractionally integrated MGARCH models

Previsione della volatilità e co-movimenti dei prezzi dei prodotti energetici mediante modelli MGARCH frazionalmente integrati

Malvina Marchese and Francesca Di Iorio

Abstract We investigate the use of fractionally integrated MGARCH models from a forecasting and a risk management perspective for energy prices. Our in-sample results show significant evidence of long memory decay in energy price returns volatilities, of leverage effects and of time-varying autocorrelations. The forecasting performance of the models is assessed by the SPA test, the Model Confidence Set and the Value at Risk.

Abstract *I modelli MGARCH frazionalmente integrati vengono impiegti in questo lavoro per lo previsione e la gestione del rischio nel mercato dei prodotti energetici. I risultati ottenuti mostrano la presenza di long-memory nella volatilità dei rendimenti dei prezzi, effetti di leverage e autocorrelatione variante nel tempo. La capacit previsiva di questi modelli è stata verificata mediante il test SPA, il Model Confidence Set e il Value at Risk.*

Key words: Multivariate GARCH models, Long memory, SPA test, MCS, VaR

1 Introduction

The effects of oil price shocks on macroeconomic variables, the impact on the financial sector of energy prices fluctuations, the degree of integration between different energy markets, have been the main fields of a increasing studies. It seems to be a growing consensus in the literature on the use of multivariate GARCH (MGARCH) models. A main reference can be Wang and Wu (2012) that compare the forecasting performances of several univariate and bivariate GARCH-type models for spot

Malvina Marchese
Cass Business School, 106 Bunhill Row London EC1Y 8TZ, e-mail: malvina.marchese@city.ac.uk

Francesca Di Iorio
Università di Napoli Federico II, via L. Rodinò 22 80138 Napoli, Italia e-mail: fdiiorio@unina.it

price returns of crude oil (WTI), conventional gasoline (NYH), heating oil (NYH) and jet fuel. Using several univariate loss functions, they find that the full and diagonal BEKK cannot be outperformed by any other model according to the superior predictive ability (SPA) test of Hansen (2005). Growing attention has been devoted to the problem of an accurate modelling and forecasting of energy price volatilities and correlations for portfolio construction, hedging purposes and risk management. The multivariate model used to these ends should capture all the statistical regularities of the volatility series and allow for time dependent correlations. While there is significant empirical evidence that univariate energy price volatilities display a strong degree of persistence, consistent with a long memory structure (e.g. Chang et al., 2010), it would seem that no attempt to include it in multivariate models has yet been made. More, choosing the most appropriate MGARCH specification must also entail comparison of the models' forecasting abilities and usefulness in a decision based framework. Most forecasting comparisons in the energy literature are based on univariate loss functions and univariate tests such as the Diebold-Mariano, that do not allow for joint evaluation of volatilities and correlations forecasts accuracy. Indeed a comprehensive conditional variance matrix forecasts comparison based on matrix loss functions seems to be lacking so far. This paper investigates and analyze the comovements across three major energy markets, namely crude oil (West Texas Intermediate-Cushing Oklaoma), conventional gasoline (New York Harbor) and heating oil (New York Harbor), by means of several multivariate GARCH-type models with particular attention on the fractionally integrated dynamic conditional correlation (FI-DCC) model. This multivariate GARCH models with long memory, asymmetries and dynamic correlations significantly improves the models' in sample and forecasting performance, and then the attractiveness in terms of risk monitoring of this class of models.

The aims of this papers are: (i) compare the in sample performances of several alternative multivariate GARCH models for the returns on the spot prices of the three series using standard information criteria; (ii) evaluate their forecasting accuracy using the Superior Predictive Ability (SPA) test of Hansen (2005) and the Model Confidence Set (MCS) method of Hansen, Lunde and Nason (2011); (iii) explore the efficiency gains in using the fractionally integrated DCC model for one step ahead Value at Risk prediction for short and long positions.

2 Data and models estimations

The series under investigation are the spot energy price returns of the crude oil (CO), conventional gasoline (CG) and heating oil (HO). We obtain 6401 valid daily observations from June 1st 1992 till June 12th 2017 from the Energy Information Administration (EIA) in the U.S. Department of Energy. Each series show the expected stylized facts, such as non normality and fat tails, and the Ljung and Box Q shows that the null hypothesis of no autocorrelation up to the 10th lag is rejected at 10% level of significance. The presence of long memory for the returns $\boldsymbol{r}_t$ is assessed by the significant estimation of the long-memory parameter d using the local Whittle estimator. To account for the serial correlation found in the data, we fit a VAR model to the returns vector whit 1 Lag as suggested by selection criteria. Only conventional gasoline displays time dependence in the mean equation and there are no evidences of spillover effects between the series. Post-estimation diagnostic tests for the residuals of the estimated VAR(1) model confirm the presence of strong GARCH effects, non Normality and no serial correlation up to lag 20. Based on the above evidences we fit the several multivariate GARCH specifications to the VAR(1) residuals. Denoting by $\boldsymbol{r}_t$ the vector of log- returns of n oil prices and θ a finite vector of parameters, the general form of a multivariate GARCH (MGARCH) model is: $\boldsymbol{r}_t = \boldsymbol{\mu}_t(\theta) + \boldsymbol{\varepsilon}_t$, with $\boldsymbol{\varepsilon}_t = \boldsymbol{H}_t^{1/2}(\theta)\boldsymbol{z}_t$ where $\boldsymbol{z}_t$ is an *i.i.d* zero mean random vector such that $Var(\boldsymbol{z}_t) = \boldsymbol{I}_n$, and $\boldsymbol{H}_t^{1/2}$ is a $n \times n$ positive definite matrix. The conditional mean of the process is $\boldsymbol{\mu}_t(\theta)$, the matrix $\boldsymbol{H}_t(\theta) = \boldsymbol{H}_t^{1/2}\boldsymbol{I}_n\left(\boldsymbol{H}_t^{1/2}\right)'$ is the conditional variance. The MGARCH specifications are based on different parameterizations of $\boldsymbol{H}_t$ which have been proposed to capture the dynamics of volatilities and correlations, avoiding the *curse of dimensionality* and to ensure positive definiteness of the covariance matrix. Comprehensive reviews of multivariate GARCH models can be found in Bauwens *et al* (2006) and Silvennoinen and Terasvirta (2009). The MGARCH models estimated in the paper and their main characteristics are summarized in Table 1. In this paper the model estimation is performed by Maximum Likelihood methods in one step under the assumption of joint normality of the vector of disturbances using Kevin Sheppard's MFE Toolbox for *Matlab*, release 2016a. Standardized residuals of estimated volatility models are fat tailed, so the assumption of Gaussianity of the innovations is not innocuous and reduces efficiency; however Gaussian likelihood retains consistency under misspecification of the conditional density, as long as the conditional mean and the conditional variance are correctly specified. We estimate $\hat{\theta}$ by maximizing the conditional log likelihood: $L_T(\theta) = c - \frac{1}{2}\sum_{t=1}^{T}\ln|\boldsymbol{H}_T| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{r}_t'\boldsymbol{H}_t^{-1}\boldsymbol{r}_t$.

Table 1 MGARCH models and their characteristics

Models	Dynamic Corr.	Asymmetries	L-Memory	Spillovers
DBEKK	x			
BEKK	x			x
ABEKK	x	x		x
AGARCH		x		x
CCC				
DCC	x			
cDCC	x			
FI-DCC	x		x	
FI-EDCC		x	x	

Full results concerning the model estimations, obtained with one-step Maximum Likelihood, are available from the authors on request. Table 2 reports the maximized log-likelihood and information criteria for all the fitted models. 'Np' is the number of estimated parameters in each model and the values in bold correspond to the best performing models. Weather the criterion is AIC or BIC, constant conditional correlation specifications are outperformed by their dynamic counterparts and symmetric specifications are outperformed by specifications including leverage effects.

Table 2 Maximized log-likelihood and information criteria

Models	Np	LogLik	AIC	BIC
DBEKK	12	-18321	36666	36737
BEKK	24	-18218	36484	36615
ABEKK	33	-17989	36044	36240
AGARCH	33	-18011	36088	36284
CCC	12	-18202	36428	36499
DCC	14	-17695	35414	**35501**
FI-DCC	17	-17684	**35402**	**35503**
FI-EDCC	23	-17661	**35368**	35505

3 The forecasting exercise

Evaluation of volatility forecasts is particularly challenging since volatility itself is latent and thus unobservable even ex post. In this case to compare model based forecasts with ex post realizations a statistical or an economic loss function as well as a proxy for the true unobservable conditional variance matrix have to be chosen. Proxy might lead to a different ordering of competing models that would be obtained if the true volatility were observed. To avoid a distorted outcome, the choice of an appropriate loss function is crucial. In this paper, we follow Bauwens et al. (2016) and use several loss functions, robust to noisy proxies. i.e. expected to provide the same forecasts ranking using the true conditional covariance or a conditionally unbiased proxy. Their definition is provided in Table 3 where H_{it} , for $i = 1, \dots, k$ denotes each model predicted covariance matrix for day t, $\hat{\Sigma}_t$ is the proxy of the conditional covariance matrix, ι is a vector of ones, T is the out of sample length and n is the the sample size. As a proxy for the conditional variance matrix at day t we use the matrix of the outer products of the daily mean forecast errors, $e_{T+1}e'_{T+1}$ which is a conditionally unbiased proxy. The forecasting ability of the set of proposed models is evaluated over a series of 630 out-of sample predictions. We compare the one day ahead conditional variance matrix forecasts based on the models estimated. We divide the full data set into two periods: the *in-sample period* from 02 August 2004 to 9 January 2014 (2430 observations), the *out of sample* with 510 observations from 10 January 2014 to 31 December 2015, used for forecasting evaluation. Forecasts are constructed using a fixed rolling window scheme: the estimation period is rolled forward by adding one new daily observation and dropping the most distant observation. Models parameters are re-estimated each day to obtain tomorrow volatility forecasts and the sample size employed to estimate is fixed and any dependence on the mean dynamics has been accounted for by fitting a VAR(1), so the mean forecasts do not depend on the models. This scheme satisfies the assumptions required by the MCS method and the SPA test and allows a unified treatment of nested and non-nested models. For each statistical loss function, we evaluate the significance of loss functions differences by means of the SPA and MCS.

Table 3 Loss function

Loss function		Type
Frobenius	$tr\left[\left(\hat{\Sigma}_t - H_{it}\right)'\left(\hat{\Sigma}_t - H_{it}\right)\right]$	Symmetric
Euclidean	$vech\left(\hat{\Sigma}_t - H_{it}\right)' vech\left(\hat{\Sigma}_t - H_{it}\right)$	Symmetric
MSFE	$\frac{1}{T} vec\left(\hat{\Sigma}_t - H_{it}\right)' vec\left(\hat{\Sigma}_t - H_{it}\right)'$	Symmetric
QLIKE	$\log\lvert H_t\rvert + vec\left(H_{it}^{-1}\hat{\Sigma}_t\right)' \iota$	Symmetric
Stein	$tr\left(H_{it}^{-1}\hat{\Sigma}_t\right) - \log\left\lvert H_{it}^{-1}\hat{\Sigma}_t\right\rvert - n$	Asymmetric
VDN	$tr(\hat{\Sigma}_t \log\hat{\Sigma}_t - \hat{\Sigma}_t \log H_{it} - \hat{\Sigma}_t + H_{it})$	Symmetric

We follow Hansen (2005) and obtain the p-values of the test by bootstrap. We implement a block bootstrap with block length equal to 2 and 10000 bootstrap samples. We find that the hypothesis of constant correlation is always rejected, as well as the hypothesis of short memory. The hypothesis of symmetry in the volatility dynamics is rejected in most benchmarks, and allowing for dynamic correlations significantly improves the models' forecasting accuracy. In the overall it appears that the most valid specification in this application is the fractionally integrated exponential DCC model that captures well the dynamics of the variance covariance matrix. The MCS methodology identifies a set of models with equivalent predictive ability which outperform all the other competing models at a given confidence level α with respect to a particular loss function. MCS determines the set of models that at a given confidence level have the best forecasting performance. We use a block bootstrap scheme to obtain the quantiles of the distribution. The block length bootstrap parameter is set equal to 2 and the number of bootstrap sample used is 10000. At the 90% confidence level the asymmetric BEKK and the fractionally integrated exponential DCC are included in the MCS resulting from the Euclidean, Frobenious, MSFE and VDN loss functions. The fractionally integrated DCC is included in the MCS deriving from the Euclidean, Frobenious and MSFE loss functions. The highest number of models (eight) is included for the Euclidean and Frobenious loss functions. The most striking result is the inclusion of the fractionally integrated exponential DCC model in the MCS of four loss functions supporting the hypothesis that the inclusion of long memory, asymmetries and time varying correlations significantly improve forecasting accuracy.

The possible efficiency gains of using long-memory asymmetric MGARCH models over short memory benchmarks for one-step ahead Value at Risk forecasting for equally weighted portfolios. To this end, we focus on the models' ability to predict the tail behavior of the returns rather than obtaining the 'best' volatility model.

We forecast the one day ahead Value at Risk for each of the models under comparison at 5%, 2.5% and 1% levels, and we assess their accuracy using statistical back-testing. We are concerned with both the long and short positions *VaR*. So we focus respectively on the left and right tail of the forecasted distribution of returns and we assess the models joint ability to delivery accurate VaR forecasts for both tails. To asses the accuracy of the *VaRs* obtained by the different models we test weather the failure rate implied by each model is statistically equal to the expected one. A popular back-testing procedure is based on the unconditional coverage test of Kupiec (e.g. Giot and Laurent, 2003). The test is a likelihood ratio test, built under the assumption that VaR violations are independent. Under the null, the test statistic is distributed as a χ^2-distribution with two degrees of freedom. Results for the short memory constant correlation models are homogenous for short and long VaRs, leading in all cases to rejection of the null hypothesis, regardless of the model structure. Models with dynamic conditional correlations perform much better passing all the tests with the occasional rejection for the most extreme quantiles. Models with dynamic conditional correlations and long memory adequately forecast *VaRs* at all levels. In conclusion for equally weighted portfolios, reliable *VaR* forecasts can be obtained under the assumption of conditionally normally standardized portfolio returns, by using DCC-type of models that include long range dependence and asymmetries in the individual volatilities.

As a general finding, fractionally integrated dynamic conditional correlation models display good in-sample fit. Using a fixed rolling window scheme, we assess the one-day ahead forecasting accuracy of the models with the MCS method and the SPA test using several matrix loss functions, robust to the choice of the volatility proxy. Short memory constant correlations models are always rejected in favour of long memory dynamic correlation models, and that the use of the latter significantly improves forecasts accuracy from a statistical as well as a risk management perspective.

References

1. Bauwens, L., Laurent, S., Rombouts, J. V., 2006. Multivariate GARCH models: a survey. Journal of applied econometrics 21, 79-109.
2. Bauwens, L., Braione, M., Storti, G., 2016. Forecasting comparison of long term component dynamic models for realized covariance matrices. Annals of Economics and Statistics/Annales d'Économie et de Statistique, (123/124), 103-134.
3. Chang, C. L., McAleer, M., Tansuchat, R., 2010. Analyzing and forecasting volatility spillovers, asymmetries and hedging in major oil markets. Energy Econ. 32, 1445–1455.
4. Hansen, P.R., 2005. A test for superior predictive ability. Journal of Business Economics and Statistics 23, 365–380.
5. Giot, P., Laurent, S., 2003. Value-at-risk for long and short trading positions. Journal of Applied Econometrics 18, 641-663.
6. Silvennoinen, A., Teräsvirta, T. , 2009. Multivariate GARCH models. Handbook of financial time series, 201-229.
7. Wang, Y., Wu, C, 2010. Forecasting energy market volatility using GARCH models: Can multivariate models beat univariate models?. Energy Economics 34, 2167-2181.

Improved bootstrap simultaneous prediction limits

Limiti di previsione simultanei migliorati basati su procedure bootstrap

Paolo Vidoni

Abstract This paper concerns the problem of constructing joint prediction regions having coverage probability equal or close to the target nominal value. In particular, the focus is on prediction regions defined using a system of simultaneous prediction limits. These regions are not necessarily of rectangular form and each component prediction interval depends on the preceding future observations. The specification of prediction regions with well-calibrated coverage probability has been considered in [2] and [5]. In this paper we consider an asymptotically equivalent procedure, which extends to the multivariate setting the bootstrap-based approach proposed in [3]. A simple application to autoregressive time series models is presented.
Abstract *In questo lavoro si considera il problema della costruzione di regioni di previsione con probabilità di copertura uguale o prossima a quella nominale, con particolare attenzione a regioni basate su limiti di previsione simultanei. Tali regioni non hanno necessariamente una forma rettangolare e ogni intervallo componente dipende dalle osservazioni future precedenti. La specificazione di regioni ben calibrate è stata studiata in [2] e [5]. In questo contributo si presenta una procedura di calcolo asintoticamente equivalente, e di facile implementazione, che estende all'ambito multivariato la procedura bootstrap introdotta in [3]. Si propone, infine, una semplice applicazione al caso dei modelli autoregressivi per serie storiche.*

Key words: bootstrap calibration, coverage, simultaneous prediction, time series

1 Introduction and preliminaries

This paper concerns the problem of constructing multivariate prediction regions having coverage probability equal or close to the target nominal value. In partic-

Paolo Vidoni
Department of Economics and Statistics, University of Udine, via Tomadini, 30/A I-33100 Udine, Italy, e-mail: paolo.vidoni@uniud.it

ular, the focus here is on multivariate prediction regions defined using a system of simultaneous prediction limits. These regions are not necessarily of rectangular form and they can be usefully considered whenever there is a natural order in the observations, such as for time series data, since each component prediction interval turns out to be influenced by the preceding future observations. With regard to time series applications, a system of simultaneous prediction intervals could be viewed as an alternative to a sequence of marginal prediction intervals at different periods into the future, which do not properly account for the actual dynamic evolution of the interest phenomenon.

Let (Y,Z) be a continuous random vector having joint density function $f(y,z;\theta)$, with $\theta \in \Theta \subseteq \mathbf{R}^d$, $d \geq 1$, an unknown d-dimensional parameter; $Y = (Y_1,\dots,Y_n)$, $n \geq 1$, is observable, while $Z = (Z_1,\dots,Z_m)$, $m \geq 1$, denotes a future, or yet unobserved, random vector. Although prediction problems may be studied from different perspectives, the aim here is to define an α-prediction region for Z, that is a random set $R(Y,\alpha) \subset \mathbf{R}^m$, depending on the observable sample Y and on the nominal coverage probability α, such that

$$P_{Y,Z}\{Z \in R(Y,\alpha);\theta\} = E_Y[P_{Z|Y}\{Z \in R(Y,\alpha)|Y;\theta\};\theta] = \alpha, \qquad (1)$$

for every $\theta \in \Theta$ and for any fixed $\alpha \in (0,1)$. The above probability is called coverage probability and it is calculated with respect to the joint distribution of (Z,Y); moreover, the expectation is with respect to Y and $P_{Z|Y}\{\cdot;\theta\}$ is the probability distribution of Z given Y.

When there exists a transitive statistics $U = g(Y)$, it is natural to consider the conditional coverage probability such that, exactly or approximately,

$$P_{Y,Z|U}\{Z \in R(Y,\alpha)|U = u;\theta\} = E_{Y|U}[P_{Z|U}\{Z \in R(Y,\alpha)|U;\theta\}|U = u;\theta] = \alpha, \quad (2)$$

where the probability and the expectation are conditioned on $U = u$. For example, if we consider an autoregressive (AR) model of order 1, the transitive statistics is $U = Y_n$. Obviously, conditional solutions satisfying (2) also satisfy (1) and, when we can not find a transitive statistic, the conditional approach is meaningless.

The easiest way for making prediction on Z is to define a prediction region by using the estimative (plug-in) predictive distribution $P_{Z|Y}\{\cdot;\hat{\theta}\}$, where the unknown parameter θ is substituted with an asymptotically efficient estimator $\hat{\theta}$ based on Y, such that $\hat{\theta} - \theta = O_p(n^{-1/2})$; we usually consider the maximum likelihood estimator or any asymptotically equivalent alternative estimator. However, estimative α-prediction regions $R_e(Y,\alpha)$ are not entirely adequate predictive solutions, since the additional uncertainty introduced by assuming $\theta = \hat{\theta}$ is underestimated and then the (conditional) coverage probability differs from α by a term usually of order $O(n^{-1})$. This lack of accuracy can be substantial for small n and/or large m.

Here we focus on a particular estimative prediction region $R_e(Y,\alpha)$ based on the system of simultaneous prediction limits defined as quantiles of the conditional distributions of the components of vector $Z = (Z_1,\dots,Z_m)$. We assume, for simplifying the exposition, that (Y,Z) follows a first-order Markovian dependence structure (so that $U = Y_n$) and then we set

$$R_e(Y,\alpha) = \{z \in \mathbf{R}^m : z_i \leq \hat{q}_i(\alpha_i), i = 1,\ldots,m\}, \tag{3}$$

where $\hat{q}_i(\alpha_i) = q_i(\alpha_i, z_{i-1}; \hat{\theta})$, $i = 1,\ldots,m$, is the α_i-quantile of the conditional distribution of Z_i given $Z_{i-1} = z_{i-1}$, evaluated at $\theta = \hat{\theta}$, with $Z_0 = Y_n$. Finally, we assume $\prod_{i=i}^m \alpha_i = \alpha$ in order to assure that $R_e(Y,\alpha)$ is an α-prediction region, namely that $P_{Z|Y_n}\{Z \in R_e(Y,\alpha)|Y_n = y_n; \hat{\theta}\} = \alpha$. Note that the conditional prediction limit $\hat{q}_i(\alpha_i)$, for each $i = 2,\ldots,m$, is obtained recursively as a function of the previous, unknown future observation z_{i-1}.

Corcuera and Giummolè [2] find that the (conditional) coverage probability of the estimative prediction region (3) is

$$P_{Y,Z|Y_n}\{Z \in R_e(Y,\alpha)|Y_n = y_n;\theta\} = E_{Y|Y_n}\left\{\int_{-\infty}^{\hat{q}_1(\alpha_1)} \cdots \int_{-\infty}^{\hat{q}_m(\alpha_m)} f_{Z|Y_n}(z|Y_n;\theta)dz \Big| Y_n = y_n;\theta\right\}$$
$$= C_m(\alpha_1,\ldots,\alpha_m;\theta,y_n) = \alpha + Q_m(\alpha_1,\ldots,\alpha_m;\theta,y_n) + O(n^{-3/2}),$$

where $f_{Z|Y_n}(z|y_n;\theta)$ is the joint conditional density of Z given $Y_n = y_n$. Moreover, after tedious calculations, an explicit expression for the $O(n^{-1})$ coverage error term $Q_m(\alpha_1,\ldots,\alpha_m;\theta,y_n)$ is also derived.

2 Improved simultaneous prediction

In order to improve the estimative predictive approach a number of solutions have been proposed. One of these strategies (see, for example, [1] and [4]) is to define an explicit modification for the estimative prediction limits, so that the associated coverage probability turns out to be equal to the target α with a high degree of accuracy. With regard to the univariate case (namely, $m = 1$, $Z = Z_1$ and $\alpha = \alpha_1$), given the estimative α_1-prediction limit $\hat{q}_1(\alpha_1)$, it is easy to prove that the modified estimative prediction limit

$$\tilde{q}_1(\alpha_1) = \hat{q}_1(\alpha_1) - \frac{Q_1(\alpha_1;\hat{\theta},y_n)}{f_{Z_1|Y_n}(\hat{q}_1(\alpha_1)|y_n;\hat{\theta})}, \tag{4}$$

reduces the coverage error to order $o(n^{-1})$. Here, $f_{Z_1|Y_n}(z_1|y_n;\theta)$ is the conditional density function of Z_1 given $Y_n = y_n$. A potential drawback of this strategy is that the evaluation of the fundamental term $Q_1(\alpha_1;\theta,y_n)$ may require complicated asymptotic calculations. To overcome this difficulty, Ueki and Fueda [3] show that the modifying term of the improved prediction limit (4) can be equivalently expressed as $\hat{q}_1(C_1(\alpha_1;\theta,y_n)) - \hat{q}_1(\alpha_1)$ and then, to the relevant order of approximation, the modified estimative prediction limit (4) corresponds to

$$\tilde{q}_1(\alpha_1) = 2\hat{q}_1(\alpha_1) - \hat{q}_1(C_1(\alpha_1;\theta,y_n)). \tag{5}$$

Therefore, the computation is greatly simplified, since we need only the value of the coverage probability $C_1(\alpha_1;\theta,y_n) = E_{Y|Y_n}\{F_{Z_1|Y_n}(\hat{q}_1(\alpha_1)|Y_n;\theta)|Y_n = y_n;\theta\}$, with $F_{Z_1|Y_n}(z_1|y_n;\theta)$ the conditional distribution function of Z_1 given $Y_n = y_n$, which can be usually estimated using a simple parametric bootstrap procedure.

The approach based on high-order analytical corrections can be extended to the multivariate case. In particular, Corcuera and Giummolè [2] specify a system of improved prediction limits $\tilde{q}_1(\alpha_1),\ldots,\tilde{q}_m(\alpha_m)$, where $\tilde{q}_1(\alpha_1)$ is defined as in (4), whereas each $\tilde{q}_i(\alpha_i)$, $i = 2,\ldots,m$, requires a further correction term in order to account for the additional dependence introduced, among the limits, by substituting θ with the same $\hat{\theta}$. This second correction term is far more complex that the first one.

In order to simplify the calculation, using a general result presented in [5], we prove that it is possible to extend the Ueki and Fueda's procedure to the multivariate setting. More precisely, an asymptotically equivalent expression for the improved simultaneous prediction limits corresponds to

$$\tilde{q}_i(\alpha_i) = 2\hat{q}_i(\alpha_i) - \hat{q}_i(C_i(\alpha_i;\theta,y_n,z_{(i-1)})),\ i = 1,\ldots,m, \tag{6}$$

with $z_{(i-1)} = (z_1,\ldots,z_{i-1})$. For $i = 1$, $C_1(\alpha_1;\theta,y_n)$ is the coverage probability of $\hat{q}_1(\alpha_1)$ as given in (5) and, for $i = 2,\ldots,m$, we consider the conditional coverage probability of $\hat{q}_i(\alpha_i)$ given $Z_{(i-1)} = z_{(i-1)}$ defined as

$$C_i(\alpha_i;\theta,y_n,z_{(i-1)}) = \frac{E_{Y|Y_n}\left\{\frac{f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|Y_n;\theta)}{f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|Y_n;\hat{\theta})} F_{Z_i|Z_{i-1}}(\hat{q}_i(\alpha_i)|z_{i-1};\theta)\Big|Y_n = y_n,\theta\right\}}{E_{Y|Y_n}\left\{\frac{f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|Y_n;\theta)}{f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|Y_n;\hat{\theta})}\Big|Y_n = y_n,\theta\right\}}, \tag{7}$$

where $F_{Z_i|Z_{i-1}}(z_i|z_{i-1};\theta)$ is the conditional distribution function of Z_i given $Z_{i-1} = z_{i-1}$ and $f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|y_n;\theta)$ is the joint conditional density of $Z_{(i-1)}$ given $Y_n = y_n$. Also the conditional coverage probability (7) can be estimated using a fairly simple bootstrap parametric approach and, since the explicit expression for the correction terms is not required, this greatly simplifies the computation of the improved limits.

3 An application to a simple autoregressive model

Let us consider a stationary AR(1) process $\{Y_j\}_{j\geq 1}$ defined as

$$Y_j = \mu + \rho\,(Y_{j-1} - \mu) + \varepsilon_j, \quad j \geq 1,$$

where $\mu \in \mathbf{R}$, $|\rho| < 1$ and $\{\varepsilon_j\}_{j\geq 1}$ is a sequence of independent normal distributed random variables with zero mean and variance $\sigma^2 > 0$. Then, using the notation introduced in Section 1, $Y = (Y_1,\ldots,Y_n)$, $Z = (Z_1,\ldots,Z_m) = (Y_{n+1},\ldots,Y_{n+m})$ and $\theta = (\theta_1,\theta_2,\theta_3) = (\mu,\rho,\sigma^2)$ is the unknown model parameter. Furthermore $\hat{\theta} =$

$(\hat{\mu}, \hat{\rho}, \hat{\sigma}^2)$ is the vector of the corresponding maximum likelihood estimators which, in this case, are explicitly known.

Since Z_i given $Z_{i-1} = z_{i-1}$, $i = 1, \ldots, m$, follows a normal distribution with mean $\mu + \rho\,(z_{i-1} - \mu)$ and variance σ^2, it is immediate to define the estimative prediction region $R_e(Y, \alpha)$ as specified by (3), with simultaneous prediction limits $\hat{q}_i(\alpha_i) = \hat{\mu} + \hat{\rho}\,(z_{i-1} - \hat{\mu}) + u_{\alpha_i}\hat{\sigma}$, $i = 1, \ldots, m$. Here, u_{α_i} is such that $\Phi(u_{\alpha_i}) = \alpha_i$, where $\Phi(\cdot)$ is the distribution function of a standard normal random variable, and $\prod_{i=1}^{m} \alpha_i = \alpha$. Using the bootstrap-based procedure outlined in Section 2, we obtain the modified simultaneous prediction limits (6), which are supposed to improve the coverage accuracy of the estimative solution.

We also consider a sequence of m marginal prediction limits, which correspond the the plug-in estimates of the α_i-quantile, for $i = 1, \ldots, m$, of the conditional distribution of Z_i given $Y_n = y_n$. Notice that the first marginal prediction limit corresponds to $\hat{q}_1(\alpha_1)$. These prediction limits are computed repeatedly one period at a time and they define a rectangular-shaped prediction region. In this case, the nominal coverage probability is not equal to $\prod_{i=1}^{m} \alpha_i = \alpha$, since the component prediction limits are not independent of each other. Furthermore, by applying a bootstrap-calibrated procedure to these marginal prediction limits, as supposed to be independent, we try to improve, also in this different situation, the coverage accuracy of the corresponding prediction region.

Table 1 presents the results of a preliminary simulation study for comparing the coverage accuracy of prediction regions based on estimative and bootstrap-calibrated simultaneous prediction limits and on estimative and bootstrap-calibrated marginal prediction limits. Conditional coverage probabilities, with nominal level $\alpha = 0.9, 0.95$, are estimated using 1,000 samples of dimension $n = 50, 100$ simulated from an AR(1) model with the last observation fixed to $y_n = 1$ and assuming $y_0 = 0$; indeed, we consider $\mu = 1$, $\sigma^2 = 1$ and (a) $\rho = 0.5$, (b) $\rho = 0.8$. The prediction regions have dimension $m = 5, 10$ and $\alpha_i = \alpha^{1/m}$, $i = 1, \ldots, m$. The bootstrap procedure is based on 1,000 conditional bootstrap samples. The results are in accordance with the theoretical findings and show that the improved bootstrap-based procedures remarkably improve on the estimative ones. The improvement is more pronounced when the dimension m of the future random vector is high with respect to n. Moreover, the bootstrap-calibrated technique seems to improve the coverage accuracy of the marginal estimative prediction limits as-well, accounting also for the dependence among the component prediction limits. This is an important point which require further attention.

Finally, we conclude this section by presenting the following Figure 1, which describes a simulated path of dimension $n = 80$ from an AR(1) Gaussian model with $y_0 = 1$, $\mu = 1$, $\sigma^2 = 1$ and $\rho = 0.5$, and a sequence of $m = 50$ future simulated observations generated from the same model. Moreover, we draw the sequence of estimative and improved simultaneous prediction intervals with level $\alpha_i = 0.9$, together with the estimative and improved marginal prediction intervals with the same nominal level. The simultaneous prediction intervals account for the actual evolution of the interest time series. Note that, using the bootstrap-based approach, the

Table 1 AR(1) Gaussian model with $\mu = 1$, $\sigma^2 = 1$ and (a) $\rho = 0.5$, (b) $\rho = 0.8$. Conditional coverage probabilities for the simultaneous (estimative and improved) and marginal (estimative and improved) prediction limits of level $\alpha = 0.9, 0.95$, with $m = 5, 10$. Estimation is based on 1,000 Monte Carlo conditional (on $y_n = 1$) samples of dimension $n = 50, 100$, with $y_0 = 0$. The bootstrap procedure is based on 1,000 conditional bootstrap samples.

			(a)				(b)			
α	n	m	Simultaneous		Marginal		Simultaneous		Marginal	
			Estimative	Improved	Estimative	Improved	Estimative	Improved	Estimative	Improved
0.9	50	5	0.860	0.905	0.885	0.910	0.860	0.908	0.878	0.897
		10	0.838	0.898	0.838	0.881	0.824	0.869	0.832	0.866
	100	5	0.884	0.907	0.888	0.896	0.886	0.890	0.886	0.890
		10	0.882	0.913	0.889	0.910	0.875	0.914	0.904	0.910
0.95	50	5	0.921	0.953	0.936	0.955	0.920	0.956	0.924	0.948
		10	0.905	0.946	0.900	0.931	0.888	0.929	0.884	0.915
	100	5	0.938	0.954	0.932	0.946	0.937	0.957	0.933	0.940
		10	0.937	0.962	0.940	0.955	0.934	0.958	0.951	0.958

estimative prediction limits are suitably calibrated in order to improve the coverage accuracy.

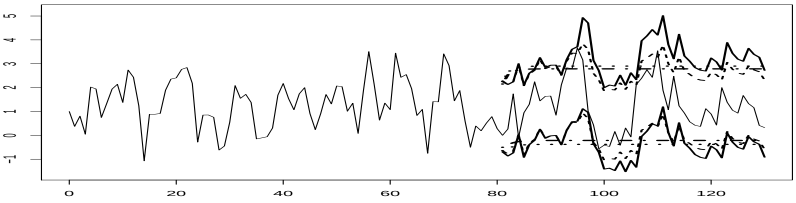


Fig. 1 Simulated observations from an AR(1) Gaussian model with $y_0 = 1$, $\mu = 1$, $\sigma^2 = 1$ and $\rho = 0.5$. Simultaneous estimative (dashed) and improved (solid) prediction intervals and marginal estimative (dotted) and improved (dot-dashed) prediction intervals with coverage probability 0.9.

Acknowledgements This research was partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003 and by the University of Udine under the PRID 2017 research grants.

References

1. Barndorff-Nielsen, O.E., Cox, D.R.: Prediction and asymptotics. Bernoulli **2**, 319–340 (1996)
2. Corcuera, J.M., Giummolè, F.: Multivariate prediction. Bernoulli **12**, 157–168 (2006)
3. Ueki, M., Fueda, K.: Adjusting estimative prediction limits. Biometrika **94**, 509–511 (2007)
4. Vidoni, P.: A note on modified estimative prediction limits and distributions. Biometrika **85**, 949–953 (1998)
5. Vidoni, P.: Calibrated multivariate distributions for improved conditional prediction. Journal of Multivariate Analysis **142**, 16–25 (2015)

Data Management

Using web scraping techniques to derive co-authorship data: insights from a case study

Utilizzo di tecniche di web scraping per derivare reti di co-authorship: evidenze da un caso studio

Domenico De Stefano, Vittorio Fuccella, Maria Prosperina Vitale, Susanna Zaccarin

Abstract The aim of the present contribution is to discuss the first results of the application of web scraping techniques to derive co-authorship data among scholars. A semi-automatic tool is adopted to retrieve metadata from a platform introduced for managing and supporting research products in Italian universities. The co-authorship relationships among Italian academic statisticians will be used as basis to analyze updated collaborations patterns in this scientific community.

Abstract *Il presente contributo riporta i primi risultati delle procedure di web scraping utilizzate per ottenere dati sulla co-autorship tra studiosi. Uno strumento semi-automatico è stato utilizzato per estrarre i metadati delle pubblicazioni da una piattaforma online introdotta per la gestione dei prodotti della ricerca delle università italiane. Le relazioni di co-authorship tra gli statistici italiani saranno utilizzate come base per individuare i patterns di collaborazione recenti.*

Key words: Bibliographic data, Web scraping, Network data mining

Domenico De Stefano
Department of Political Science, University of Trieste e-mail: ddestefano@units.it

Vittorio Fuccella
Department of Informatics, University of Salerno e-mail: vfuccella@unisa.it

Maria Prosperina Vitale
Department of Economics and Statistics, University of Salerno e-mail: mvitale@unisa.it

Susanna Zaccarin
Department of Economics, Business, Mathematics and Statistics ”B. de Finetti”, University of Trieste e-mail: susanna.zaccarin@deams.units.it

1 Introduction

Several bibliographic sources are available online to retrieve co-authorship relations to analyze scientific collaboration among groups of scholars. Usually, scientific collaboration studies are based on international databases (e.g. ISI-WoS or Scopus) containing high-impact publications. Whereas the interest is to describe collaboration patterns among scholars involved in a field and/or affiliated to an institution, these sources can provide a partial coverage of their scientific production. Hence, it emerges the need to integrate these results by using local bibliographic archives. In addition, the complexities related to the scraping data process from heterogeneous online sources at international and national level, often including different data format, are well-recognized (7; 8; 6). Data mining tools have been introduced for this topic highlighting the strengths and weaknesses of each of them (7).

In this scenario, scientific production of the Italian academic scholars can be retrieved from general and thematic international bibliographic archives as well as national ones (2). To describe collaboration in the national academic community, publications can be collected from individual web pages ("sito docente"), managed by the Italian Ministry of University and Research (MIUR) and the Cineca consortium. Unfortunately, the access to this database is not freely available, due to privacy policies.

The recent introduction of the Institutional Research Information System (IRIS) developed by Cineca consortium seems to furnish a unique platform in Italy for managing and supporting research in academic and research institutions. Within this system, it is available an open archive module for the repository of the research products allowing the storage, the consultation and the enhancement of these outputs. Thanks to this tool, the affiliated universities can access to a system able to communicate with the national (i.e. "sito docente" of MIUR) and international databases for the management and dissemination of scholars' scientific publications. Considering the list of institutions affiliated to the IRIS platform,[1] 65 Italian universities out of 97 [2] (in which 67 are public universities, 19 private universities and 11 private online universities) adopt the IRIS platform for publications data storage.

The aim of the present contribution is to discuss the main results of the first step of web scraping procedures, before to obtain clean data to derive co-authorship relationships among scholars. A semi-automatic tool is adopted to retrieve publication metadata from the IRIS platform with the purpose of automatically extract the data from the system in order to obtain a good coverage of the author scientific production and reducing the manual adjustments to manage errors. The derived co-authorship relationships among Italian academic statisticians, considering the publications until 2017, will be used as basis to derive updated collaborations patterns in this scientific community. The contribution offers also the possibility to compare

[1] For detail see the IRIS webpage https://www.cineca.it/en/content/IRIS-institutional-research-information-system.

[2] For detail see http://www.miur.gov.it/istituzioni-universitarie-accreditate.

the results with the previous network analysis carried on the same target population with data updated to 2010 (2; 3).

The paper is organized as follows. Section 2 reports details on the population under analysis and on the web scraping procedure adopted to extract data from the IRIS system. Section 3 traces the directions of the further network analysis.

2 Web scraping techniques to extract publications of Italian statisticians

The present contribution deals with the retrieval of proper data to construct co-authorship network in Statistics starting from the IRIS online platform. In particular, we focus on the academic statisticians in Italy, that is, those scientists classified as belonging to one of the five subfields established by the governmental official classification: Statistics (Stat), Statistics for Experimental and Technological Research (Stat for E&T), Economic Statistics (Economic Stat), Demography (Demo), and Social Statistics (Social Stat). The target population is composed of the 721 statisticians who have permanent positions in Italian universities, as recorded in the MIUR database at July 2017. Table 1 reports the composition of the statisticians by Statistics subfields, gender, academic ranking and university geographic location. A comparison with the number of scholars by Statistics subfields with a permanent position in March 2010 is also given (Table 1 in De Stefano et al. 2, p. 373).

Table 1 Italian academic statisticians by Statistics subfields (%) in 2017 and (total) and 2010 comparison. Source: MIUR, March 2010 and July 2017

	All	Stat	Stat for E&T	Economic Stat	Demo	Social Stat
Gender						
Female	47.2	49.4	30.0	37.2	58.6	47.7
Male	52.8	50.6	70.0	62.8	41.4	52.3
Academic ranking						
Researcher	33.8	33.7	45.0	33.1	38.6	27.7
Associate professor	38.3	39.2	35.0	35.2	34.3	44.6
Full professor	27.9	27.1	20.0	31.7	27.1	27.7
University geographic location						
North	42.7	46.3	15.0	38.6	40.0	40.0
Center	25.7	24.0	20.0	31.7	27.1	23.1
South	31.6	29.7	65.0	29.7	32.9	36.9
Total (July 2017)	721	421	20	145	70	65
Total (March 2010)	792	443	30	160	85	74
Relative difference 2017-2010 (%)	-9.0	-5.0	-33.3	-9.4	-17.6	-12.2

Starting from scraping data techniques, a semi-automated tool based on two main steps is used to retrieve the publication metadata of the population of Italian academic statisticians in the IRIS platform. Indeed, each author has a page from which it is possible to access to the data of his/her publications. The tool is implemented in Java. Besides Java standard libraries to download Web pages, the Tagsoup library[3] is used for parsing well-formed or even unstructured and malformed HTML. This tool is programmed with the aim of automatically extract the data from the system obtaining a good coverage of the author publications and reducing the manual adjustments to manage errors or uncertainty conditions.

The input information is a table containing references (name, surname and academic institution) of the 721 statisticians.

In the *first step*, the URL of IRIS page is retrieved for each author. It is worth noting that each institution hosts a different deployment of the system, thus each statistician is linked to the index page of the IRIS deployment of his/her institution. Then a query is launched on a specific search by author interface available on the system. The interface responds by outputting a Web page containing a list of authors indicated by name and surname, each associated with a link to the author's page. The last name of the author is used as a query string and both author's name and surname are considered to match an item in the list. In the case of a single match, the link is directly captured. In case of no match or multiple matches, the procedure returned an error. As a result of the first step, the complete database of publications records for each author is available. The author's page contains the list of publications of which the person is co-author. If an author has more than 100 publications, these are necessarily split into multiple pages. Each publication in the list is associated to a link to a new page containing the details of the publication (title, authors, venue, year of publication and various identifiers –URL, DOI, ISI codes WoS and Scopus and so on).

In the *second step*, the proposed web scraping procedure retrieved and followed the links of each author publication in order to download these metadata. In a few cases, it will be necessary to manually retrieve the link to the author's publication pages to check the aforementioned errors and to integrate the retrieved metadata for the authors not found by the tool. The complete database of publication records for each author derived at step 2 reports information for the entire population of statisticians. It could contain many duplications in presence of publications co-authored by statisticians affiliated to different institutions. Indeed, in each IRIS system the same publication can be reported with a different format. To manage this issue, it is therefore necessary a further phase of record linkage, in which the records corresponding to the same publication are automatically reconciled. Where available, the correspondence of any of the aforementioned authors' identifiers allowed a sure identification of the match. In case of lack of unique identifiers, the records can be reconciled using information related to the title, list of authors and year of publication, following the procedure described in Fuccella et al. (3).

The aforementioned procedure allowed us to obtain a good coverage of the target population of authors in the first data extraction phase from IRIS platform. The metadata of the publications of around 80% of all statisticians is available from the

[3] For details see https://hackage.haskell.org/package/tagsoup.

extraction made in April 2018. After a manual check to integrate the retrieved metadata, we improved the coverage rate for all subfields (Table 2). This result seems almost in line with the authors' coverage obtained by using three different archives (Table 2 in De Stefano et al. 2, p. 374).

The tool returned zero publications for some authors, especially belonging to the Statistics and Economic Stat subfields. These errors are mainly due to the absence of the IRIS platform in some institutions (mainly private and online universities) and limitations to its access (i.e., a password is required to access to the system). Errors happened also when the author search returned more than one match.

Table 2 Statisticians found before and after manual check and descriptive statistics of the retrieved publications by Statistics subfields

	Authors			Publications				
	Found	Manual check	% found	Min.	Max.	Median	Average	Dev.ST.
All	555	82	88.3	1	333	50.0	56.4	38.0
Stat	319	54	88.6	1	292	49.0	54.9	34.6
Stat for E&T	18	2	100.0	8	126	46.0	55.5	34.1
Economic Stat	110	11	83.4	2	196	42.0	45.2	29.1
Demo	55	8	90.0	12	314	55.0	68.2	48.6
Social Stat	53	7	92.3	8	333	68.5	75.6	50.1

Focusing on the retrieved bibliographic data, several issues affect the data quality. First, the IRIS platform content is different and independent for each university and there is no automatic procedure that allows to match the same publication co-authored by authors enrolled in different institutions. Therefore for each co-authored publication, a number of duplication of this product equal to the number of the co-authors hired by different universities can be found. Second, the imputation of some crucial fields for the network construction –e.g. the name of the external authors– is up to the individual researcher and for this reason it can happen that same author names can be typed in several ways.

Both product and author name duplications should be addressed before the co-authorship network construction by adapting to this context the procedure proposed in Fuccella et al. (3).

3 Final remarks and future lines of research

The availability of the IRIS archive is certainly a promising tool but a lot of issues must be managed during the data collection process from this source. The unavailability of IRIS platform in some universities, the private access to the IRIS system in some cases, the different publication data format, and the presence of more than one record found for the same author are examples of errors obtained after the extraction of information. In addition, record linkage and author disambiguation processes need to be taken into account to obtain co-authorship data among scholars affiliated in different universities.

After the data cleaning to reconcile publication records, the detection of duplicates and the recognition of internal and external authors of the same publications, the co-authorship networks will be derived. In line with the findings discussed in the previous contributions, the advancements of the current study will be devoted to two main directions of analysis. First, the interest is in discovering clusters of scholars through community detection algorithms, comparing results from two well-known community detection algorithms (4; 1), and a new proposed method based on an adaptation of modal clustering procedure (5). Second, the stability of research groups and of collaboration behaviors will be analyzed in order to capture the effect of research assessment exercises, introduced in Italy to evaluate researchers and their scientific production.

References

[1] Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment. **2008**, P10008 (2008)

[2] De Stefano, D., Fuccella, V., Vitale, M. P., Zaccarin, S.: The use of different data sources in the analysis of co-authorship networks and scientific performance. Social Networks. **35**, 370-381 (2013)

[3] Fuccella, V., De Stefano, D., Vitale, M. P., Zaccarin, S.: Improving co-authorship network structures by combining multiple data sources: evidence from Italian academic statisticians. Scientometrics. **107**, 167-184 (2016)

[4] Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the national academy of sciences. **99**, 7821-7826 (2002)

[5] Menardi, G., De Stefano, D.: Modal clustering of social network. In: Cabras, S. and Di Battista, T. and Racugno, W. (eds) Proceedings of the 47th SIS Scientific Meeting of the Italian Statistical Society, CUEC Editrice, Cagliari (2014)

[6] Mitchell, R.: Web scraping with Python: collecting data from the modern web. Packt Publishing, Birmingham (2015)

[7] Murthy, D., Gross, A., Takata, A., Bond, S.: Evaluation and development of data mining tools for social network analysis. In: Ozyer, T., Erdem, Z., Rokne, J., Khoury, S. (eds.) Mining Social Networks and Security Informatics, pp. 183-202. Springer, Dordrecht (2013)

[8] Vargiu, E., Urru, M.: Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artificial Intelligence Research. **2**, 44-54 (2013)

Dealing with Data Evolution and Data Integration: An approach using Rarefaction

Luca Del Core, Eugenio Montini, Clelia Di Serio, Andrea Calabria

Abstract Heterogeneity and unreliability of data negatively influence the effectiveness and reproducibility of the results in all fields involving sampling techniques. Heterogeneity is mainly due to technological advances which imply improvements in measurements resolution. Unreliability or under-representativeness in data may be due to machine/software or human variances/errors, or other unidentifiable external factors. In the era of big data, technological evolution, and continuous data integration, scientists are increasingly facing with the problems of how to (1) identify and filter-out unreliable data, and (2) harmonize samples gauged with different platforms improved over time. This work is aimed at developing a new statistical framework to address both issues, showing results in real case scenarios.

Luca Del Core
San Raffaele Telethon Institute for Gene Therapy (SR-Tiget),
IRCCS San Raffaele Scientific Institute,
Via Olgettina 58, 20132, Milano, Italy
e-mail: delcore.luca@hsr.it

Eugenio Montini
San Raffaele Telethon Institute for Gene Therapy (SR-Tiget),
IRCCS San Raffaele Scientific Institute,
Via Olgettina 58, 20132, Milano, Italy
e-mail: montini.eugenio@hsr.it

Clelia Di Serio
University Vita-Salute San Raffaele,
University Centre of Statistics in the Biomedical Sciences,
Via Olgettina 58, 20132, Milano, Italy
e-mail: diserio.clelia@unisr.it

Andrea Calabria
San Raffaele Telethon Institute for Gene Therapy (SR-Tiget),
IRCCS San Raffaele Scientific Institute,
Via Olgettina 58, 20132, Milano, Italy
e-mail: calabria.andrea@hsr.it

Key words: Heterogeneity, Rarefaction, Filtering, Abundance Models, Species Pooling, Generalized Nonlinear Models, Hurlbert-Heck Model, Entropy, Integration Sites, Gene Therapy.

Motivation and background

Nowadays, with the advent of high throughput technologies, new statistical challenges are aimed at combining large heterogeneous datasets produced under different platforms having different efficiency, reliability and resolution. This is the case of biomedical data, where the follow-ups of a clinical cohort of patients under treatment may last over several decades and the monitoring of patient health-care must benefit by the biotechnological improvements continuously consolidated. Therefore, new statistical methods are required to understand how to consider, within a unique framework, time series that are observed over a long period, and how to distinguish whether the increased number of events is attributable to the change in technology rather than to the disease change itself. Thus, it is crucial to obtain reliable and harmonized time-course data, addressing the problems of (1) the identification and filtering of unreliable data, and (2) how to scale heterogeneous integrated data to obtain consistent results in the application domain. This work is proposed as a first step towards a mathematical/statistical solution of both problems.

Material and Methods

The filtering problem is addressed through the expected richness estimation via the Hurlbert-Heck (HH) curve[1, 2] and the Species Pooling (SP) methods[3, 4, 5] for an estimation of the unseen species. The base statistical methods have been previously applied in ecological and population-based studies. We exploited the Generalized Nonlinear Model (GNM) as an estimator of the HH curve properly rescaled. Then, using an empirical approach, we identified a minimum threshold for the richness over the whole cohort of data to filter out under-representative observations. To address the problem of data integration for reproducible results over continuous technological improvements and the scaling problem, we used a Rarefaction Method[1, 2].Both methodologies have been applied in biomedical science using molecular data (retroviral vector integration sites, IS) of Gene Therapy (GT) clinical trials, a good case study for the presence of heterogeneous data. The filtering technique is used basing on the richness in distinct number of ISs. The rarefaction approach allowed improving data integration of IS by rescaling data in order to obtain rarefied population measures (such as entropy indexes[6, 7]) that are more robust and homogeneous than the un-scaled ones, thus potentially improving the assessment of safety and long term efficacy of the treatment. A discussion of results is finally presented.

Results

Filtering Unreliable Data

Dealing with biological and molecular data, such as IS in GT studies, means dealing with high variability in data collection and sampling, due to the high variability of the available biological material (for example the different amount of DNA used in each test). Thus, the number of retrieved IS from each patient at different time points (IS_s Richness) may vary. Therefore, we have to evaluate the level of richness for each sample and filter-out those samples with an insufficient level of IS richness. To overcome this problem, the percentage $S_{\%}$ of richness in IS_s observed over the total can be defined as the ratio between the observed richness S and an estimator of the whole theoretical richness $\hat{S}_{tot}$, namely $S_{\%} = S/\hat{S}_{tot}$. The theoretical richness $\hat{S}_{tot}$ can be estimated in different ways, depending on the chosen Sampling Pooling (SP) technique[3, 4, 5]. In this work, the SP estimator is chosen based on the best Ranked Abundance Distribution (RAD)[8, 9] together with a p-leave out cross validation. Namely, if the general lognormal (gln) curve is chosen as the best AIC_c[10] RAD model among the candidates[1], or if, according to a χ^2 Goodness of Fit test, the gln distribution can be used in place of the optimal one, then the Preston estimator $\hat{S}^{tot}_{Preston}$[3] is used. Otherwise the performance of Chao and ACE estimators is compared[11] via a uniform leave-$p = .3$-out cross validation: the whole sample is considered as the sampling universe with a known total richness S_{obs} and the $[0,1]$-bounded quantity[2]

$$A^{est}_{absolute} = \min\{S_{obs}, \hat{S}_{est}\}/\max\{S_{obs}, \hat{S}_{est}\}$$

is used to compare the performance of the two non-parametric estimators. The more the accuracy of the estimator, the greater $A^{est}_{absolute}$ is: therefore, the estimator $best = argmin_{est}\{A^{est}_{absolute}\}$ is chosen if the gln distribution is rejected as RAD model. Also, in order to assess the accuracy of the species pooling estimator chosen among the candidates, during the $nFld = 100$ p-leave-out cross simulations, three additional accuracy indexes are calculated and compared with $A^{best}_{absolute}$. These are defined as

$$A_{effective} = 1 - |.7 - S^{0.7}_{obs}/\hat{S}^{0.7}_{tot}| \qquad A_{relative} = 1 - |S^{0.7}_{obs}/\hat{S}^{0.7}_{tot} - S^{1}_{obs}/\hat{S}^{1}_{tot}|$$

$$A_{cumulative} = \min\{\hat{S}^{0.7}_{tot}, \hat{S}^{1}_{tot}\}/\max\{\hat{S}^{0.7}_{tot}, \hat{S}^{1}_{tot}\}$$

where $S^{0.7}_{obs}$, $\hat{S}^{0.7}_{tot}$, S^{1}_{obs}, $\hat{S}^{1}_{tot}$ are the observed and estimated (using the chosen estimator) total richness in the 70% fold and in the whole sample respectively. By definition, they are $[0,1]$ bounded and the more the accuracy, the greater they are. As an

[1] Geometric Series, MacArthur's Broken Stick, Zipf-Mandelbrodt, Zipf, General lognormal are the candidate RAD models in the case study. All these distributions are fitted with the Maximum Likelihood Estimation (MLE) technique.

[2] $\hat{S}_{est}$ is the estimation of the total richness S_{obs} obtained using the the estimator $est \in \{Chao, ACE\}$ using the 70-random subsample.

example, these indexes are calculated in the case study, together with $A^{est}_{absolute}$, and the results are shown in Fig.1(a). Furthermore, in order to check robustness with respect to little variations, the model selection and inference are performed on a fixed number *nRnd* of binomial randomizations of the original data.

Therefore, a set of $S_{\%}$'s is collected among all the samples, and a minimum threshold for this quantity is identified to filter out under-representative observations. Without any ground-truth available, we used an empirical approach in which the shape of the $S_{\%}$'s empirical cumulative distribution (eCDF) is analyzed, and we selected as threshold (if existing) the main concave/convex inflection point over the eCDF curve[3], which could be interpreted as a signal of multimodal distribution. Then, a Generalized Pareto Distribution (GPD) is fitted on the excesses above that threshold (POT method[13]) and a QQ-plot between the empirical and GPD quantiles could provide feedbacks on the quality of the fitting such that an higher quality corresponds to a better overlap of the curve to the main diagonal. This method was applied in our case study and the results are shown in Fig.1(b,c).

Another interesting question to deal with is the saturation problem: finding the total abundance Ab_{tot} needed to reach a certain percentage level p of richness, say the 90%. For this reason, the Hurlbert-Heck (HH) rarefaction curve[1, 2] $E(S)$ is calculated, over a properly chosen grid of rarefaction levels of total abundance, as a pointwise estimation of the expected richness associated with each rarefaction level. Then a family of generalized nonlinear models defined by a log-linear mixture regression function $E(Y) = g^{-1}(\alpha log(X) + \beta X)$ and binomial distribution for the response variable with *probit*, *logit*, *cloglog* as candidate link functions g are applied on the percentage ratio $S^{\%} = E(S)/\hat{S}_{tot}$ over Ab_{tot}, and the one with the maximum R^2 index is chosen. Also, in order to calculate the total abundance $Ab^{p\%}_{tot}$ associated with a certain percentage level p of richness, the regression function was inverted via the numerical resolution of the Lambert function $W(z)e^{W(z)}, z \in \mathbb{C}$. In Fig.1(d) a graphical representation of the percentage HH curve and its regression estimator are shown for one sample of the case study.

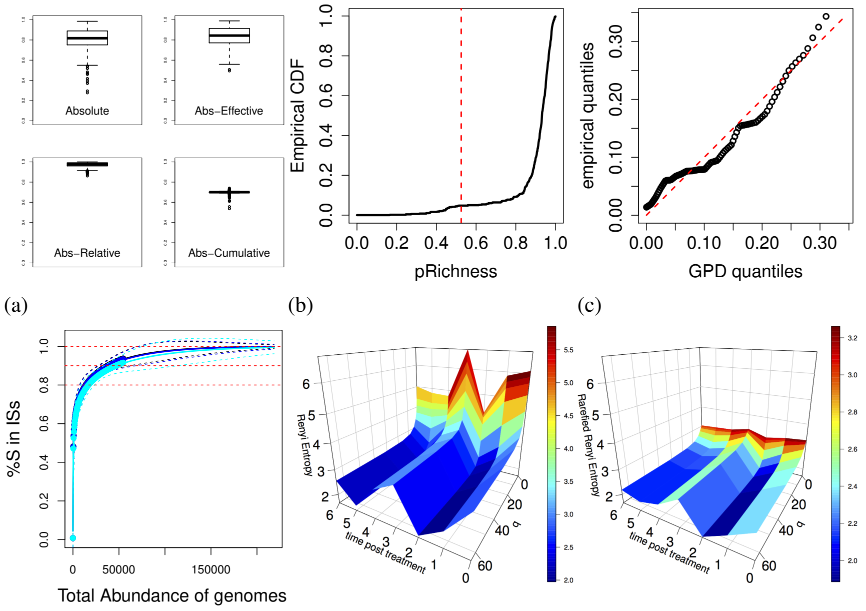
Scaling of the Heterogeneous Data

Another problem regarding the IS_s data being analyzed in the case study concerns the reliability of data interpretation due to the different orders of magnitude in total abundance Ab_{tot}, and indeed in richness S_{obs}, of IS_s reached in each sample mainly due to the variation in resolution of the gauge instruments adopted during time.

Therefore, in order to compare any measure of safety obtained in each sample (e.g. an entropy index), the whole cohort of data should be first rescaled to the same magnitude level of total abundance. In this work, the minimum total abundance level $Ab_{raremax} = \min_{samples} Ab_{tot}$ among the samples is used as rarefaction level.

[3] The eCDF inflection points are found via the Extremum Distance Estimator (EDE) algorithm [12].

Then, a rarefaction technique[2], which essentially consists in random subsampling with proportional abundances $p_i = Ab_i/Ab_{tot}$ as probability weights with Ab_i the abundance of the i-th species in the original sample, is applied in order to generate a rarefied version of that sample. This results in a more homogeneous pool of samples which can be used for entropy measures comparison during time of therapy. As an example, the Renyi Entropy Spectre (RES)[7] is calculated on the IS_s data of the case study, before and after rarefaction. The results are shown in Fig.1(e,f).



(a) (b) (c)

(d) (e) (f)

Fig. 1: (**a**) From top-left, the boxplot of the Absolute, Effective, Relative and Cumulative accuracy indexes are shown respectively. (**b**) The empirical CDF associated with the collected sample of $S^{\%}$. The vertical red highlighted line represents the threshold $S^{\%}_{thr}$ estimated via the EDE algorithm. (**c**) The scatterplot between the empirical quantiles associated with the excesses below that threshold and the GPD quantiles fitted on the same quantities. (**d**) The Hurlbert-Heck curve rescaled by the sampling pooling species estimators $S^{tot}_{Preston}$, S^{tot}_{Chao}, S^{tot}_{ACE} are drawn in black, blue and cyan. The observed and predicted ratio $S^{\%} = S_{obs}/S^{tot}$ are respectively represented by the thick and thin lines. The 80%, 90% and 100% richness thresholds are also shown as red dotted lines. This figure is related to a single sample. (**e**) The Renyi Entropy Spectre is shown during time of therapy on the heterogeneous (un-rarefied) and (**f**) homogeneous (rarefied) samples respectively.

Discussion and Conclusion

As a result of the filtering method, some of the correlations expected to be biologically relevant (e.g. between DNA nanograms DNA_{ng} and total abundance Ab_{tot} in IS_s) slightly increased, suggesting to further characterize the discarded samples.

Moreover, the comparison between the non-rarefied and rarefied entropy curves shows the positive effect in data harmonization by reducing the fluctuations in results due to change in sampling technology. These findings shed lights on reliability and reproducibility in continuous data integration over improvements in technological changes, critical challenges in the era of big data and improvements in high-throughput technologies. Species diversity could also be better addressed using the Renyi Entropy Spectre, where the effect of most abundant clones is visible at higher levels of q.

In conclusion, this work provided new methods to address the data integration and rescaling from technological sources continuously evolving and the problem of filtering unreliable data. Both problems approach the reproducibility of results in science even over time, and data accuracy and reliability.

References

1. S. H. Hurlbert, "The Nonconcept of Species Diversity: A Critique and Alternative Parameters," *Ecology*, vol. 52, no. 4, pp. 577–586, 1971.
2. K. L. Heck, G. van Belle, and D. Simberloff, "Explicit Calculation of the Rarefaction Diversity Measurement and the Determination of Sufficient Sample Size," *Ecology*, vol. 56, no. 6, pp. 1459–1461, 1975.
3. F. W. Preston, "The Commonness, And Rarity, of Species," *Ecology*, vol. 29, no. 3, pp. 254–283, 1948.
4. A. Chao, "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability," *Biometrics*, vol. 43, no. 4, p. 783, 1987.
5. R. B. O'Hara, "Species richness estimators: How many species can dance on the head of a pin?," *Journal of Animal Ecology*, vol. 74, no. 2, pp. 375–386, 2005.
6. T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*. 2006.
7. J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. No. XIV, 2010.
8. R. H. Whittaker, "Dominance and diversity in land plant communities," *Science*, vol. 147, no. 3655, pp. 250–260, 1965.
9. J. B. Wilson, "Methods for fitting dominance / diversity curves," *Journal of Vegetation Science*, vol. 2, no. 1, pp. 35–46, 1991.
10. K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," 2004.
11. A. E. Magurran and B. J. McGill, "Biological diversity: frontiers in measurement and assessment," *Challenges*, p. 368, 2011.
12. D. T. Christopoulos, "Developing methods for identifying the inflection point of a convex/concave curve," pp. 1–29, 2012.
13. G. Salvadori, C. De Michele, N. T. Kottegoda, and R. Rosso, *Extremes in Nature: An approach using Copulas*. 2005.

Monitoring event attendance using a combination of traditional and advanced surveying tools

Monitoraggio della partecipazione agli eventi tramite la combinazione di strumenti tradizionali e nuove tecnologie

Mauro Ferrante, Amit Birenboim, Anna Maria Milito, Stefano De Cantis

Abstract This paper will describe the research stages and tools used for monitoring participants' attendance at the European Researchers' Night, held in Palermo in September 2017. A combination of traditional survey instruments and new technologies was effected in order to analyse participants' behaviour during the event. The results derived from these different data sources were also integrated and analysed in order to evaluate the success of the event from social and economic points of view. Data relating to participants' mobility during the event will be described and clusters of participants proposed, based on their mobility behaviour.

Abstract *Il presente articolo descrive le fasi della ricerca e gli strumenti utilizzati per lo studio del comportamento dei visitatori alla Notte Europea dei Ricercatori, svoltasi a Palermo nel Settembre 2017. Viene proposta l'integrazione di strumenti di rilevazione tradizionali con l'utilizzo di nuove tecnologie per lo studio del comportamento dei visitatori durante l'evento. I dati derivanti dall'integrazione di più fonti informative sono stati analizzati al fine di valutare il successo dell'iniziativa dal punto di vista economico e sociale. Vengono inoltre descritti i dati relativi alla mobilità dei vistatori e vengono proposti dei cluster di partecipanti sulla base del loro comportamento in termini di mobilità.*

Key words: Event behaviour, GPS technologies, mobility, event management

Mauro Ferrante
Department of Culture and Society – University of Palermo e-mail: mauro.ferrante@unipa.it

Amit Birenboim
Department of Geography and the Human Environment, Tel Aviv University, Tel Aviv, Israel e-mail: abirenboim@tauex.tau.ac.il

Anna Maria Milito
Department of Culture and Society – University of Palermo e-mail: annamaria.milito@unipa.it

Stefano De Cantis
Department of Economics, Business and Statistics – University of Palermo e-mail: stefano.decantis@unipa.it

1 Introduction

The monitoring and measurement of participant attendance at events is particularly relevant for a number of reasons. Beyond providing information to organisers and sponsors regarding the popularity of an event, it is a fundamental prerequisite for measuring the economic, social and environmental impacts of events [1]. Given these premises, the aim of this work is to describe the various research stages, the survey instruments used, and the main results of a survey performed during the European Researchers' Night, held in Palermo in September 2017. An integration of traditional survey instruments and new technologies was implemented, the aim of which was to collect information regarding the event attendance. Data relating to the participants' profile and their behaviour during the event were collected and analysed in order to characterize participants' behaviour, motivation, and satisfaction during the event. The results are of relevance for both sponsors and organisers alike for planning and management similar initiatives in the future.

2 Data and Methods

On the occasion of the second edition of the *Sharper* European Researchers' Night project, a research team from the University of Palermo planned a survey during the event, the aim of which was to monitor and evaluate participants attendance, behaviour and satisfaction at the event. This event is supported by the European Commission as a part of the Marie Skłodowska- Curie actions and funded by the Horizon 2020 programme, a stated aim of which is to boost the careers of researchers. The survey made use of an integration of several survey tools, including: i) an infra-red beam counter; ii) a questionnaire distributed at the beginning of the event (the *opening questionnaire*); iii) a smartphone App 'Sensometer'; iv) GPS devices and v) a questionnaire distributed at the close of the event (the *closing questionnaire*).

There was no charge to enter the event. In order to measure the number of participants attending the event, an infra-red beam counter was placed in the main entrance of the Botanical Garden, the main venue, to count the total number of those entering and exiting during the event. This information were used to monitor the smooth running of the event and to obtain ex-post information useful for sampling purposes. The opening questionnaire was administered to a sample of participants and aimed at collecting information regarding participants' profile, motivation and expectations. This sample was selected according to a pseudo-systematic procedure, with a sampling interval of 1 every 20 participants joining the event. It was requested that every sampled participant download and use a smartphone app 'Sensometer'. The app collected information on participants' mobility during the event and it enabled pictures of the most 'liked' aspects of the event to be sent to the research team. A sub-sample of participants also received a GPS device which collected more accurate information on participants' mobility during the event. In order to easily recognize each sampled unit at the end of the visit, a numbered badge was provided to each participant. At the end of the visit, every sampled participant was asked to return the GPS device (where given), and to answer a closing questionnaire. The aim of the latter was to: collect post-visit information relating to satisfaction of the visit, the intention to revisit, the intention to recommend the event to friends and relatives, and a question relating to the willingness to pay for similar events in the future. Finally, the impact of the event on the participants' opinion of the University of Palermo was evaluated by means of pre-visit and a post-visit questions. In order to analyse the results, the data obtained from the various survey instruments used (two questionnaires, the smartphone app, and GPS devices) and data quality evaluation were merged. The GPS trajectories of participants were compared and participant clusters were created which were based on the similarities of their trajectories. A Dynamic Time Warping (DTW) algorithm was used to measure the distance among trajectories. This is an algorithm for measuring distance between two temporal sequences, and it was applied to obtain a distance between GPS tracking [4]. The major advantage of DTW over Euclidean distance is its ability to take into account the stretching and compression of sequences. As a result, this method produces a distance-like metric, which is independent of the velocity of the two temporal sequences. The DTW algorithm remaps the time indexes of two series X and Y to produce series of the same length T, which are termed warping curves $\phi(k) = (\phi_x(k), \phi_y(k))$ with $k = (1, \dots, T)$. The monotonicity of the warping curve is usually imposed as a constraint in order to preserve time ordering. The DTW distance can, therefore, be defined as the minimum Euclidean distance between each warping curve:

$$DTW = min_{\phi} \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k))$$

However, when comparing alignments between time series of different lengths, it is usually appropriate to manage an average per-step distance along the warping curve. Therefore, the DTW distance is divided by the number of steps of the warping curve T [3]. Having obtained the distance matrix among the trajectories, in order to segment participants in relation to their behaviour, an average linkage hierarchical clustering was implemented [2]. More peculiarly, an agglomerative approach was used by initially assigning each observation to its own cluster, then by computing the similarity between each cluster and recursively joining the two most similar.

3 Results

The fact that there was one entrance to the event (Fig. 1a) facilitated the use of an infra-red beam counter, which counted the number of participants attending the event: a total of 1,815 entrances and 1,819 exits were recorded throughout the period 18:30 to 00:00. Regrettably, the device malfunctioned for 16.4 minutes, representing 5% of the total time recorded (332 minutes). By considering the difference between numbers of exits and entrances, it is possible to determine the number of people attending the event over time. The number of people attending the event in a 10 minute time interval is reported in Fig. 1b. A peak of attendance was observed at around 22:00, and, at 23:00 (the time when the event officially closes), more than 400 people were still present until midnight when the event closed its doors. Table 1 reports the descriptive statistics for the main variables, which were collected from the opening and closing questionnaires. The majority of participants were young, well-educated and residing in Palermo. Approximately 20% of participants interviewed were returnees with a relatively high share of participants (62%) agreeing to download the mobile app. The majority of participants expressed a high degree of satisfaction with the event, many of whom felt they could recommend the event to friends and relatives. A positive impact of the event on the opinion of the University of Palermo was observed, with 44.1% of participants interviewed declaring an improvement in their opinion after the event.

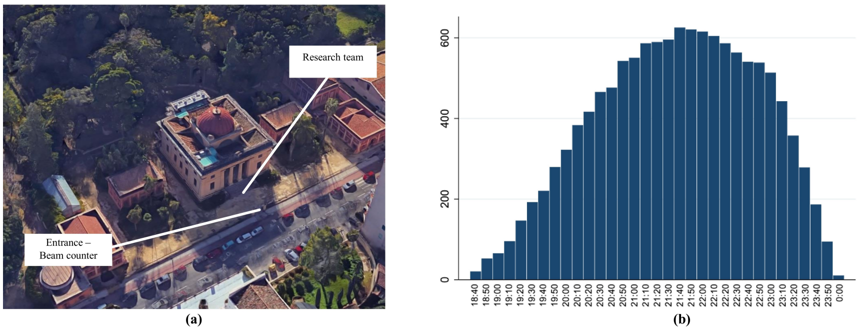


Fig. 1 Event site and survey setting (a), and number of people attending the event at 10 minutes time interval (b)

In terms of participants' behaviour at the event, a comparison of the results obtained from the mobile app and GPS devices demonstrated clearly better performances for the latter compared to the former. Indeed, the low quality of the data collected with smartphone app (missed observations, many observations with the same coordinates, irregular data collection time interval, etc.) led the authors of this paper only to analyse GPS tracking, albeit from a limited number of participants (approximately 20). After the pre-processing of GPS tracking data, the degree of similarities between each pair of trajectories was measured with the DTW algorithm. The implementation of hierarchical clustering and visual examinations of clusters of trajectories, led us to choose four clusters. Patterns of participants' mobility belonging to each cluster are reported in Fig. 2, in which the height of each polygon is proportional to the average time spent in each cell, and the colour is related to the number of participants in each cell. Differences among the identified mobility cluster patterns may be highlighted. Cluster 4 comprises those participants who had did not remain for long at the event. They made short stops at the various event sites, with little or no deviation from the circular area around the main building in the Botanical Garden. These participants spent about 1 hour at the event, and their trajectories were on average 1.7 kilometres length. Participants belonging to cluster 3 remained the longest at the event (about 2 hours) and walked the most (2.7 kilometres). The height of these polygons highlights the many stops made at several sites in the event. Similar considerations can be made for clusters 1 and 2, although their trajectories differ among each other.

Table 1 Opening and ending questionnaires: descriptive statistics (n=101).

Variable	**Categories**	%	**Variable**	**Categories**	%
Gender	Male	32.0	Took part to last year's night?	Yes	20.8
	Female	68.0		No	79.2
Place of Residence	Palermo	81.2	Willingness to use mobile app	Yes	62.0
	Other	18.8		No	38.0
Age	18-25	30.7	Company type	Partner/spouse	37.6
	26-35	20.8		Parents	3.0
	36-45	25.7		Children	25.7
	46 or more	22.8		Friends/relatives	54.5
Job type	Student	36.6	Opinion about the University before the event	Very good	8.9
	School teacher	14.9		Good	40.6
	University	4.0		Neither	44.6
	Freelance	24.8		Bad	3.0
	Other	19.8		Very bad	3.0
Main motivation	Visit to friends/relatives	19.8	Degree of satisfaction	Low	5.4
	Interest for the event	71.3		Medium	38.7
	Other	8.9		High	55.9
Education	Mid-school	5.0	Opinion about the University after the event	Improved	44.1
	High-school	39.6		Remained equal	53.8
	Bachelor	47.5		Got worse	2.2
	Master or Ph.D.	7.9	Willingness to recommend	Yes	97.8
				Don't know	2.2

4 Conclusion

Events are often used strategically to contribute to the economic development of a region. The monitoring of events is particularly useful when evaluating their cultural, social, economic and environmental impact. Furthermore, it is helpful for event managers to obtain information relating to where, when and how activities take place, and the degree of satisfaction of these experiences. The monitoring of participants' space-time behaviour can, therefore, be said to be critical to supply management. Consequently, the development and implementation of new data collection tools and methods is particularly relevant for the sustainable management

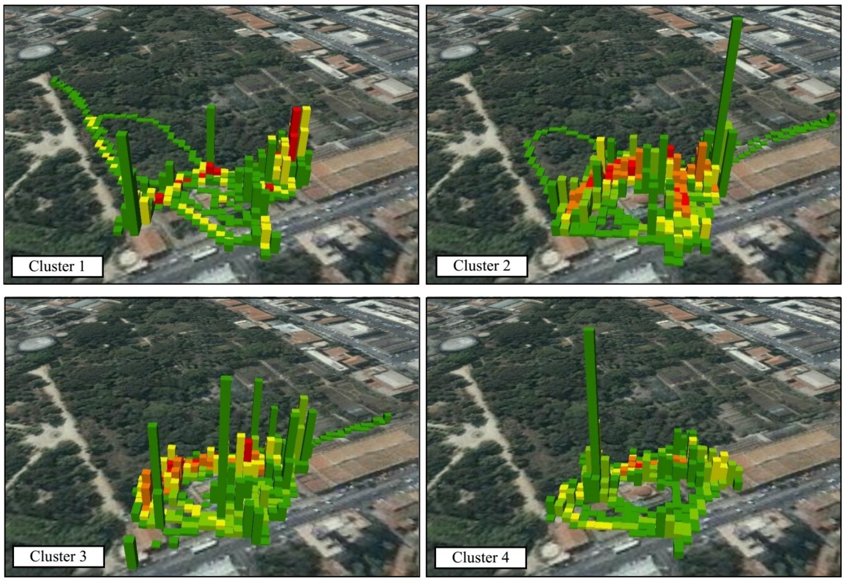


Fig. 2 Patterns of trajectories of clusters of participants

of events. The empirical application presented in this paper highlights the potential deriving from new data collection tools (e.g. GPS devices, electronic counters, etc.) for monitoring and analysing event attendance. The single entry-exit point allowed for the implementation of the presented survey strategy, a discussion on potential solutions in open events, in terms of sampling strategies and counting procedures, represents a clear research line. Moreover, an analysis of participants' characteristics as potential determinants of their behaviour at events represents an important topic for future research. An improved knowledge of participants' behaviour at events may be useful for the implementation of event management policy.

References

1. Davies, L., Coleman, R., Ramchandani, G.: Measuring attendance: issues and implications for estimating the impact of free-to-view sports events. International Journal of Sports Marketing and Sponsorship, 12(1), 6-18 (2010)
2. Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863-14868 (1998)
3. Giorgino, T.: Computing and visualizing dynamic time warping alignments in R: the dtw package. Journal of statistical Software, 31(7), 1-24 (2009)
4. Johnson, D., Trivedi, M.M.: Driving style recognition using a smartphone as a sensor platform. Intelligent Transportation Systems (ITSC), 14th International IEEE Conference on IEEE, pp. 1609-1615 (2011)

Indefinite Topological Kernels

Kernel Topologici Non Definiti

Tullia Padellini and Pierpaolo Brutti

Abstract Topological Data Analysis (TDA) is a recent and growing branch of statistics devoted to the study of the shape of the data. Motivated by the complexity of the object summarizing the topology of data, we introduce a new topological kernel that allows to extend the TDA toolbox to supervised learning. Exploiting the geodesic structure of the space of Persistence Diagrams, we define a geodesic kernel for Persistence Diagrams, we characterize it, and we show with an application that, despite not being positive semi–definite, it can be successfully used in regression tasks.
Abstract Topological Data Analysis (TDA) è una branca della statistica volta allo studio della "forma" dei dati. Data la complessità delle summaries topologiche, introduciamo una nuova famiglia di kernels per estendere TDA a problemi di apprendimento supervisionato. Sfruttando la geodesica dello spazio delle summaries topologiche, in questo lavoro definiamo un kernel geodesico, lo caratterizziamo e mostriamo con un applicazione le sue performance in un problema di regressione.

Key words: Topological Data Analysis, Kernel Methods, Indefinite Kernels

1 Introduction to Topological Data Analysis (TDA)

Topological Data Analysis TDA is a new area of research in statistics consisting of techniques aimed at recovering the topological structure of data [1, 3]. The interest in the topological structure of data stems from the immediate interpretation of the characterization provided by topological invariants: 0-dimensional features represent connected components, 1-dimensional features represent loops, 2-dimensional are voids and so on. These are all quantity of interest in statistical analysis, as for

Tullia Padellini, Pierpaolo Brutti
Sapienza Universit di Roma, Piazzale Aldo Moro, 5, 00185 Roma,
e-mail: {tullia.padellini, pierpaolo brutti}@uniroma1.it

example connected components correspond to clusters [2] and loops represent periodic structures [7].

If data come in the form of point–cloud $\mathbb{X}_n = \{X_1, \dots, X_n\}$, however, it is not possible to compute such invariants directly. A point–cloud $\mathbb{X}_n$, in fact, has a trivial topological structure, as it is composed of as many connected components as there are observations and no higher dimensional features. `TDA` provides a framework for estimating the topological structure of $\mathbb{X}_n$ by enriching it, encoding data into a filtration. The most common way to do so is to grow each observation point X_i into a ball

$$B(X_i, \varepsilon) = \{x \,|\, d_{\mathbb{X}}(x, X_i) \leq \varepsilon\},$$

of fixed radius ε.

As the resolution ε changes, the topological structure of the *cover*

$$\mathbb{X}_n^{\varepsilon} = \bigcup_{i=1}^{n} B(X_i, \varepsilon),$$

changes as well. When ε is very small, $\mathbb{X}_n^{\varepsilon}$ is topologically equivalent to $\mathbb{X}_n$; as ε grows, however, balls of the cover start to intersect, "giving birth" to loops, voids and other topologically interesting structures. At some point, when connected components merge, loops are filled and so on, these structures start to "die". Eventually when ε reaches the diameter of the data $\mathbb{X}_n$, $\mathbb{X}_n^{\varepsilon}$ is topologically equivalent to a point ball, and again retains no information. The "lifetime" of the generic i-th feature can be represented by a "birth–time" b_i representing the first value ε for which the i-th feature appears in the data, and a "death–time" d_i corresponding to when the feature disappear, i.e. the first value ε for which $\mathbb{X}_n^{\varepsilon}$ does not retain the i-th feature anymore. Birth and death times for all the features in the cover are then summarized in the *Persistence Diagram* $D = \{(b_i, d_i), i = 1, \dots m\}$.

Points that are close to the diagonal have a "short life", in the sense that the features they represent appear and disappear almost immediately and hence may be neglected; on the other hand the "longer" a feature lives, the more important it is in characterizing the structure of $\mathbb{X}_n$.

The space of Persistence Diagram $\mathscr{D}$ is a metric space when endowed with the Wasserstein distance.

Definition 1 (Wasserstein Distance between Persistence Diagrams). Given a metric d, called *ground distance*, the Wasserstein distance between two persistence diagrams D and D' is defined as

$$W_{d,p}(D, D') = \left[\inf_{\gamma} \sum_{x \in D} d\big(x, \gamma(x)\big)^p\right]^{\frac{1}{p}},$$

where the infimum is taken over all bijections $\gamma : D \mapsto D'$.

Depending on the choice of the ground distance d, Definition 1 defines a family of metrics; we focus on the L^2–norm, especially in the case $p = 2$, for which [10] proved that $W_{L^2,2}$ is a geodesic on the space of persistence diagrams.

1.1 Geodesic Topological Kernels

As most statistical learning tools are defined for inner product spaces, the metric structure of the space of persistence diagrams may be limiting, we thus turn to a kernel approach. Roughly speaking a kernel K on a space $\mathscr{M}$ is a symmetric binary function $K : \mathscr{M} \times \mathscr{M} \mapsto \mathbb{R}^+$ which represent a measure of similarity between two elements of $\mathscr{M}$. As every kernel is associated to an inner product space [9], we can use them to implicitely define an inner product space in which it is possible to perform most statistical tasks, from classification to testing, through regression.

Previous attempts in this direction (such as [8]) built kernels on persistence diagrams by considering each point of the diagram individually, thus loosing the structure of the object. In order to consider the diagram as a whole, we propose a kernel which, being based on the Wasserstein distance, preserves information about how points in the diagram are related to each others.

One popular family of kernels for a geodesic metric space $(\mathbb{X}, d)$ is the *exponential kernel*

$$k(x,y) = \exp\left\{d(x,y)^p/h\right\} \qquad p,h > 0$$

where $h > 0$ is the bandwidth parameter; for $p = 1$ this is the Laplacian kernel and for $p = 2$ this is the Gaussian kernel. It is straightforward to use this class to define a *Topological kernel* to be used for statistical learning.

Definition 2 (Geodesic Topological Kernel). Let $\mathscr{D}$ be the space of persistence diagrams, and let $h > 0$, then the Geodesic Gaussian Topological (GGT) kernel $K_{\text{GG}} : \mathscr{D} \times \mathscr{D} \mapsto \mathbb{R}^+$ is defined as

$$K_{\text{GG}}(D,D') = \exp\left\{\frac{1}{h} W_{L^2,2}(D,D')^2\right\} \qquad \forall D,D' \in \mathscr{D}.$$

Analogously, the Geodesic Laplacian Topological Kernel (GLT), K_{GL} is defined as:

$$K_{\text{GL}}(D,D') = \exp\left\{\frac{1}{h} W_{L^2,2}(D,D')\right\} \qquad \forall D,D' \in \mathscr{D}.$$

As opposed to their euclidean counterparts, the Geodesic Laplacian and Gaussian kernels are not necessarily positive definite; as shown in [4], in fact, a Geodesic Gaussian kernel on a metric space is positive definite only if the space is flat, but this is not the case for the space of Persistence Diagram, which has been proved to be curved [10].

2 Application - Fullerenes

Buckyballs fullerenes are spherical pure carbon molecules whose main trait is that atoms' linkage can form either pentagons or hexagons. We will show that our topological kernel can be exploited to predict the Total Strain Energy (measured in Ev) of a molecule from the shape of the molecule, as our Topological Kernel allows us to use Persistence Diagrams as covariates. given a sample $\{X_1,\ldots,X_n\}$ of Fullerenes we model their Total Strain Energy, Y as a function of their Persistence Diagrams $\{D_1,\ldots,D_n\}$:

$$Y_i = m(D_i) + \varepsilon_i$$

where ε_i is the usual 0–mean random error. As in standard nonparametric regression, we can estimate the regression function $m(\cdot)$ with the Nadaraya–Watson estimator[5], which does not require a positive definite kernel.

Moreover, in order to take into account the group structure naturally induced by the isomers, we also considered a model with a fixed group intercept, i.e:

$$Y_{ij} = \alpha_j + m(D_{ij}) + \varepsilon_{ij},$$

where D_{ij} denotes the persistence diagram of the $i^{\mathtt{th}}$ isomer of the $j^{\mathtt{th}}$ molecule. We fit the resulting partially linear model using Robinson's trimmed estimator, as detailed in [6]. We fit the models using data from $n = 535$ molecules of 10 different types of Fullerenes. For each molecule, the data (freely available at http://www.nanotube.msu.edu/fullerene/fullerene-isomers.html consists of the coordinates of the atoms taken from Yoshida's Fullerene Library and then re–optimized with a Dreiding–like forcefield. We focus on features of dimension 1, which recover the structure of the molecule; as we can see from Figure 1, loops in the diagrams are, in fact, clearly clustered around two centers, which represent the pentagons and the hexagons formed by the carbon atoms. Interestingly enough, the Wasserstein distance and, hence, both the geodesic kernels, fully recover the class structure induced by the isomers, as we can see in Figure 3.

	Geodesic Gaussian Kernel	Geodesic Laplacian Kernel
Nonparametric regression	339.89	342.14
Semiparametric regression	1049.02	331.04

Table 1 Residual Sum of Squares.

After choosing the bandwidth h via Leave–One–Out cross validation, we compare the different models in terms of Residual Sum of Squares (RSS). As we can see from Table 1, the two kernels yield similar results when used in a fully nonparametric estimator, while the Laplacian kernel performs better when adding the group intercept to the model. This can be understood by looking at the kernel matrices (Figure 3); the Gaussian Kernel has a sharper block structure than the Laplace Kernel, which makes it better at discriminating the 10 molecule classes. However, when

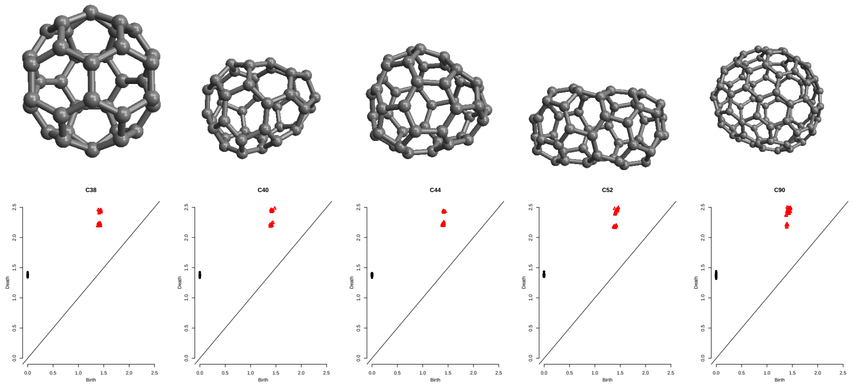


Fig. 1 Topological configurations of some fullerenes (top) and corresponding persistence diagrams (bottom). From left to right: C38(C2v), C40(C1), C44(C1), C52(C2), C90(C1).

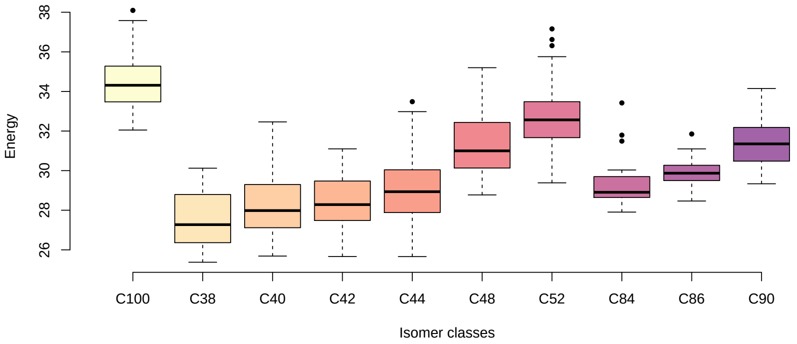


Fig. 2 Energies for the 10 different classes of isomers. It is worth noticing that Fullerenes with higher numbers of atoms do not necessarily have higher energy.

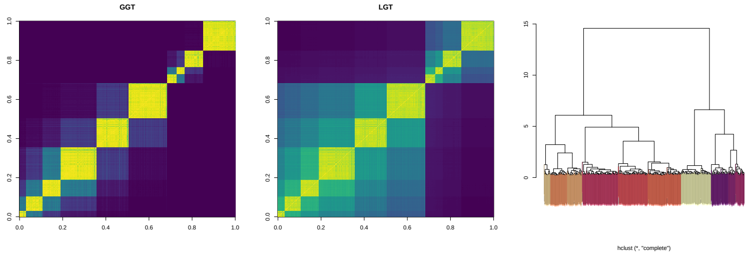


Fig. 3 Kernel Matrix for the Geodesic Gaussian Kernel (left), Geodesic Laplacian Kernel (center), Hierarchical Clustering built from the Wasserstein distance with complete linkage (right). Colors represent the different isomer classes as shown in Figure 2.

the group structure is taken into account by the model itself, this clustered structure leads to worse prediction.

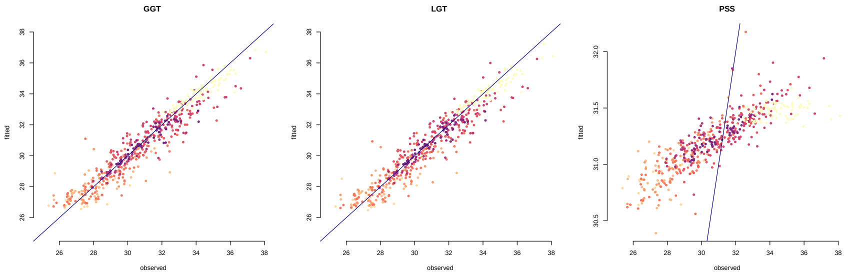


Fig. 4 Observed vs fitted plot for the fully nonparametric model fitted with the Geodesic Gaussian (left), Geodesic Laplacian (center) and the Persistence Scale Space kernel (right). Colors represent the different isomer classes as shown in Figure 2.

Finally, we compare the performance of our geodesic kernels with the Persistence Scale Space kernel K_{PSS}; as we can clearly see from the fitted-vs-observed plots in Figure 4, the positive definiteness of the PSS kernel does not result in more accurate prediction, as both K_{GG} and K_{LG} outperform it.

References

1. Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society **46**(2), 255–308 (2009)
2. Chazal, F., Guibas, L.J., Oudot, S.Y., Skraba, P.: Persistence-based clustering in riemannian manifolds. Journal of the ACM (JACM) **60**(6), 41 (2013)
3. Fasy, B.T., Kim, J., Lecci, F., Maria, C., Rouvreau, V.: Tda: Statistical tools for topological data analysis. Availabl e at https://cran. r-project. org/web/packages/TDA/index. html (2014)
4. Feragen, A., Lauze, F., Hauberg, S.: Geodesic exponential kernels: When curvature and linearity conflict. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3032–3042 (2015)
5. Härdle, W.K., Müller, M., Sperlich, S., Werwatz, A.: Nonparametric and semiparametric models. Springer Science & Business Media (2012)
6. Li, Q., Racine, J.S.: Nonparametric econometrics: theory and practice. Princeton University Press (2007)
7. Perea, J.A., Harer, J.: Sliding windows and persistence: An application of topological methods to signal analysis. Foundations of Computational Mathematics **15**(3), 799–838 (2015)
8. Reininghaus, J., Huber, S., Bauer, U., Kwitt, R.: A stable multi-scale kernel for topological machine learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4741–4748 (2015)
9. Scholkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2001)
10. Turner, K., Mileyko, Y., Mukherjee, S., Harer, J.: Fréchet means for distributions of persistence diagrams. Discrete & Computational Geometry **52**(1), 44–70 (2014)

Data Integration in Social Sciences: the earnings intergenerational mobility problem

Integrazione di dati nelle Scienze Sociali: il problema della mobilità intergenerazionale dei redditi

Veronica Ballerini, Francesco Bloise, Dario Briscolini and Michele Raitano

Abstract Merging operations on two or more datasets has become an usual need for Statistical Institutes in the last few decades. Nowadays social sciences offer the opportunity of a new application of data integration techniques. In this short paper we deal with the problem related to the estimation of the intergenerational earnings elasticity when proper datasets are not available. We compare the classical Two Samples Two Stages Least Squares with "Record Linkage" and "Matching" procedures.
Abstract *Abstract in Italian* La fusione di due o più dataset è ormai da qualche decennio una necessità per gli Istituti di Statistica. Oggi le scienze sociali offrono un'opportunità di nuova applicazione delle tecniche di integrazione di dati. In questo lavoro verrà trattato il problema legato alla stima dell'elasticità intergenerazionale del reddito quando datasets adeguati non sono disponibili. In particolare verrà confrontato l'approccio del Two Samples Two Stages Least Squares con le procedure di Record Linkage e Matching.

Key words: TSTSLS, Record Linkage, Matching

Veronica Ballerini
Sapienza Università di Roma, Via del Castro Laurenziano 9, Roma 00161,Italy
e-mail: `veronica.ballerini@uniroma1.it`

Francesco Bloise
Sapienza Università di Roma, Via del Castro Laurenziano 9, Roma 00161, Italy
e-mail: `francesco.bloise@uniroma1.it`

Dario Briscolini
Università di Padova, Via Giustiniani 2, Padova 35128, Italy
e-mail: `dario.briscolini@unipd.it`

Michele Raitano
Sapienza Università di Roma, Via del Castro Laurenziano 9, Roma 00161, Italy
e-mail: `michele.raitano@uniroma1.it`

1 Introduction

Empirical studies on economic mobility are usually intended to evaluate to what extent economic opportunities of children are associated with those of their parents, by estimating the so called intergenerational income elasticity (IGE) . This measure of mobility is the estimated coefficient of an OLS regression in which the dependent variable Y is the logarithm of permanent earnings (or incomes) of children and the regressor X is the logarithm of permanents earnings of parents. Unfortunately, data which allow to link earnings of sons and their fathers when the two generations were aged approximately the same are barely available in many countries. This the reason why scholars exploit another methodological approach which is called two-samples-two-stages least squares (TSTSLS) (see Björklund and Jäntti [1]). According to this methodology, it is possible to estimate IGEs by exploiting a sample of sons who report some socio-economic characteristics of their fathers and an independent sample of pseudo-fathers i.e. generic individuals taken in the pasts with approximately the same age of actual fathers. The problem of TSTSLS is that it may be biased. In this short paper we propose two different approaches to this issue, “Record Linkage” and Matching”. The paper is organized as follows. The second section describes the formal statistical framework; the third reviews the TSTSLS methodology; the fourth and the fifth illustrate our alternatives to the classical approach. Finally we show an application on a real dataset.

2 The basic statistical framework

Let us suppose we have two data sets, say F_1 of size $N_1 \times (h+1)$ and F_2 of size $N_2 \times (h+p)$. Records in each data set consist of several variables which may be observed together with a potential amount of measurement error. Let us denote the observed variables in F_1 by $(Y, Z_{*1}, Z_{*2}, \ldots, Z_{*h})$, whereas the observed variables in F_2 are $(X_1, X_2, \ldots, X_p, Z_{*1}, Z_{*2}, \ldots, Z_{*h})$. Also suppose that one is interested in performing a linear regression analysis of Y on $X_1, X_2, \ldots, X_p$ restricted to those pairs of records which are declared matches after a record linkage or a matching analysis based on variables $(Z_{*1}, \ldots, Z_{*h})$. The goal of the “merging procedures” is to detect all the pairs of units (j, j'), with $j \in F_1$ and $j' \in F_2$, such that j and j' actually refer to the same unit or j' is the unit in F_2 mostly similar to j .

3 The classical approach

The TSTSLS consists of two sequential steps. In the first stage, earnings of pseudo fathers are regressed on socio-economic characteristics in the main sample according to the following regression:

$$X_{j'} = \alpha + \theta_1 Z^{pf}_{j'*} + v_{j'} \tag{1}$$

where $X_{j'}$ are earnings of pseudo-father j', $Z^{pf}_{j'*}$ is a vector of the socio-economic characteristics of j', α is the intercept and $v_{j'}$ is the usual disturbance. The estimated coefficient $\hat{\theta}_1$ is then used to predict missing fathers earnings by merging the two samples according to child-reported characteristics of actual fathers. The intergenerational earnings (or income) elasticity β is thus estimated in the second stage:

$$Y_j = \alpha + \beta \hat{X}_j + \varepsilon_j \tag{2}$$

where Y_j is the logarithm of son j earnings and $\hat{X}_j = \hat{\theta}_1 Z^{f}_{j*}$ is the prediction of the logarithm of his father earnings (see Jerrim et al. [6]). The more the socio-economic characteristics perform well at predicting fathers economic status, the less estimated elasticities will be biased. More specifically, when one tries to impute fathers economic status, he is likely to make some errors in measuring their income. This reduces estimated elasticities under the assumption of classical measurement error. Moreover, if the set of socio-economic characteristics is not able to capture other characteristics of individuals which are positively correlated across generations, then the elasticity will be again downward biased.

4 Record Linkage

The issues raised in the previous section suggest the necessity of alternative strategies and perspectives. One of this possibilities is represented by direct modeling of the linkage uncertainty. In contrast to heuristic methods of data integration the introduction of a probabilistic model on the linkage step allows to evaluate the quality of the results. Moreover it may represent a more natural and realistic way to deal with the absence of joint information on the variables of interest.

Regression with linked data is well documented in Lahiri and Larsen [7]. Tancredi and Liseo [9] have proposed a Bayesian approach for Record Linkage (see also Briscolini et al. [2]). Let $\tilde{z}_{ijl}$ be the true latent value for field l of record j in data set Z_i and let $\tilde{Z}_i$ $(i = 1, 2)$ be the corresponding unobserved data matrix. Assume the "hit and miss' model by Copas and Hilton [3]:

$$p(Z_1, Z_2 | \tilde{Z}_1, \tilde{Z}_2, v) = \prod_{ijl} p(z_{ijl} | \tilde{z}_{ijl}, v_l) = \prod_{ijl} \left[v_l I(z_{ijl} = \tilde{z}_{ijl}) + (1 - v_l) \xi(z_{ijl}) \right] \tag{3}$$

The above expression is a mixture of two components: the former is degenerate at the true value while the latter can be any distribution whose support is the set of all possible values of the variable Z_{*l}, with $l = 1, 2, ..., h$.

As in Tancredi and Liseo [9], let C be a $N_1 \times N_2$ matrix whose unknown entries are either 0 or 1, where $C_{jj'} = 1$ represents a match, $C_{jj'} = 0$ denotes a non-match.

Assume that each data set does not contain replications. Also, assume that the joint distribution of $\tilde{Z}_1$ and $\tilde{Z}_2$ depends on C and on a probability vector θ which describes the distribution of the true values of Z_1 and Z_2. In detail, assume that

$$p(\tilde{Z}_1,\tilde{Z}_2|C,\theta) = \prod_{j:C_{jj'}=0,\forall j'} p(\tilde{z}_{1j}|\theta) \prod_{j':C_{jj'}=0,\forall j} p(\tilde{z}_{2j'}|\theta) \prod_{jj':C_{jj'}=1} p(\tilde{z}_{1j},\tilde{z}_{2j'}|\theta) \quad (4)$$

where $p(\tilde{z}_{ij}|\theta)$ and $p(\tilde{z}_{1j},\tilde{z}_{2j'}|\theta)$ are specific probability distributions depending on θ. The record linkage model has to be combined with the regression model in order to estimate the IGE. To complete the model, a prior distribution must be elicited for the matrix C, ν, θ and the regression parameters. The posterior distribution is not available in closed form: MCMC samples are required.

5 Statistical Matching

A valid alternative to the record linkage procedures described above is statistical matching. Although this integration technique may seem similar to record linkage, they address two different problems. On the one hand, contrarily to record linkage, matching procedures deal with observed units that are not overlapping. On the other hand, statistical matching does not take into account the possibility of measurement errors. To the aim of this work, we only focus on the so called micro approach of statistical matching (see D'Orazio [4]), i.e. the approach whose purpose is the generation of a synthetic dataset with complete information on both the variables observed only in one of the files and those observed in common. In practice, through the micro approach we are able to match records according to the related key variables. We assume that the samples belonging to the different files are generated from the same unknown distribution $f(Y,X,Z)$. In this framework, an example of statistical matching method belonging to the micro approach is the Distance Hot Deck (among others, see Rodgers [8]). In the two samples case, the first step consists in assigning to one of the files the role of recipient, i.e. the file that receives information from the so called donor. Secondly, for each record in the recipient file we measure the distance d among the h matching variables of each record in the donor; the pairs of records with the minimum distance become matches. In other words, (j,j'), $j \in F_1$,$j' \in F_2$, is a match if:

$$d_{(jj')} = \min_{j'} |z_{1jl} - z_{2j'l}|, \; j = 1,...,N_1, j' = 1,...,N_2, l = 1,...,h \quad (5)$$

This is valid in the case of continuous matching variables. Ordinal categorical might be associated to continuous variables, yet considering a proper weighting system, accordingly to the meaning and the role of the variable. Instead, distances between non ordinal may be computed assigning 1 if the value of the key variable of the recipient is equal to that of the donor, and 0 otherwise. At the end of the procedure, we obtain a complete dataset. According to the times each record in the donor file can be used as donor, we distinguish the methods of unconstrained (more than once) and constrained (only once) distance hot deck, whose main advantage is to maintain the marginal distribution of the imputed variable.

6 The application

We use the dataset AD-SILC that has been built merging the cross sectional 2005 wave of IT-SILC (the Italian component of the Eu-SILC) with the administrative records - collected by the Italian National Social Security Institute (INPS) - about working episodes and earnings of all individuals interviewed in IT-SILC 2005 since the beginning of their career. To our aims, IT-SILC 2005 includes a specific section about intergenerational mobility where many aspects of family background of the respondents are recorded. Sons, selected in the period 1970-1974, are followed since age 35 up to 39 in the period 2005-2013. Pseudo-fathers are selected in the period 1980-1988 and followed since age 40 up to 44: they were born in the period 1940-1944. Earnings of both generations are averaged. According to the notation introduced in section 2, let F_1 and F_2 be the matrices of records of the younger and the older generation respectively. The sizes of the files are $1509 \times (1+5)$ and $17245 \times (1+5)$ respectively. For each individual belonging to F_1, his full wage (Y), and a set $Z = (Z_{*1}, Z_{*2}, ..., Z_{*5})$[1] of characteristics about his father are recorded. Yet for each individual of F_2, his full wage (X) is recorded along with 5 personal characteristics - the same of Z.

$$F_1 = \begin{pmatrix} Y_1 & Z_{*1,1} & \cdots & Z_{*1,5} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1509} & Z_{*1509,1} & \cdots & Z_{*1509,5} \end{pmatrix}, F_2 = \begin{pmatrix} X_1 & Z_{*1,1} & \cdots & Z_{*1,5} \\ \vdots & \vdots & \ddots & \vdots \\ X_{17245} & Z_{*17245,1} & \cdots & Z_{*17245,5} \end{pmatrix}.$$

Z_{*1} is used as blocking variable in record linkage steps and, similarly, as donor class in matching procedure (as suggested by D'Orazio [5]). In the latter we assign to Z_{*4} a double role, given its structure: it is first a stratification variable and then it has been used to compute distances. Firstly, the ISCO-08 macro classification of occupation categories is used to create a blocking variable that assumes 9 values. Secondly, through Z_{*4} we compute distances within the macro categories. Z_{*2}, Z_{*3} and Z_{*5} are matching variables.

7 Discussion

As we expected, the estimates obtained through the standard approach are not the same as those provided by the data integration techniques (table 1). Differently from

[1] Z_{*1}: region of residence (ISTAT codification), Z_{*2}: year of birth, Z_{*3}: level of education (ISCED classification), Z_{*4}: occupation category (according to ISCO-08 classification), Z_{*5}: dummy assuming value 1 if the individual is self employed and 0 otherwise.

Table 1 Comparison of different estimators of the regression coefficient IGE

Method	**Point Estimate**	**Standard deviation**
TSTSLS	0.499	0.055
Record Linkage	0.272	0.040
Matching	0.256	0.023

the proposed methods, TSTSLS does not preserve the whole variability of X; the regression coefficient may be overestimated. Since X is only a proxy of the true fathers' income, further analyses including the measurement error seem opportune.

Certainly, record linkage allows to better account for linkage uncertainty with respect to the TSTSLS. Furthermore, it deals with measurement error in the matching variables; this is relevant since sons could have misreported some key information about their fathers. From this point of view, record linkage should be preferred also to statistical matching. On the other hand, as already said in section 5, matching procedures apply to non-overlapping samples. With this respect, statistical matching seems to be more appropriate than linkage procedures. Nevertheless, improving the information provided by the matching variables increases the efficiency of data integration techniques.

References

1. Björklund A., Jäntti M. (2011). *Intergenerational income mobility and the role of family background*. The Oxford Handbook of Economic Inequality.
2. Briscolini D., Di Consiglio L., Liseo B., Tancredi A., Tuoto T. (2017). *New methods for small area estimation with linkage uncertainty*. To appear on International Journal of Approximate Reasoning (IJAR).
3. Copas, J., Hilton, F. (1990). *Record linkage: statistical models for matching computer records*. Journal of the Royal Statistical Society, A, 153, pp. 287–320.
4. D'Orazio M., Di Zio M., Scanu M. (2004) *Statistical matching: theory and practice*, Wiley Series in Survey Methodology, John Wiley & Sons, Ltd.
5. D'Orazio M. (2017) *Statistical matching and imputation of survey data with StatMatch for the R environment*, R package vignette https:cran.r-project.orgwebpackages StatMatchvignettesStatistical_Matching_with_StatMatch.pdf
6. Jerrim, J., Choi, A. and Rodriguez Simancas, R. (2016) *Two-Sample Two-Stage Least Squares (TSTSLS) estimates of earnings mobility: how consistent are they?* Survey Research Methods, Vol. 10, No. 2, pp. 85-102
7. Lahiri P., Larsen, M.D. (2005). *Regression Analysis With Linked Data.* Journal of the American Statistical Association, 100, 222–230.
8. Rodgers W.L. (1984) *An evaluation of statistical matching*, Journal of Business and Economic Statistics 2, pp. 91102
9. Tancredi, A. and Liseo, B. (2015) *Regression Analysis with linked data: Problems and possible solutions*. Statistica, 75,1:19–35.

An innovative approach for the GDPR compliance in Big Data era

Un approccio innovativo per la conformità al regolamento GDPR nell'era dei Big Data

M. Giacalone, C. Cusatelli, F. Fanari, V. Santarcangelo, D.C. Sinitò[1]

Abstract The present work shows a preliminary overview of the Big Data Analytics scenario, introducing the related privacy issues considered by the new General Data Protection Regulation, better known by its acronym GDPR. The work then introduces an innovative index to assess the compliance of a company with this regulation on the protection of personal data, in terms of privacy by design and privacy by default.

Abstract *Il presente lavoro mostra una preventiva panoramica in merito allo scenario del Big Data Analytics, introducendo le relative problematiche di privacy considerate dal nuovo Regolamento Generale sulla Protezione dei Dati, meglio noto con l'acronimo inglese GDPR. Il lavoro introduce quindi un innovativo indice per valutare la conformità di un'azienda a tale regolamento sulla tutela dei dati personali, in termini di privacy by design e privacy by default.*

Key words: GDPR, privacy, Big Data

1 Introduction

We live the historical moment of the Big Data boom, but even more often we realize that those who use this phrase do not really know its meaning: to understand well what Big Data is, we need to understand the deep meaning of the expression and how its influence can be noticed in everyday life. It is important to start by saying that the same phrase "Big Data" is somewhat misleading as it suggests the enormous amount of data available today in different sectors and, automatically, leads to the conclusion that Big Data revolution means opportunities to have so much information available

[1] Massimiliano Giacalone, Department of Economics and Statistics, University of Naples "Federico II"; massimiliano.giacalone@unina.it
Carlo Cusatelli, Ionian Department, University of Bari "Aldo Moro"; carlo.cusatelli@uniba.it
Fabio Fanari, iInformatica S.r.l.s., ffanari@iinformatica.it
Vito Santarcangelo, Department of Mathematics and Informatics, University of Catania, santarcangelo@dmi.unict.it
Diego Carmine Sinitò, iInformatica S.r.l.s., disinito@iinformatica.it

for business. This conclusion is only partially true, because there are sectors where data, although in large quantities, are not always available to everyone and, above all, are not always shared.

If Information Technology (IT) represents for Big Data the point from which to start with the necessary tools such as cloud computing, search algorithms, etc., on the other hand Big Data are necessary and useful in the most disparate business sectors as automotive, medicine, astronomy, biology, pharmaceutical chemistry, finance, gaming, commerce.

In the public sphere, there are many other types of Big Data applications:

- the deployment of police forces where and when the offenses are more likely to occur;
- the study of associations between air quality and health;
- genomic analysis to improve the resistance to drought of rice crops;
- the creation of models to analyze data coming from living beings in the biological sciences.

2 General Data Protection Regulation

As a consequence, the need arises to regulate the use of Big Data with the help of European legislation: the EU 679/2016 General Data Protection Regulation (GDPR) was born from this need, and the aim of this work is to provide an overview of the new legislation and to introduce a new index to measure GDPR compliance (Corrales M. et al, 2017). The GDPR, approved by the European Parliament in April 2016, will enter into force on May 25, 2018. The goal is to harmonize the laws on the confidentiality of information and privacy of all European countries and keep safe the sensitive user data processed by companies, and to limit uses according to the principles of (Anisetti et al., 2018):

- lawfulness, correctness and transparency: data must be processed in such ways;
- limitation of purposes: they must be determined, explicit and legitimate, then clearly identified;
- data minimization: data must be adequate, relevant and limited;
- accuracy: the data must be updated;
- restriction of storage: data must be kept for a limited period of time to achieve the purposes;
- integrity and confidentiality: adequate security of personal data must be guaranteed.

The GDPR, replacing the regulations of the individual European countries that differ from one another, represents an important step forward in terms of standardizing European policies and data protection at the continental level (Torra V., 2017). What changes is the extension of the jurisdiction to all companies that process personal data of subjects residing in the European Union, regardless of the geographical location of the company or the place where the data are managed and processed. Non-European companies that process data of European citizens will also have to appoint an EU representative (Terry N., 2017). It is essential that European companies identify immediately how to adapt to the new legislation, thus avoiding being unprepared to face what is considered the most significant change in the history of data protection over the last 20 years.

It is necessary that companies immediately review their internal processes, placing user privacy as a primary element to guarantee priority and precedence. It is also necessary for companies to strengthen internal corporate communication through specific training programs so that anyone in a position that implies access to personal data of users correctly knows the extent to which they can carry out their profession. The concept of "privacy by design", a fundamental point on which the GDPR is concentrated (D'Acquisto, G., & Naldi, M., 2017) establishes that the data protection measures must be planned with the relative supporting IT applications starting from the planning of the business processes. This implies that only the data that are really indispensable for the performance of one's professional duties are processed and that access to information is limited only to those who have to carry out the processing. Another important point of the legislation concerns the "Breach Notification": data breach notifications are mandatory where the violation can put at risk the rights and freedoms of individuals. The notification must be made within 72 hours from the time the violation is verified and the customers are required to be informed "without undue delay".

The changes that the GDPR will bring are not only linked to the relationship between companies and users, but also concern the internal structure of the company: the new legislation will give greater prominence to the IT team and the company CIOs, making their tasks, nevertheless many managers still consider the GDPR as a waste of money and time, not understanding the importance of data protection today. (Mittal S. & Sharma P., 2017). With the GDPR, the figure of the Data Protection Officer (DPO) is established within the company with the task of monitoring the internal processes of the structure and acting as a consultant: the controllers of the monitoring and data processing activities are still required to notify their activities to local Data Protection Advisors (DPAs) which, for example within multinationals, can be a real bureaucratic nightmare, since each Member State has different notification requirements (Bertino, E., & Ferrari, E., 2018). With the introduction of the DPO, appointed on the basis of professional quality, expert in the field of law and data protection practices and equipped with the appropriate resources, the control of internal data management processes will be simplified.

The new legislation pays particular attention, in addition to what has already been said, to the requests for consent made to the subjects (Cohen M., 2017): the GDPR wants the requests to be submitted to the user in an "intelligible and easily accessible" manner, so that it is immediately clear what is the purpose of data processing. The companies will also have to guarantee users the right to delete personal data (Right to be forgotten), the possibility to request information about their treatment and to obtain a free copy in electronic format. The new regulation will be the cause of severe sanctions for companies that do not respect it, with fines of up to 4% of the total annual turnover or € 20 million, whichever is the greater of the two. But the consequences will not be only economic: failure to comply with the new rules will also have repercussions on the reputation and image of the company, which will not be considered as attentive to the privacy of users and their sensitive data.

The GDPR has shed light on the issues of Data Protection (McDermott Y., 2017), a theme that, also due to the latest cyber attacks, requires ever more attention. It is well known that the threats against IT security and data protection are not going to decrease: just think of the recent attack of the WannaCry ransomware that hit more than 150 countries between Europe and Asia causing serious damage all over the world. Such a serious attack makes us understand the skills of today's hackers, always in search of flaws and inadequacies in IT systems, which must also be protected with the help of specialists in the sector. (Beckett P., 2017)

By taking advantage of effective security solutions, companies can protect themselves completely, thus guaranteeing their users that their data is always safe and that there is no risk of it being lost.

3 Innovative approach to GDPR compliance

To this end, it is essential to set up an IT infrastructure capable of analyzing the corporate GDPR compliance by analyzing in real time the various factors that feed the system. In particular, the system must be able to consider whether the activities of privacy by design and privacy by default are actually respected in the company. By privacy by design we mean that the company in planning a new service will have to ask itself if in this new service personal data will be processed, and if these data are ordinary or particular (sensitive): in this second case it will be necessary to express how these data will be protected. It is therefore necessary, from planning, to make all assessments concerning personal data if they are processed, the so called Data Protection Impact Assessment (DPIA).

For particular data we mean:

- patrimonial data, those related to income tax returns and other taxes and duties, etc.;
- any personal data that could potentially harm the dignity of the person or affect his natural right to privacy without legal reason.

On the other hand, the principle of privacy by default establishes that by default companies should only process personal data to the extent necessary and sufficient for the intended purposes and for the period strictly necessary for such purposes. It is therefore necessary to design the data processing system, ensuring that the collected data are not excessive. However, the UE 2016/679 does not present quantitative metrics to implement compliance about privacy by design and privacy by default. Then, it is very difficult to evaluate objectively if there is compliance on an infrastructure, and it lends itself to a heuristic implementation. Our approach introduces a possible methodology to evaluate the conformance about GDPR considering a metric to calculate privacy by design and by default considering the IT infrastructure of the company.

The risk level related to the GDPR of a business can therefore be defined as the following index:

GDPR_risk = risk_by_design + risk_by_default

If for convenience we indicate the two aforementioned addends with *rbds* and *rbdf*,

rbds = n_new/n_DPIA

where: *n_new* represents the number of services / products / processes that treat personal data activated by the business under study, *n_DPIA* represents the number of impact assessments on data protection.

A value close to 1 of *rbds* shows a high risk by design value.

Considering i the i-th asset, we can define rbdf with the following approach:

$$rbdf = \sum\left(\frac{c(i)}{BC(i) + DR(i)}\right) + VASS + PTEST$$

where the risk by default is given by various contributions: the first is related to asset management and is given by the sum of the impact of the individual assets (i), considering the ratio between the complexity (c_i) of the asset (i) and the value given by the presence of Business Continuity (BC_i) and Disaster Recovery (DR_i) for the considered asset (i), each contributes a value of 1; the incidence of the single asset (i) is thus mapped: in the case of low complexity such as desktop operating computers without specific data, the incidence of the asset is 0, in the case of medium complexity (data storage server) is equal to 1, in the case of high complexity (server for economic transactions) it is 2; the value of vulnerability assessment (*VASS*) is given by 0 in the case of established security measures (e.g. IPS, IDS), and penetration testing (*PTEST*) is dimensioned on the scale from 1 to 10 according to the Common Vulnerability Scoring System (CVSS).

A value of GDPR_risk below 10 is low (high compliance), 10 to 50 is medium (intermediate compliance), over 50 is high (no compliance).

This index is the basis of the GDPR_COMPLIANCE software developed by the young Sicilian company IINFORMATICA S.R.L.S. (first and unique innovative SME company of Trapani), free of charge for educational and academic purposes, that is a very useful tool to evaluate the GDPR compliance of one’s system.

The following example represents the calculation of the *GDPR_risk* in an Italian SME with 2 locations (P1, P2) related to administration (which do not manage data locally on the machine), 1 data server (S1), 1 accounting server (S2) and 1 server for Disaster Recovery (S3). There is no backup of the 2 stations, but 1 backup of the data server and the accounting server are provided. 5 personal data treatment processes/services have been introduced with implementation of 5 DPIA. The data server and the accounting server process business data, but also special data (e.g. payroll), therefore fall within the average semantic category (average value equal to 1). There is an IPS device which then performs activities to guarantee a good Vulnerability Assessment (risk value 0), and a CVSS score equal to 1 has been found.

The value of the *risk_by_design* is given by 5/5 = 1.

The value of the *risk_by_default* is given by the following calculation:

$$rbdf = r(P1) + r(P2) + r\left(\frac{S_1}{S_3}\right) + r\left(\frac{S_2}{S_3}\right) + VASS + PTEST$$

$$rbdf = 0 + 0 + \left(\frac{1}{1}\right) + \left(\frac{1}{1}\right) + 0 + 1 = 3$$

GDPR_risk = *risk_by_design* + *risk_by_default* = 4.

The company has a low level of risk.

This index would be further reduced in the case of adoption of ISO 27001:2013 (information security management system) and ISO 22301:2012 (business continuity).

4 Conclusions

As far as data are really in unspeakable amounts, the real revolution referred to Big Data is the ability to use all this information to process, analyze and find objective evidence on different themes: the Big Data revolution refers precisely to what can be done with this amount of information, that is, to the algorithms capable of dealing with so many variables in a short time and with few computational resources. Until recently, to analyze a mountain of data that today we would call Small or Medium Data, a scientist would have taken a long time and would have used extremely expensive mainframe computers. Today, with a simple algorithm, the same information can be processed within a few hours, perhaps using a simple laptop to access the analysis platform. This presupposes new capacities to connect information to each other to provide a visual approach to data, suggesting patterns and models of interpretation so far unimaginable.

References

1. Anisetti, M., Ardagna, C., Bellandi, V., Cremonini, M., Frati, F., & Damiani, E. (2018). Privacy-aware Big Data Analytics as a Service for Public Health Policies in Smart Cities. Sustainable Cities and Society.
2. Beckett, P. (2017). GDPR compliance: your tech department's next big opportunity. Computer Fraud & Security, 2017(5), 9-13.
3. Bertino, E., & Ferrari, E. (2018). Big Data Security and Privacy. In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years (pp. 425-439). Springer, Cham.
4. Cohen, M. (2017). Fake news and manipulated data, the new GDPR, and the future of information. Business Information Review, 34(2), 81-85.
5. Corrales, M., Fenwick, M., & Forgó, N. (2017). New Technology, Big Data and the Law. Springer.
6. D'Acquisto, G., & Naldi, M. (2017). Big Data e Privacy by design (Vol. 5). G Giappichelli Editore.
7. McDermott, Y. (2017). Conceptualising the right to data protection in an era of Big Data. Big Data & Society, 4(1).
8. Mittal, S., & Sharma, P. (2017). General Data Protection Regulation (GDPR). Asian Journal of Computer Science And Information Technology, 7(4).
9. Terry, N. (2017). Existential challenges for healthcare data protection in the United States. Ethics, Medicine and Public Health, 3(1), 19-27.
10. Torra, V. (2017). Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer International Publishing.

Developments in Graphical Models

An extension of the glasso estimator to multivariate censored data

Un'estensione dello stimatore glasso per dati censurati multivariati

Antonino Abbruzzo and Luigi Augugliaro and Angelo M. Mineo

Sommario Glasso is one of the most used estimators for inferring genetic networks. Despite its diffusion, there are several fields in applied research where the limits of detection of modern measurement technologies make the use of this estimator theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. In this paper we propose an extension to censored data.
Sommario *Lo stimatore glasso è uno degli stimatori più diffusi per fare inferenza sulle reti generiche. Nonostante la sua diffusione, vi sono molti campi della ricerca applicata dove i limiti di misurazione dei moderni strumenti di misurazione rendono teoricamente infondato l'utilizzo di questo stimatore, anche quando l'assunzione sulla distribuzione gaussiana multivariata è soddisfatta. In questo lavoro, proponiamo un'estensione dello stimatore glasso ai dati censurati.*

Key words: Censored data, Gaussian graphical model, glasso estimator.

1 Introduction

An important aim in genomics is to understand interactions among genes, characterized by the regulation and synthesis of proteins under internal and external signals. These relationships can be represented by a genetic network, i.e., a graph where

Antonino Abbruzzo
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: antonino.abbruzzo@unipa.it

Luigi Augugliaro
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: luigi.augugliaro@unipa.it

Angelo M. Mineo
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: angelo.mineo@unipa.it

nodes represent genes and edges describe the interactions among them. Gaussian graphical models [3] have been widely used for reconstructing a genetic network from expression data. The reason of such diffusion relies on the statistical properties of the multivariate Gaussian distribution which allow the topological structure of a network to be related with the non-zero elements of the concentration matrix, i.e., the inverse of the covariance matrix. Thus, the problem of network inference can be recast as the problem of estimating a concentration matrix. The glasso estimator [8] is a popular method for estimating a sparse concentration matrix, based on the idea of adding an ℓ_1-penalty function to the likelihood function of the multivariate Gaussian distribution.

Despite the widespread literature on the glasso estimator, there is a great number of fields in applied research where modern measurement technologies make the use of this graphical model theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. A first example of this is Reverse Transcription quantitative Polymerase Chain Reaction (RT-qPCR), a popular technology for gene expression profiling. This technique relies on fluorescence-based detection of amplicon DNA and allows the kinetics of PCR amplification to be monitored in real time, making it possible to quantify nucleic acids with extraordinary ease and precision. The analysis of the raw RT-qPCR profiles is based on the cycle-threshold, defined as the fractional cycle number in the log-linear region of PCR amplification in which the reaction reaches fixed amounts of amplicon DNA. If a target is not expressed or the amplification step fails, the threshold is not reached after the maximum number of cycles (limit of detection) and the corresponding cycle-threshold is undetermined. For this reason, the resulting data is naturally right-censored data. In this paper we propose an extension of the glasso estimator that takes into account the censoring mechanism of the data explicitly.

2 The censored glasso estimator

Let $\boldsymbol{X} = (X_1,\ldots,X_p)^\top$ be a p-dimensional random vector. Graphical models allow to represent the set of conditional independencies among these random variables by a graph $\mathscr{G} = \{\mathscr{V},\mathscr{E}\}$, where $\mathscr{V}$ is the set of nodes associated to $\boldsymbol{X}$ and $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ is the set of ordered pairs, called edges, representing the conditional dependencies among the p random variables [3]. The Gaussian graphical model is a member of this class of models based on the assumption that $\boldsymbol{X}$ follows a multivariate Gaussian distribution with expected value $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_p)^\top$ and covariance matrix $\Sigma = (\sigma_{hk})$. Denoting with $\Theta = (\theta_{hk})$ the concentration matrix, i.e., the inverse of the covariance matrix, the density function of $\boldsymbol{X}$ can be written as follows

$$\phi(\boldsymbol{x};\boldsymbol{\mu},\Theta) = (2\pi)^{-p/2}|\Theta|^{1/2}\exp\{-1/2(\boldsymbol{x}-\boldsymbol{\mu})^\top\Theta(\boldsymbol{x}-\boldsymbol{\mu})\}. \tag{1}$$

As shown in [3], the off-diagonal elements of the concentration matrix are the parametric tools relating the pairwise Markov property to the factorization of the density

(1). Formally, two random variables, say X_h and X_k, are conditionally independent given all the remaining variables if and only if θ_{hk} is equal to zero. This result provides a simple way to relate the topological structure of the graph $\mathscr{G}$ to the pairwise Markov property, i.e., the undirected edge (h,k) is an element of the edge set $\mathscr{E}$ if and only if $\theta_{hk} \neq 0$, consequently the graph specifying the factorization of the density (1) is also called concentration graph.

Let $\boldsymbol{X}$ be a (partially) latent random vector with density function (1). In order to include the censoring mechanism inside our framework, let us denote by $\boldsymbol{l} = (l_1,\ldots,l_p)^\top$ and $\boldsymbol{u} = (u_1,\ldots,u_p)^\top$, with $l_h < u_h$ for $h = 1,\ldots,p$, the vectors of known left and right censoring values. Thus, X_h is observed only if it is inside the interval $[l_h,u_h]$ otherwise it is censored from below if $X_h < l_h$ or censored from above if $X_h > u_h$. Under this setting, a rigorous definition of the joint distribution of the observed data can be obtained using the approach for missing data with nonignorable mechanism [4]. This requires the specification of the distribution of a p-dimensional random vector, denoted by $R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u})$, used to encode the censoring patterns. Formally, the hth element of $R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u})$ is defined as $R(X_h;l_h,u_h) = I(X_h > u_h) - I(X_h < l_h)$, where $I(\cdot)$ denotes the indicator function. By construction $R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u})$ is a discrete random vector with support the set $\{-1,0,1\}^p$ and probability function $\Pr\{R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r}\} = \int_{D_{\boldsymbol{r}}} \phi(\boldsymbol{x};\boldsymbol{\mu},\Theta)d\boldsymbol{x}$, where $D_{\boldsymbol{r}} = \{\boldsymbol{x} \in \mathbb{R}^p : R(\boldsymbol{x};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r}\}$.

Given a censoring pattern, we can simplify our notation by partitioning the set $\mathscr{I} = \{1,\ldots,p\}$ into the sets $o = \{h \in \mathscr{I} : r_h = 0\}, c^- = \{h \in \mathscr{I} : r_h = -1\}$ and $c^+ = \{h \in \mathscr{I} : r_h = +1\}$ and, in the following of this paper, we shall use the convention that a vector indexed by a set of indices denotes the corresponding subvector. For example, the subvector of observed elements in $\boldsymbol{x}$ is denoted by $\boldsymbol{x}_o = (x_h)_{h\in o}$ and, consequently, the observed data is the vector $(\boldsymbol{x}_o^\top, \boldsymbol{r}^\top)^\top$. The joint probability distribution of the observed data, denoted by $\varphi(\boldsymbol{x}_o,\boldsymbol{r};\boldsymbol{\mu},\Theta)$, is obtained by integrating $\boldsymbol{X}_{c^+}$ and $\boldsymbol{X}_{c^-}$ out of the joint distribution of $\boldsymbol{X}$ and $R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u})$, which can be written as the product of the density function (1) and the conditional distribution of $R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u})$ given $\boldsymbol{X} = \boldsymbol{x}$. Formally

$$\varphi(\boldsymbol{x}_o,\boldsymbol{r};\boldsymbol{\mu},\Theta) = \int \phi(\boldsymbol{x}_o,\boldsymbol{x}_{c^-},\boldsymbol{x}_{c^+};\boldsymbol{\mu},\Theta)\Pr\{R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{x}\}d\boldsymbol{x}_{c^-}d\boldsymbol{x}_{c^+}. \quad (2)$$

Density (2) can be simplified by observing that $\Pr\{R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{x}\}$ is equal to one if the censoring pattern encoded in $\boldsymbol{r}$ is equal to the pattern observed in $\boldsymbol{x}$, otherwise it is equal to zero, i.e.,

$$\Pr\{R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u}) = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{x}\} = I(\boldsymbol{x}_{c^-} < \boldsymbol{l}_{c^-})I(\boldsymbol{l}_o \leq \boldsymbol{x}_o \leq \boldsymbol{u}_o)I(\boldsymbol{u}_{c^+} < \boldsymbol{x}_{c^+}),$$

where the inequalities in the previous expressions are intended elementwise. From this, $\varphi(\boldsymbol{x}_o,\boldsymbol{r};\boldsymbol{\mu},\Theta)$ can be rewritten as

$$\varphi(\boldsymbol{x}_o,\boldsymbol{r};\boldsymbol{\mu},\Theta) = \int_{\boldsymbol{u}_{c^+}}^{+\infty}\int_{-\infty}^{\boldsymbol{l}_{c^-}} \phi(\boldsymbol{x}_o,\boldsymbol{x}_{c^-},x_{c^+};\boldsymbol{\mu},\Theta)d\boldsymbol{x}_{c^-}d\boldsymbol{x}_{c^+}I(\boldsymbol{l}_o \leq \boldsymbol{x}_o \leq \boldsymbol{u}_o)$$

$$= \int_{D_c} \phi(\boldsymbol{x}_o, \boldsymbol{x}_c; \boldsymbol{\mu}, \Theta) d\boldsymbol{x}_c I(\boldsymbol{l}_o \leq \boldsymbol{x}_o \leq \boldsymbol{u}_o), \tag{3}$$

where $D_c = (-\infty, \boldsymbol{l}_{c^-}) \times (\boldsymbol{u}_{c^+}, +\infty)$ and $c = c^- \cup c^+$. Suppose we have a sample of size n; in order to simplify our notation the set of indices of the variables observed in the ith observation is denoted by $o_i = \{h \in \mathscr{I} : r_{ih} = 0\}$, while $c_i^- = \{h \in \mathscr{I} : r_{ih} = -1\}$ and $c_i^+ = \{h \in \mathscr{I} : r_{ih} = +1\}$ denote the sets of indices associated to the left and right-censored data, respectively. Denoting by $\boldsymbol{r}_i$ the realization of the random vector $R(\boldsymbol{X}_i; \boldsymbol{l}, \boldsymbol{u})$, the ith observed data is the vector $(\boldsymbol{x}_{io_i}^\top, \boldsymbol{r}_i^\top)^\top$. Using the density function (3), the observed log-likelihood function can be written as

$$\ell(\boldsymbol{\mu}, \Theta) = \sum_{i=1}^{n} \log \int_{D_{c_i}} \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta) d\boldsymbol{x}_{ic_i} = \sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{io_i}, \boldsymbol{r}_i; \boldsymbol{\mu}, \Theta), \tag{4}$$

where $D_{c_i} = (-\infty, \boldsymbol{l}_{c_i^-}) \times (\boldsymbol{u}_{c_i^+}, +\infty)$ and $c_i = c_i^- \cup c_i^+$. Although inference about the parameters of this model can be carried out via the maximum likelihood method, the application of this inferential procedure to real datasets is limited for three main reasons. Firstly, the number of measured variables is often larger than the sample size and this implies the non-existence of the maximum likelihood estimator even when the dataset is fully observed. Secondly, even when the sample size is large enough, the maximum likelihood estimator will exhibit a very high variance [5, 7]. Thirdly, empirical evidence suggests that gene networks or more general biochemical networks are not fully connected [2]. In terms of Gaussian graphical models this evidence translates in the assumption that Θ has a sparse structure, i.e., only few θ_{hk} are different from zero, which is not obtained by a direct (or indirect) maximization of the observed log-likelihood function (4).

All that considered, we propose to estimate the parameters of the Gaussian graphical model by generalizing the approach proposed in [8], i.e., by maximizing a new objective function defined by adding a lasso-type penalty function to the observed log-likelihood (4). The resulting estimator, called censored glasso (cglasso), is formally defined as

$$\{\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho\} = \arg\max_{\boldsymbol{\mu}, \Theta \succ 0} \frac{1}{n} \sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{io_i}, \boldsymbol{r}_i; \boldsymbol{\mu}, \Theta) - \rho \sum_{h \neq k} |\theta_{hk}|. \tag{5}$$

Like in the standard glasso estimator, the non-negative tuning parameter ρ is used to control the amount of sparsity in the estimated concentration matrix $\widehat{\Theta}^\rho = (\hat{\theta}_{hk}^\rho)$ and, consequently, in the corresponding estimated concentration graph $\widehat{\mathscr{G}}^\rho = \{\mathscr{V}, \widehat{\mathscr{E}}^\rho\}$, where $\widehat{\mathscr{E}}^\rho = \{(h,k) : \hat{\theta}_{hk}^\rho \neq 0\}$. When ρ is large enough, some $\hat{\theta}_{hk}^\rho$ are shrunken to zero resulting in the removal of the corresponding link in $\widehat{\mathscr{G}}^\rho$; on the other hand, when ρ is equal to zero and the sample size is large enough the estimator $\widehat{\Theta}^\rho$ coincides with the maximum likelihood estimator of the concentration matrix, which implies a fully connected estimated concentration graph.

Tabella 1 Results of the simulation study: for each measure used to evaluate the behaviour of the considered methods we report average values and standard deviation between parentheses

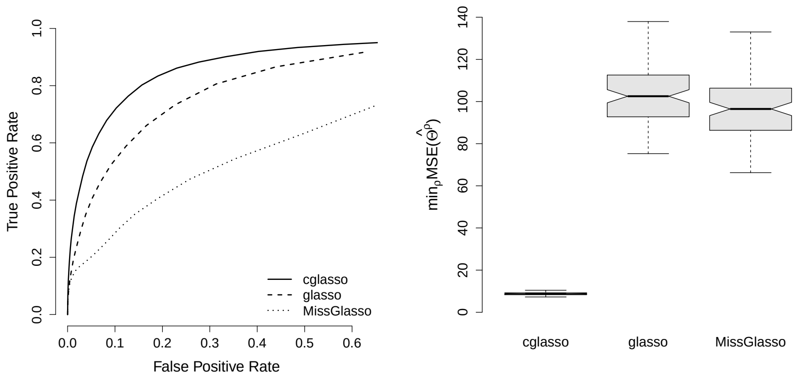
Model	$\min_\rho \text{MSE}(\hat{\boldsymbol{\mu}}^\rho)$		$\min_\rho \text{MSE}(\widehat{\Theta}^\rho)$			AUC		
H/p	cglasso	MissGlasso	cglasso	glasso	MissGlasso	cglasso	glasso	MissGlasso
0.5	0.47	14.50	8.76	103.35	96.75	0.60	0.46	0.37
	(0.11)	(0.69)	(0.64)	(14.43)	(16.01)	(0.02)	(0.02)	(0.02)
0.7	0.48	21.00	10.11	139.76	131.99	0.58	0.39	0.25
	(0.10)	(0.76)	(0.84)	(15.94)	(18.81)	(0.02)	(0.02)	(0.02)

3 Simulation study

By a simulation study, we compare our proposed estimator with MissGlasso [6], which performs ℓ_1-penalised estimation under the assumption that the censored data are missing at random, and with the glasso estimator [1], where the empirical covariance matrix is calculated by imputing the missing values with the limit of detection. These estimators are evaluated in terms of both recovering the structure of the true concentration graph and the mean squared error.

Our study is based on a multivariate Gaussian distribution with $p = 50$ and sparse concentration matrix simulated by a random structure, i.e., the probability of observing a link between two nodes is 0.05. To simulate a censored sample we use the following procedure: we set the mean $\boldsymbol{\mu}$ in such a way that $\mu_h = 40$ for the H censored variables, i.e. $\Pr\{R(X_h; -\infty, 40) = +1\} = 0.50$, while for the remaining variables μ_h is sampled from a uniform distribution on the interval $[10; 35]$. The quantity H is used to evaluate the effects of the number of censored variables on the behaviour of the considered estimators. In particular, we consider $H \in \{25, 35\}$. At this point, we simulate a sample from the latent p-variate Gaussian distribution and treat all values greater than 40 as censored. We use the previous procedure to simulate 100 samples and in each simulation we compute the coefficients path using cglasso, MissGlasso and glasso. Each path is computed using an equally spaced sequence of 30 ρ-values. Table 1 reports the summary statistics $\min_\rho \text{MSE}(\hat{\boldsymbol{\mu}}^\rho)$, $\min_\rho \text{MSE}(\widehat{\Theta}^\rho)$ and the Area Under the Curve (AUC) for network discovery.

The results on the AUC suggest that cglasso can be used as an efficient tool for recovering the structure of the true concentration matrix. The distribution of the minimum value of the mean squared errors shows that, not only our estimator is able to recover the structure of the graph but also outperforms the competitors in terms of both estimation of $\boldsymbol{\mu}$ and Θ. We did not report $\min_\rho \text{MSE}(\hat{\boldsymbol{\mu}}^\rho)$ for glasso since this method does not allow to estimate the parameter $\boldsymbol{\mu}$. Figure 1 shows a graphical representation of the results obtained with $H = 25$.



(a) (b)

Figura 1 Results of the simulation study with $H = 25$. Panel (a) shows the ROC curves; Panel (b) shows the box-plots of the behaviour of quantity $\min_\rho \mathrm{MSE}(\widehat{\Theta}^\rho)$ for the considered estimators.

4 Conclusions

In this paper, we have proposed an extension of the glasso estimator to multivariate censored data. An extensive simulation study showed that the proposed estimator overcomes the existing estimators both in terms of parameter estimation and of network recovery.

Riferimenti bibliografici

1. Friedman, J. H., Hastie T., Tibshiran, T.: Sparce inverse covariance estimation with the graphical lasso. **9**(3), 432–441 (2008)
2. Gardner T. S., di Bernardo D., Lorenz D., Collins J. J.: Inferring genetic networks and identifying compound mode of action via expression profiling. Science. **301**, 102–105 (2003)
3. Lauritzen, S. L.: Graphical Models. Oxford University Press, Oxford (1996)
4. Little, R. J. A., Rubin, D. B.: Statistical Analysis with Missing Data. John Wiley & Sons, Inc., Hoboken (2002)
5. Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology. **4**(1). (2005)
6. Städler, N., Bühlmann, P.: Missing values: sparse inverse covariance estimation and an extension to sparse regression. Stat. Comput. **22**(1), 219–235 (2012)
7. Uhler C.: Geometry of maximum likelihood estimation in Gaussian graphical models. Ann. Statist. **40**(12), 238–261 (2012)
8. Yuan M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika. **94**(1), 19–35 (2007)

Bayesian Estimation of Graphical Log-Linear Marginal Models

Stima Bayesiana di Modelli Grafici Log-Lineari Marginali

Claudia Tarantola, Ioannis Ntzoufras and Monia Lupparelli

Abstract Bayesian methods for graphical log-linear marginal models have not been developed as much as traditional frequentist approaches. The likelihood function cannot be analytically expressed in terms of the marginal log-linear interactions, but only in terms of cell counts or probabilities. No conjugate analysis is feasible, and MCMC methods are needed. We present a fully automatic and efficient MCMC strategy for quantitative learning, based on the DAG representation of the model. While the prior is expressed in terms of the marginal log-linear interactions, the proposal is on the probability parameter space. In order to obtain an efficient algorithm, we use as proposal values draws from a Gibbs sampling on the probability parameters.

Abstract *I metodi bayesiani per l'analisi di modelli grafici log-lineari marginali non sono stati sviluppati allo stesso modo di quelli frequentisti. La funzione di verosimiglianza non può essere espressa analiticamente attraverso i parametri log-lineari marginali, ma solamente in termini di frequenze o probabilità di cella. Non è possibile effettuare analisi coniugata, rendendo necessario l'utilizzo di metodi MCMC. Presentiamo una strategia MCMC per l'apprendimento quantitivo, completamente automatica ed efficiente, basata sulla rappresentazione del modello in termini di DAG. Mentre la prior è espressa in termini dei parametri marginali log-lineari, la proposal è sullo spazio delle probabilità. Al fine di ottenere un algoritmo efficiente, usiamo come proposal i valori ottenuti applicando un campionamento di Gibbs sullo spazio delle probabilità.*

Key words: DAG Representation, Marginal Log-Linear Parameterisation, Markov Chain Monte Carlo, Probability Based Sampler

Claudia Tarantola
University of Pavia, Italy, e-mail: claudia.tarantola@unipv.it

Ioannis Ntzoufras
Athens University of Economics and Business, Greece, e-mail: ntzoufras@aueb.gr

Monia Lupparelli
University of Bologna, Italy, e-mail: monia.lupparelli@unibo.it

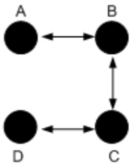
1 Introduction

Statistical models defined by imposing restrictions on marginal distributions of contingency tables have received considerable attention in economics and social sciences; for a thorough review see [2]. In particular standard log-linear models have been extended by [1] to allow the analysis of marginal distributions in contingency tables. This wider class of models is known as the class of marginal log-linear models. In these models, the log-linear interactions are estimated using the frequencies of appropriate marginal contingency tables, and are expressed in terms of marginal log-odds ratios. Following [1], the parameter vector λ of the marginal log-linear interactions can be obtained as

$$\lambda = C \log\big(M \mathrm{vec}(p)\big), \tag{1}$$

where $\mathrm{vec}(p)$ is the vector of joint probabilities, C is a contrast matrix and M specifies from which marginal each element of λ is calculated. A standard log-linear model is obtained from (1) setting M equal to the identity matrix and C to the inverse of the design matrix. Marginal log-linear models have been used by [4] to provide a parameterisation for discrete graphical models of marginal independence. A graphical log-linear marginal model is defined by zero constraints on specific log-linear interactions. It can be represented by a bi-directed graph like the one in Figure 1, where a missing edge indicates that the corresponding variables are marginally independent; for the related notation and terminology see [3] and [4] .

Fig. 1 A bi-directed graph



Despite the increasing interest in the literature for graphical log-linear marginal models, Bayesian analysis has not been developed as much as traditional methods. The main reasons are the following. Graphical log-linear marginal models belong to curved exponential families that are difficult to handle from a Bayesian perspective. Posterior distributions cannot be directly obtained, and MCMC methods are needed. The likelihood cannot be analytically expressed as a function of the marginal log-linear interactions, but only in terms of cell counts or probabilities. Hence, an iterative procedure should be implemented to calculate the cell probabilities, and consequently the model likelihood. Another important point is that, in order to have a well-defined model of marginal independence, we need to construct an algorithm which generates parameter values that lead to a joint probability distribution with compatible marginals.

A possibility is to follow the approach presented in [5], where a Gibbs sampler based on a probability parameterisation of the model is presented. Even if using this approach one can obtain as a by-product the distribution of the log-linear interactions, if the focus is on marginal log-odds a prior should be directly specified for these parameters. Additionally, and more importantly, if any prior information exists for log-odds then we need to work directly using the log-linear parameterisation. For instance, symmetry constraints, vanishing high-order associations or further prior information about the joint and marginal distributions can be easily specified by setting linear constraints on marginal log-linear interactions, instead of non-linear multiplicative constraints on the probability space. For the previous reasons, in [6] a novel MCMC strategy (the probability based sampler) is introduced. In the probability based sampler the prior is expressed in terms of the marginal log-linear interactions, while the proposal is defined on the probability parameter space. Efficient proposal values are obtained via the conditional conjugate approach of [5]. The corresponding proposal density on the marginal log-linear interactions is obtained by implementing standard theory about functions of random variables. For more details on the methodology and the obtained results, see the extended version of this work ([6]).

2 Probability Based independence Sampler

Following the notation of [5], we can divide the class of graphical log-linear marginal models in two major categories: homogeneous and non-homogeneous models. Both type of models are shown to be compatible, in terms of independencies, with a certain DAG representation (augmented DAG). Nevertheless, while homogeneous models can be generated via a DAG with the same vertex set, for non-homogeneous ones it is necessary to include some additional latent variables. The advantage of the augmented DAG representation is that the joint probability over the augmented variable space (including both observed and latent variables) can be written using the standard DAG factorisation. We parameterise the augmented DAG via a set Π of marginal and conditional probability parameters on which, following [5], we implement a conjugate analysis based on products of Dirichlet distributions. Once a suitable prior is assigned on the marginal log-linear interaction parameters, a Metropolis-Hastings algorithm can be used to obtain a sample from the posterior distribution. For $t = 1,\dots,T$, we repeat the following steps

1. propose Π' from $q(\Pi'|\Pi^{(t)})$, where $\Pi^{(t)}$ is the value of Π at t iteration;
2. from Π' calculate via marginalisation the proposed joint probabilities p' for the observed table;
3. from p', calculate λ' using (1) and then obtain the corresponding non-zero elements $\overrightarrow{\lambda}'$;
4. set $\xi' = \Pi'_{\xi}$; where Π'_{ξ} is a pre-specified subset of Π' of dimension $\dim(\Pi) - \dim(\overrightarrow{\lambda})$;
5. accept the proposed move with probability $\alpha = \min(1, A)$

$$A = \frac{f(n|\Pi')f(\overrightarrow{\lambda}')q(\Pi^{(t)}|\Pi')}{f(n|\Pi^{(t)})f(\overrightarrow{\lambda}^{(t)})q(\Pi'|\Pi^{(t)})} \times \text{abs}\left(\frac{\mathscr{J}\left(\Pi^{(t)}, \overrightarrow{\lambda}^{(t)}, \xi^{(t)}\right)}{\mathscr{J}\left(\Pi', \overrightarrow{\lambda}', \xi'\right)}\right), \qquad (2)$$

where abs$(\cdot)$ stands for the absolute value, and $\mathscr{J} = \mathscr{J}(\Pi, \overrightarrow{\lambda}, \xi)$ is the determinant of the jacobian matrix of the transformation $\Pi = g(\overrightarrow{\lambda}, \xi)$. The construction of the Jacobian matrix is facilitated by the augmented DAG representation of the model. Note that the ratio $f(\xi')/f(\xi^{(t)})$ cancels out from the acceptance rate since we set $f(\xi_t) = I_{\{0<\xi_t<1\}}$.

6. If the move is accepted, we set $\Pi^{(t+1)} = \Pi'$, $\xi^{(t+1)} = \xi'$, and $\overrightarrow{\lambda}^{(t+1)} = \overrightarrow{\lambda}'$ otherwise we set $\Pi^{(t+1)} = \Pi^{(t)}$ and $\overrightarrow{\lambda}^{(t+1)} = \overrightarrow{\lambda}^{(t)}$.

In order to obtain a high acceptance rate it is crucial the choice of the proposal density $q(\Pi'|\Pi^{(t)})$. As discussed in [6], an efficient proposal is $q(\Pi'|\Pi^{(t)}) = f_q\left(\Pi'|n^{\mathscr{A}}\right)f\left(n^{\mathscr{A}}|\Pi^{(t)}, n\right)$, where $n^{\mathscr{A}}$ is an augmented table. Exploiting the conditional conjugate approach of [5] we consider as a "prior" $f_q(\Pi)$ a product of Dirichlet distributions obtaining a conjugate "posterior" distribution $f_q(\Pi'|n'^{\mathscr{A}})$. The acceptance rate in (2) becomes equal to

$$A = \frac{f\left(n^{\mathscr{A}(t)}|\Pi'\right)f\left(\lambda'\right)f_q\left(\Pi^{(t)}|n^{\mathscr{A}(t)}\right)}{f\left(n'^{\mathscr{A}}|\Pi^{(t)}\right)f\left(\lambda^{(t)}\right)f_q\left(\Pi'|n'^{\mathscr{A}}\right)} \times \text{abs}\left(\frac{\mathscr{J}\left(\Pi^{(t)}, \lambda^{(t)}, \xi^{(t)}\right)}{\mathscr{J}\left(\Pi', \lambda', \xi'\right)}\right).$$

In the following, we will refer to this approach as the *probability-based independence sampler* (PBIS). Although PBIS simplifies the MCMC scheme, the parameter space is still considerably extended by considering the augmented frequency table $n^{\mathscr{A}}$. This algorithm can be further simplified by using as proposal a random permutation of the MCMC output obtained applying the Gibbs sampling of [5]. The acceptance rate becomes

$$A = \frac{f(\lambda')f_q(\Pi^{(t)})}{f(\lambda^{(t)})f_q(\Pi')} \times \text{abs}\left(\frac{\mathscr{J}\left(\Pi^{(t)}, \lambda^{(t)}, \xi^{(t)}\right)}{\mathscr{J}\left(\Pi', \lambda', \xi'\right)}\right).$$

This sampler is named the *prior-adjustment* algorithm (PAA) due to its characteristic to correct for the differences between the prior distributions used under the two parameterisations.

3 Simulation study

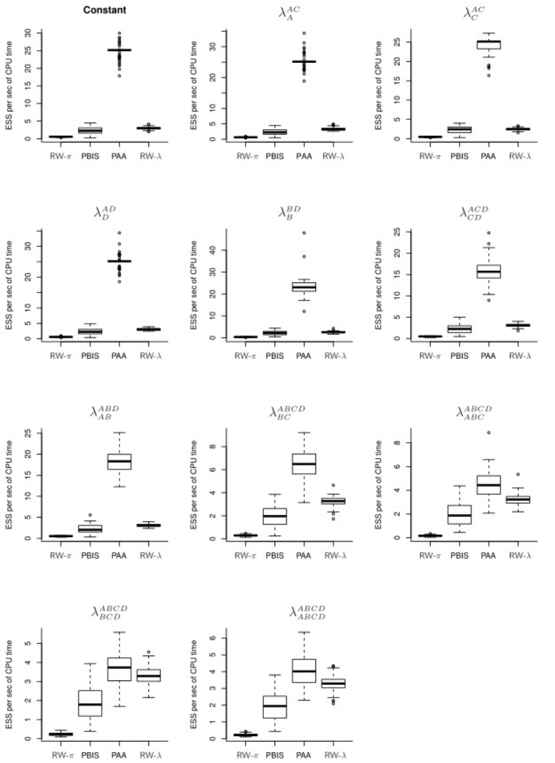
We evaluate the performance of the algorithms presented in Section 2 via a simulation study. We generated 100 samples from the marginal association model represented by the bi-directed graph of Figure 1, and true log-linear interactions given in Table 1. In addition to the algorithms described in Section 2, for comparative purposes, we consider random walks on marginal log-linear interactions λ and on logits of probability parameters π (RW-λ and RW-π respectively). We compare the examined methods in terms of Effective Sample Size (ESS) and Monte Carlo Error (MCE).

Table 1 True Effect Values Used for the Simulation Study

Marginal	Active interactions	Zero interactions
AC	$\lambda_{\emptyset}^{AC} = -1.40, \lambda_{A}^{AC}(2) = -0.15, \lambda_{C}^{AC}(2) = 0.10,$	$\lambda_{AC}^{AC} = 0$
AD	$\lambda_{B}^{AD}(2) = 0.12,$	$\lambda_{BD}^{BD}(2,2) = 0$
BD	$\lambda_{D}^{BD}(2) = -0.09,$	$\lambda_{AD}^{AD}(2,2) = 0$
ACD	$\lambda_{CD}^{ACD}(2,2) = 0.20,$	$\lambda_{ACD}^{ACD}(2,2,2) = 0$
ABD	$\lambda_{AB}^{ABD}(2,2) = -0.15,$	$\lambda_{ABD}^{ABD}(2,2,2) = 0$
ABCD	$\lambda_{BC}^{ABCD}(2,2) = -0.30, \lambda_{ABC}^{ABCD}(2,2,2) = 0.15,$ $\lambda_{BCD}^{ABCD}(2,2,2) = -0.10, \lambda_{ABCD}^{ABCD}(2,2,2) = 0.07.$	

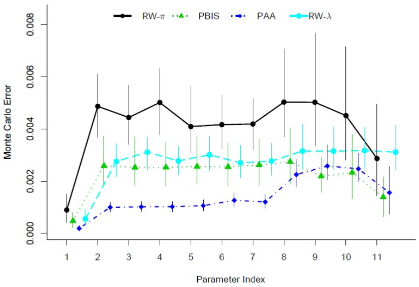
In Figure 2 we report the distribution of the ESS per second of CPU time. PAA is clearly the most efficient among the four methods under consideration.

Fig. 2 ESS per second of CPU time



In Figure 3, for all methods and all marginal log-linear interactions, we represent the time adjusted MCEs for the posterior means. In the graph we represent the 95% error bars of the average time adjusted MCEs for the posterior means. We notice that PAA performs better than all competing methods, since the corresponding MCEs are lower for almost all interactions.

Fig. 3 MCEs for posterior Mean adjusted for CPU time for the simulation study



For a more detailed analysis and a real data application, see Section 5 of [6].

4 Concluding remarks

In this paper we have presented a novel Bayesian methodology for quantitative learning for graphical log-linear marginal models. The main advantages of this approach are the following. It allows us to incorporate in the model prior information on the marginal log-linear interactions. It leads to an efficient and fully automatic setup, and no time consuming and troublesome tuning of MCMC parameters is needed. The authors are planning to extend the method to accommodate fully automatic selection, comparison and model averaging.

References

1. Bergsma, W. P. and Rudas, T. (2002). Marginal log-linear models for categorical data. *Annals of Statistics*, **30**, 140-159.
2. Bergsma, W. O., Croon, M. A. and Hagenaars, J. A. (2009). *Marginal Models. For Dependent, Clustered, and Longitudinal Categorical Data*, Springer
3. Drton, M. and Richardson, T. S. (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society*, Ser. B , **70**, 287-309.
4. Lupparelli, M., Marchetti, G. M., Bersgma, W. P. (2009). Parameterization and fitting of bi-directed graph models to categorical data. *Scandinavian journal of Statistics* **36**, 559-576.
5. Ntzoufras, I. and Tarantola, C. (2013). Conjugate and Conditional Conjugate Bayesian Analysis of Discrete Graphical Models of Marginal Independence. *Computational Statistics & Data Analysis*, **66**, 161-177.
6. Ntzoufras, I. Tarantola, C. and Lupparelli, M. (2018). Probability Based Independence Sampler for Bayesian Quantitative Learning in Graphical Log-Linear Marginal Models, *DEM Working Paper Series*.

Statistical matching by Bayesian Networks
L'uso delle reti Bayesiane nello Statistical Matching

Daniela Marella and Paola Vicard and Vincenzina Vitale

Abstract The goal of statistical matching is the estimation of the joint distribution of variables not jointly observed in a sample survey but separately available from independent sample surveys. The lack of joint information on the variables of interest leads to uncertainty about the data generating model. In this paper we propose the use of Bayesian networks to deal with the statistical matching problem since they admit a recursive factorization of a joint distribution useful for evaluating the statistical matching uncertainty in the multivariate context.

Abstract Lo scopo dello statistical matching è stimare una distribuzione congiunta di variabili osservate separatamente in due campioni indipendenti. La mancanza di osservazioni congiunte sulle variabili di interesse causa incertezza sul modello che ha generato i dati: l'informazione campionaria non è in grado di discriminare tra un insieme di modelli plausibili. In questo lavoro il problema dello statistical matching è analizzato utilizzando le reti Bayesiane che consentono non solo di descrivere la struttura di dipendenza in distribuzioni multivariate ma ammettono una fattorizzazione della distribuzione congiunta utile ai fini della valutazione dell'incertezza.

Daniela Marella
Dipartimento di Scienze della Formazione, via del Castro Pretorio 20, 00185 Roma, e-mail: daniela.marella@uniroma3.it

Paola Vicard
Dipartimento di Economia, Via Silvio D'Amico 77, 00145 Roma, e-mail: paola.vicard@uniroma3.it

Vincenzina Vitale
Dipartimento di Economia, Via Silvio D'Amico 77, 00145 Roma, e-mail: vincenzina.vitale@uniroma3.it

1 Introduction

Statistical matching aims at combining information obtained from different non-overlapping sample surveys, referred to the same target population. The main target is in constructing a complete synthetic data set where all the variables of interest are jointly observed, see [3].

Formally, let (X,Y,Z) be a random variable (rv) with joint discrete distribution P. Without loss of generality, let $X = (X_1,\dots,X_H)$, $Y = (Y_1,\dots,Y_K)$ and $Z = (Z_1,\dots,Z_T)$ be vectors of rvs of dimension H, K, T, respectively. Furthermore, let A and B be two independent samples of n_A and n_B independent and identically distributed records from (X,Y,Z). Assume that (X,Y) are observed in sample A while (X,Z) are observed in sample B. The main goal of statistical matching consists in estimating the joint distribution of (X,Y,Z) from the samples A and B. Such a distribution is not identifiable due to the lack of joint information on Z and Y given X.

In order to overcome this problem, the following approaches have been considered. The first approach uses techniques based on the conditional independence assumption between Y and Z given X (henceforth CIA assumption) see, e.g., [9]. The second approach uses techniques based on external auxiliary information regarding the statistical relationship between Y and Z , e.g. an additional file C where (X,Y,Z) are jointly observed is available, as in [12].

However, it is possible that neither case is appropriate, then the third group of techniques addresses the so called *identification problem*. The lack of joint information on the variables of interest is the cause of uncertainty about the model of (X,Y,Z). In other terms, the sample information provided by A and B is actually unable to discriminate among a set of plausible models for (X,Y,Z). For instance, in a parametric setting and for $K = T = 1$ the estimation problem cannot be "pointwise", only ranges of values containing all the pointwise estimates obtainable by each model compatible with the available sample information can be detected. Such intervals are uncertainty intervals. Uncertainty in statistical matching is analyzed in [8],[11], [4],[1] and [2].

In this paper we propose the use of Bayesian networks (BNs) to deal with statistical matching in the identification problem framework for multivariate categorical data. The first attempt in such direction is in [5] where the CIA is assumed.

The use of BNs is motivated by the following advantages: (i) BNs are widely used to describe dependencies among variables in multivariate distributions; (ii) BNs admit convenient recursive factorizations of their joint probability useful for the uncertainty evaluation in a multivariate context.

The paper is organized as follows. In section 2 the concept of uncertainty in statistical matching when BNs are used is discussed.

2 Uncertainty in Statistical Matching using graphical models

BNs are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG), see [10]. The network consists of two components: the DAG where each node corresponds to a random variable, while edges represent direct dependencies; the set of all parameters in the network. For instance, with regard to the random vector $X = (X_1,\dots,X_H)$, a BN encodes the joint probability distribution of X by specifying: (i) the set of conditional independence statements by means of a DAG and (ii) the set of conditional probability distributions associated to the nodes of the graph. The joint probability distribution can be factorized according to the DAG as follows

$$P(X_1,\dots,X_H) = \prod_{h=1}^{H} P\left(X_h|\text{pa}(X_h)\right)$$

where $P\left(X_h|\text{pa}(X_h)\right)$ is the probability distribution associated to node X_h given its parents $\text{pa}(X_h)$, $h = 1,\dots,H$. Given two nodes $X_{h'}$ and X_h, linked by an arrow pointing from to $X_{h'}$ to X_h, $X_{h'}$ is said parent of X_h, and X_h is said child of $X_{h'}$. We say that two vertices X_h and $X_{h'}$ are adjacent if there is an edge connecting them. Let $fa(X_h) = X_h \bigcup pa(X_h)$ then the clan of X_h is defined as $clan(X_h) = fa(X_h \bigcup ch(X_h))$ where $ch(X_h)$ is the set of all children of X_h.

The non identifiability of a statistical model for (X,Y,Z) implies that both the DAG and its parameters can not be estimated from the available sample information. Two kinds of uncertainty can be distinguished: 1) uncertainty regarding the DAG, that is the dependence structure between the variables of interest; 2) uncertainty regarding the network parameters, given the DAG, *i.e.* the joint probability factorization.

2.1 Uncertainty in the dependence structure

Let P be the joint probability distribution of (X,Y,Z) associated to the DAG $G_{XYZ} = (V,E)$ consisting of a set of vertices V and a set E of directed edges between pairs of nodes. Let us denote by $G_{XY} = (V_{XY},E_{XY})$ and $G_{XZ} = (V_{XZ},E_{XZ})$ the DAGs estimated on sample A and B, respectively. As in [5] G_{XY} and G_{XZ} are estimated subject to the condition that the association structure of the common variables X is fixed. In particular, the DAG G_X is estimated on the overall sample $A \bigcup B$. Given G_X, we proceed to estimate the association structure between (X,Y) and (X,Z) on the basis of sample data in A and B, respectively.

As far as P is concerned, unless special assumptions are made, one can only say that it lies in the class of all joint probability distributions for (X,Y,Z) satisfying the estimate collapsibility over Y and Z, respectively. Formally, we say that the joint probability distribution P is *estimate collapsible* over Z_t if

$$\widehat{P}(X,Y,Z\backslash\{Z_t\}) = \widehat{P}_{G_{XYZ\backslash\{Z_t\}}}(X,Y,Z\backslash\{Z_t\}). \tag{1}$$

That is, the estimate $\widehat{P}(X,Y,Z\backslash\{Z_t\})$ of $P(X,Y,Z\backslash\{Z_t\})$ obtained by marginalizing the maximum likelihood estimate (MLE) of $\widehat{P}(X,Y,Z)$ under the original DAG model (G_{XYZ},P) coincides with the MLE under the DAG model $(G_{XYZ\backslash\{Z_t\}})$, see [7]. Estimate collapsibility over a set Z is defined similarly. In terms of graphs a concept equivalent to estimate collapsibility is the c-removability. A vertex Z_t is c-removable from G_{XYZ} if any two vertices in $clan(Z_t)$ are adjacent, except when both vertices belong to $pa(Z_t)$. Further, the set $Z = (Z_1,\dots,Z_T)$ is sequentially c-removable if all vertices in Z can be ordered so that they can be c-removed according to that ordering. An analogous condition is required for the estimate collapsibility over Y. The class of plausible joint distributions for (X,Y,Z) can be described as follows

$$\mathscr{P}_{XYZ} = \{P : \widehat{P}(X,Y) = \widehat{P}_{G_{XY}}(X,Y), \widehat{P}(X,Z) = \widehat{P}_{G_{XZ}}(X,Z)\} \tag{2}$$

or equivalently by using the graph of the model structure, the class can also be defined as the class of plausible DAGs G_{XYZ} where the variables Z and Y are removable, respectively. Formally

$$\mathscr{G}_{XYZ} = \{G_{XYZ} : Z \quad is \quad removable, \quad Y \quad is \quad removable\} \tag{3}$$

The most favorable case, that for instance happens under CIA, occurs when the class (2) is composed by a single joint probability distribution defined as $P(X,Y,Z) = P(X)P(Y|X)P(Z|X)$. In an equivalent manner, this means that the class (3) collapses into a single graph given by $G^{CIA}_{XYZ} = G_{XY} \bigcup G_{XZ}$ where Y and Z are d-separated by the set X. Note that such a network always belongs to the class (3). Under the CIA, both the dependence structure and the BN parameters are estimable from the sample data.

Clearly, when the CIA does not hold, in order to choose a plausible DAG from the class $\mathscr{G}_{XYZ}$, it is important to have extra-sample information on the dependence structure. This is generally available or can be elicited by experts. As stressed in [6], *qualitative dependencies among variables can often be asserted with confidence, whereas numerical assessments are subject to a great deal of hesitacy*. For example, for $K = T = 1$ an expert may willingly state that the variable Y is related to variable Z, however he/she would not provide a numeric quantification of this relationships.

2.2 Uncertainty in the parameter estimation

Suppose that a DAG G^*_{XYZ} has been selected from the class $\mathscr{G}_{XYZ}$. Let P^* the joint probability distribution associated to G^*_{XYZ}. According to G^*_{XYZ} the distribution P^* can be factorized into local probability distributions some of which can be estimated from the available sample information while other not. In the case of categorical variables, uncertainty is dealt with in [4] where parameters uncertainty is estimated according to the maximum likelihood principle. The parameter estimate maximizing the likelihood function is not unique and the set of maximum likelihood estimates is called likelihood ridge.

Assume that, X_h, Y_k and Z_t are discrete rvs with I, J and L categories, respectively and that their joint distribution is multinomial with vector parameter $\theta^* = \{\theta^*_{ijl}\}$, for $i = 1,\ldots,I$, $j = 1,\ldots,J$, and $l = 1,\ldots,L$. Suppose that from the factorization of P^*, the unique parameter that can not be estimated is the joint probability $P(X_h,Y_k,Z_t)$. Analogously to (2), as far as θ^* is concerned, one can only say that it lies in the following set:

$$\Theta^* = \{\theta^* : \sum_l \theta^*_{ijl} = \widehat{\theta}_{ij.}, \sum_j \theta^*_{i.l} = \widehat{\theta}_{i.l}, \theta^*_{ijl} \geq 0, \sum_{ijl} \theta^*_{ijl} = 1\} \tag{4}$$

where

$$\widehat{\theta}_{ij.} = \frac{n^A_{ij.}}{n^A_{i..}} \frac{n^A_{i..} + n^B_{i..}}{n_A + n_B}, \quad \widehat{\theta}_{i.l} = \frac{n^A_{i.l}}{n^A_{i..}} \frac{n^A_{i..} + n^B_{i..}}{n_A + n_B} \tag{5}$$

are the marginal distribution ML estimates of (X_h,Y_k) and (X_h,Z_t) from samples A and B, respectively. The maximum of the observed likelihood in θ^*_{ijl} is not unique, all the distributions in the likelihood ridge are equally informative, given the data. For details, see [4].

In order to exclude some parameter vectors in Θ^* it is important to introduce constraints characterizing the phenomenon under study. These constraints can be defined in terms of structural zero ($\theta^*_{ijl} = 0$ for some (i,j,l)) and inequality constraints between pairs of distribution parameters ($\theta^*_{ijl} < \theta^*_{i'j'l'}$ for some $(i,j,l),(i',j',l')$). Their introduction is useful for reducing the overall parameter uncertainty. Clearly, the amount of reduction depends on the informativeness of the imposed constraints. The problem of the likelihood function maximization when constraints are imposed may be solved through a modified EM algorithm, see [13].

Example Suppose that an expert can elicit the association structure between the variables of interest (X_1,Y_1,Y_2,Z_1,Z_2). The BN is reported in Figure 1. Note that, $Y = (Y_1,Y_2)$ is sequentially c-removable according to the ordering (Y_1,Y_2), and $Z = (Z_1,Z_2)$ is sequentially c-removable according to the ordering (Z_1,Z_2).

The joint distribution P can be factorized according to the graph as follows

$$P(X,Y,Z) = P(Z_1)P(X_1|Z_1)P(Y_1|X_1)P(Z_2|X_1,Z_1)P(Y_2|X_1,Z_1,Z_2) \tag{6}$$

The parameter $P(Y_2|X_1,Z_1,Z_2)$ can not be estimated from the sample available information in A and B. Nevertheless, such a distribution can be estimated following the approach described in Section 2.2 using an iterative procedure starting by $P(Z_1,Y_2|X_1)$ and ending with $P(Y_2,Z_2|X_1,Z_1)$.

Clearly, the larger is the number of directed edges between the components of Y and Z, the larger is the number of uncertain parameters needed to be estimated in the factorization of the joint distribution P^*.

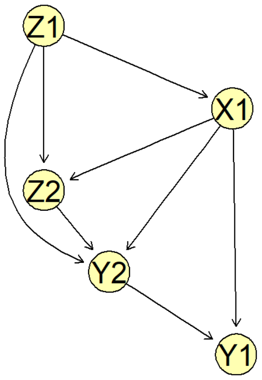


Fig. 1 BN for $(X_1, Y_1, Y_2, Z_1, Z_2)$

References

1. Conti, P.L., Marella, D., Scanu, M.: Uncertainty analysis in statistical matching.*Journal of Official Statistics*, 28, 69-88, (2012)
2. Conti, P.L., Marella, D., Scanu, M.: Statistical matching analysis for complex survey data with applications. Journal of the American Statistical Association. DOI:10.1080/01621459.2015.1112803, (2015)
3. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching: Theory and Practice. Chichester: Wiley, (2006)
4. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *Journal of Offcial Statistics*, 22, 137–157, (2006)
5. Endres, E., Augustin, T.: Statistical matching of Discrete Data by Bayesian Networks. JMLR: Workshop and Conference Proocedings, **52**, 159–170, (2016)
6. Geiger, D., Verma, T., Pearl, J.: Identifying Independence in Bayesian Networks. *Networks*, **20**, 507–534, (1990)
7. Kim, S.H., Kim, S.H.: A Note on Collapsibility in DAG Models of Contingency Tables. Scandinavian Journal of Statistics, **33**, 575-590, (2006)
8. Moriarity, C., Scheuren, F.: Statistical Matching: A Paradigm of Assessing the Uncertainty in the Procedure. Journal of Official Statistics, **17**, 407–422, (2001)
9. Okner, B.: Constructing a new data base from existing microdata sets: the 1966 merge file. Annals of Economic and Social Measurement, **1**, 325–342, (1972)
10. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, (1998)
11. Rässler, S.: Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Springer, New York, (2002)
12. Singh, A.C., Mantel, H., Kinack, M., Rowe, G.: Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. Survey Methodology, **19**, 59–79, (1993).
13. Winkler, W.E.: Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the American Statistical Association, Section on Survey Research Methods, 274–279, (1993).

Sparse Nonparametric Dynamic Graphical Models

Un modello Grafico non parametrico, dinamico e sparso

Fabrizio Poggioni, Mauro Bernardi, Lea Petrella

Abstract We propose a Sparse Nonparametric Dynamic Graphical Model for financial application. We base our model on multiple CAViaR quantile regression models, and we address the issue of the quantile crossing for this type of semi-parametric models. We show how to jointly estimate the multiple quantile levels by exploiting the conditions on the parameters and setting the estimation as a linear constrained optimization problem. We employ the defined non-crossing Multiple CAViaR model as non-parametric estimation of the marginal distributions to get a sparse dynamic graphical model .

Abstract *Proponiamo un modello grafico non parametrico e dinamico per applicazioni finanziarie. Stimiamo modelli di regressione quantilica multipla di tipo CAViaR, ed affrontiamo il tema del non crossing pe modelli semi parametrici. Mostriamo come stimare congiuntamente i diversi livelli garantendo la propriet di non crossing e come usare questi come stime di distribuzioni marginali per ottenere un modello grafico che risulti sparso e dinamico.*

Key words: Multiple Quantile, Non-Crossing, Dynamic Graphical Model,. . .

1 Introduction

In recent years the theme of graphic models has developed in literature. A *Graphical Model* exploits the graph theory as well as the statical theory to describe the

Fabrizio Poggioni
MEMOTEF, Sapienza University of Rome, Italy e-mail: fabrizio.poggioni@uniroma1.it

Mauro Bernardi
Department of Statistical Sciences, University of Padua, Padua, Italy e-mail: mauro.bernardi@unipd.it

Lea Petrella
MEMOTEF, Sapienza University of Rome, Italy e-mail: lea.petrella@uniroma1.it

dependency and conditional dependence ratios of a set of random variables. In financial applications the Graphical model tool allows the advantage of being able to describe complex structures in a simple way. Following the crises of the last decade, it has been very successful as a tool to describe systemic risk as systems of connections in financial markets. Quantile Regression is another very widespread tool in literature used to analyze dependency in financial markets, see for example [1]. We propose in this work an application of graphical models in which we model the time conditional CDF with dynamic quantile regression. Estimating multiple quantile regression involves numerical problems, such as the so called non-crossing problem of the estimated quantiles and the consequent violation of one of the basic principles of the inverse distributions functions: the monotone property. Even if quantile crossing problem is a finite sample problem and should be negligible when the sample size is sufficiently large and the model is correctly specified, for a large number of estimated quantities and perhaps with non-linear specifications of the model quantile crossing may remain a relevant issue. A vast literature exists about the quantile crossing problem: [7] address the quantile-crossing problem using a support vector regression approach for nonparametric models; [4] propose a method for non parametric models in which they use an initial estimate of the conditional distribution function in a first step and solve the problem of inversion and monotonization simultaneously; [9] propose a stepwise method for estimating multiple quantile regression functions without crossing for linear and kernel quantile regression models; [3] propose a method to address the non-crossing problem by rearranging the original estimated non-monotone curve into a monotone rearranged curve; [2] propose a constrained version of quantile regression to avoid the crossing problem for both linear and nonparametric quantile curves. We focus our analysis on the so called semi-parametric Quantile Regression models, specifically on the Conditional Autoregressive Value at. Risk (CAViaR) Quantile Regression model specification introduced by [5], and its multi quantile extension. The existing methods for the non crossing that are applicable to semi-parametric regression models often force the estimate procedure to be step by step, thus less efficient than a joint estimation, and, mostly, these methods do not guarantee that the estimated parameters belong to a ”non crossing” parametric space. We believe that in the case of a regression model for which we have parametrical assumption, the quantile crossing issue can be dealt with more efficiently, moreover we find that ignoring the parametric conditions related to the non-crossing property can lead to serious estimation errors.

2 Non-Crossing MQ-CAViaR as a constrained problem

We show in this chapter how to include the non crossing conditions in the estimation problem as linear constraints to the regression parameters. For the assumptions necessary for the correct specification and estimation of the models we refer to [8]. We consider here a *Symmetric Caviar* specification for two different quantile levels:

$$q_{t_i,\tau_A} = w_A + \phi_A q_{t_{i-1},\tau_A} + \gamma_A |x_0|, \tag{1}$$

$$q_{t_i,\tau_B} = w_B + \phi_B q_{t_{i-1},\tau_B} + \gamma_B |x_0|. \tag{2}$$

Let $\tau_A > \tau_B$, the non crossing condition would be, trivially, $q_{t_i,\tau_A} > q_{t_i,\tau_B}$, $i = 1,2,...,T$. Putting directly this condition as a constraint would not be efficient, contrariwise we consider the assumption made on the stochastic process X_t in order to find necessary conditions on the parameters that satisfy the non crossing issue. Let consider the time series y_t with $t = 1,\dots,T$ and a grid of p different quantile levels, the estimation problem is:

$$\begin{array}{l} \min\limits_{\omega,\gamma,\phi} \sum_{j=1}^{p} \sum_{t=1}^{T} w_{t,\tau_j}(y_t - q_{t,\tau_j}) \\ \text{s.t.} \\ \quad \gamma_{\tau_j} \le \gamma_{\tau_{j+1}}, \quad \forall j \\ \quad \phi_{\tau_j} \ge \phi_{\tau_{j+1}}, \quad j < j^* \\ \quad \phi_{\tau_j} \ge 0, \quad \forall j \end{array} \tag{3}$$

where w_{t,τ_j} is the check function for the quantile level τ_j at time t, $\tau \in [0,1]$ levels are sorted so that $\tau_j < \tau_{j+1}$, and j^* corresponds to τ^*, the quantile level where the autoregressive terms reach their possible minimum ϕ^*. It is necessary to specify that not all the conditions are included in the constraints of the estimation problem, it remains a necessary condition: $w_A(\sum_{1=0}^{n-1}\phi_A^i) - w_B(\sum_{1=0}^{n-1}\phi_B^i) + (\phi_A^n A_0 - \phi_B^n B_0) > 0$, where n is the number of the observations. This last condition can be written more simply, however ensuring this relationship for each integer n means adding a non linear constrain to the estimation problem. Following the tests carried out we choose to proceed this way: we first solve the previous linear constrained problem, then we check if the estimated parameters also meet the non included conditions. If they don't, we estimate the fully constrained problem. Most of the time we did not need to repeat the estimation, thus we believe this way can be more efficient instead of solving the fully constrained problem. It is worth noting that in the "not fully constrained" problem we don't check if our estimated quantiles exhibit crossing to verify the requirement of monotonicity for CDF, whose absence would be just a necessary condition for the monotonicity requirement, instead we check the estimated parameters whose conditions are necessary and sufficient for the non-crossing (monotonicity) requirements. Indeed, forcing the non crossing condition with the constraints $q_{t_i,\tau_A} > q_{t_i,\tau_B}$, or adjusting the estimates that exhibits crossing in this sense, does not guarantee that the non-crossing conditions are satisfied. In the case in which a model with a complex parametric structure is used, for which the non crossing conditions are not exactly known we recommend using simulations of the estimated model to ensure that no crossing occurs instead of just checking that the estimates does not exhibits crossings.

3 The Graphical model: a Sparse Gaussian Copula VAR

A *Graphical Model* exploits the graph theory as well as the statical theory to describe the dependency and conditional dependence ratios of a set of random variables. Nowadays the literature about Graphical Models is huge ,a wide but not exhaustive overview on the subject is [6]. Beyond the obvious advantages of being able to graphically represent dependency structures, we must be able to say something about the dependency and the conditional dependence of random variables. Formally, if X; Y; Z are continuous random variables which admit a joint distribution, we say that X is conditionally independent of Y given Z , and write $X \perp Y|Z \iff f_{XY|Z}(x,y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)$. A graph $G = (V,E)$ is a *conditional independence graph* if it respects this property. In order to exploit the property of conditioned independence we will use the Gaussian copula tool as a graphical model (Gaussian Graphical Model). The Copula approach allows to model multivariate associations and dependencies separately from the univariate marginal distributions of the observed variables. The theorem by [10] shows that every multivariate distribution can be represented in terms of a copula function, which *couples* the univariate marginal distributions, i.e. $F(Y_1,\dots,Y_d) = C(F_1(Y_1),\dots,F_d(Y_d))$. In order to exploit the property of conditioned independence we will use the Gaussian copula tool as a graphical model (Gaussian Graphical Model), the we assume the copula function to be: $C(Y_1,\dots,Y_d) = \Phi_d(\Phi^{-1}(F_1(Y_1)),\dots,\Phi^{-1}(F_d(Y_d)))$, where Φ^{-1} is the univariate standard Gaussian quantile function, and Φ_d is an n-variate Gaussian CDF with mean 0_p and covariance matrix P. Moreover , in the light of the complex relationships that link financial returns we want to describe the conditional dependence of the random variables *between* and *within* time. We choose a VAR specification for the multivariate Gaussian Copula-regression model, thus the structure of the model is of the type: $Y = XB + E$. Where Y denotes the $n \times q$ random response matrix X represents a $n \times k$ regression coefficient matrix containing lagged values of Y, B is the $k \times q$ regression matrix and E is the $n \times k$ error terms matrix. Under Gaussian assumption the multivariate series E is assumed to be distributed as a $N(0,\Sigma)$. As it is known, the estimation of multivariate regression matrices can lead to numerical problems especially for high number of marginals. We employ here the MRCE (multivariate regression with covariance estimation) methodology proposed by [11] which consists in a joint estimation of the parameters in B and in Σ, adding two penalties to the negative log-likelihood function g to obtain sparse estimates for both the matrices. The penalty is of the Lasso type, the penalized estimation problem for $(\hat{B},\hat{\Omega})$ is:

$$\min_{\Omega,\Pi}\{g(B,\Omega) + \lambda_1 \sum_{j' \neq j} |\omega_{j',j}| + \lambda_2 \sum_{j=1}^{p}\sum_{i=1}^{q} |\phi_{j,k}|\} \tag{4}$$

where $\lambda_1 \geq 0, \lambda_2 \geq 0$ are the tuning parameters, $\omega_{j',j}$ are entries of Ω^{-1} and $\phi_{j,k}$ are entries of B. The problem in 4 is not convex, but solving for either Ω or P with the other fixed is a convex problem. Then [11] proposed a cyclical-coordinate descent algorithm for an efficient computation, the authors also implemented an R-package for *MRCE* methodology called MRCE. The MRCE methodology allows us to get sparse estimates of the parameters of a Copula VAR model, this will improve the interpretability of the results and the stability of the predictions.

4 Empirical application

We collected 4000 daily observation of 24 USA banking institutions. We consider a grid of 103 quantile levels $\tau = 0.0001, 0.001, 0.01, 0.02, \ldots, 0.98, 0.99, 0.999, 0.9999$ and for these levels we estimate a symmetric Non Crossing-CAViaR model for each series by solving the constrained problem and checking the non crossing parametric condition as explained in section 2. Then we can use the estimated multiple conditional quantile to get the estimation $\hat{F}(y_t)$ for $t = 1, \ldots, T$. To get the estimate $\hat{F}(y_t)$ we can proceed directly using the 103 estimated quantile at time t to get the estimated probability $\hat{u}_t = \hat{F}(y_t)$ by linear interpolation. Alternatively we can try to get more precision by estimating $\hat{F}(y_t)$ with a smoothing spline.

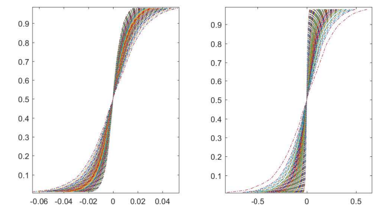


Fig. 1 Some of the time conditional density functions estimated on 103 quantile levels. A period of low volatility in the left graph,a period of high volatility in the right.

Once obtained the series $u_{i,t} = F_{i,t}^{-1}(Y_{i,t})$ with $t = 1, \ldots, 4000$ and $p = 1, \ldots, 24$, we can proceed with the estimation of the parameters of the Copula-VAR function employing the MRCE algorithm to get sparse estimates. We choose a Lag-2 VAR specification and we use the cross validation for the chose of the LASSO shrinking parameters λ_1 and λ_2. Fig. 1 shows some of the estimated cumulative marginal distribution, in Fig.2 the estimated Sparse Graphical model.

5 Conlusions and Discussion

We have identified the parametric space that ensures the non crossing condition for some of the models belonging to the CAViaR specification: this allow the possibility of jointly estimating a considerable number of semi parametric quantile regression models and use them to get non parametric estimation of the marginals for a Copula model. Despite the advantages of knowing the exact parametric space that ensure non-crossing, depending on the parametric assumption of the Multiple Quantile re-

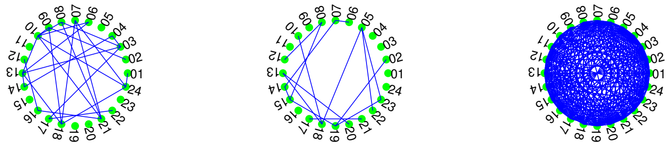


Fig. 2 A circular net representation of the estimated Copula-VAR parameters. From left to right: the regression parameters for the LAG 1, the regression parameters for the LAG 2, and the estimated correlation matrix. Note that the first two networks are intended to be directional.

gression model the constraints to the estimation problem can make the computational part very hard, it would be useful to find algorithms that can efficiently satisfy these conditions. Finally, it is possible to enrich the analysis by adding in the specifications of the models some additional exogenous variable, maintaining the property of non-crossing. The Multivariate part can also be extended by adding exogenous variables, in this case maintaining the same proposed estimation method.

References

1. Bernardi, Mauro and Gayraud, Ghislaine and Petrella, Lea and others:Bayesian tail risk interdependence using quantile regression. Bayesian Analysis & International Society for Bayesian Analysis (2015)
2. Bondell, Howard D and Reich, Brian J and Wang, Huixia: Non-crossing non-parametric estimates of quantile curves. Biometrika & JSTOR (2010)
3. Chernozhukov, Victor and Fernández-Val, Iván and Galichon, Alfred: Quantile and probability curves without crossing. Econometrica & Wiley Online Library (2010)
4. Dette, Holger and Volgushev, Stanislav: Non-crossing non-parametric estimates of quantile curves. Journal of the Royal Statistical Society & Wiley Online Library (2008)
5. Engle, Robert F and Manganelli, Simone: CAViaR: Conditional autoregressive value at risk by regression quantiles. Journal of Business & Economic Statistics & Taylor & Francis (2004)
6. Koller, Daphne and Friedman, Nir: Probabilistic graphical models: principles and techniques. MIT press (2009)
7. Takeuchi, Ichiro and Furuhashi, Takeshi: Non-crossing quantile regressions by SVM. MIT press (2009)
8. White Jr, Halbert L and Kim, Tae-Hwan and Manganelli, Simone: Modeling autoregressive conditional skewness and kurtosis with multi-quantile CAViaR. ECB Working Paper, (2008)
9. Wu, Yichao and Liu, Yufeng: Stepwise multiple quantile regression estimation using non-crossing constraints. Statistics and Its Interface (2009)
10. Sklar, M: Fonctions de répartition á n dimensions et leurs marges. Université Paris 8 (1959)
11. Rothman, Adam J and Levina, Elizaveta and Zhu, Ji : Journal of Computational and Graphical Statistics. & Taylor & Francis (2010)

Non-communicable diseases, socio-economic status, lifestyle and well-being in Italy: An additive Bayesian network model

Malattie croniche, status socio-economico, stile di vita e benessere: un approccio basato sui Modelli di Rete Bayesiani additivi

Laura Maniscalco and Domenica Matranga

Abstract The aim of the paper is to investigate the statistical association, on a sample of Italian subjects, extracted by Survey of Health, Ageing and Retirement in Europe (SHARE) dataset, between chronic diseases (occurrence or number of chronic diseases) and socio-economic and behavioural determinants (lifestyle indicators, QoL indicators, cognitive functioning variables). To this aim, additive Bayesian network (ABN) analysis was used. The resulting ABN model shows that better-educated individuals have better health outcomes, age is direct and gender is an indirect determinant of the number of chronic diseases. Furthermore, self-perceived health is associated with lower number of chronic diseases, lower physical limitations and higher quality of life and these indicators can be considered within a unitary vision to represent well-being of elderly people, as they share a similar distribution by gender and age.

Abstract *Lo scopo dello studio è stato quello di verificare, con dati su un campine italiano, l'associazione statistica tra la malattia cronica ed i determinanti che rigurdano lo status socio-economico, gli indicatori dello stile di vita e della qualità di vita e la funzione cognitiva. Per raggiungere lo scopo posto, sono stati usati i modelli di rete Bayesiani additivi. Il modello ABN scelto mostra che gli individui più istruiti hanno una migliore salute, tenendo costante le altre esplicative, l'età è un fattore determinante diretto e il sesso è un fattore determinante indiretto del numero di malattie croniche. Inoltre, la salute percepita è associata a un minor numero di malattie croniche, a minori limitazioni fisiche e ad una migliore qualità di vita, questi indicatori possono essere considerati nel loro insieme perché rappresentano il benessere degli anziani e condividono una distribuzione simile per genere ed età.*

Laura Maniscalco
Università degli Studi di Palermo , Dipartimento di Chirurgia Neurosensoriale e Motoria, Medicina Orale con Odontoiatria per pazienti a rischio, e-mail: maniscalco.laura92@gmail.com

Domenica Matranga
Università degli Studi di Palermo, Dipartimento di Scienze per la Promozione della Salute e Materno-Infantile "G. D'Alessandro" e-mail: domenica.matranga@unipa.it

Key words: GLM, Additive Bayesian Network, lifestyle, well-being

1 Introduction

In Italy, the main non-communicable diseases (NCDs) all together account for heavy disease burden, due to the increasing population aging. Data from the Italian National Institute of Statistics in 2015 show that the mortality rate for cardiovascular diseases is 512 per 100000 people, while mortality for other diseases is lower but still worrying (294 per 100000 for cancer, 137 per 100000 for chronic respiratory diseases and 82 per 100000 for mental disorders). The WHO NCD global surveillance strategy is based on a multidimensional view of NCD determinants, including physiological and lifestyle influences and environmental and social factors. People with low socioeconomic status (SES) have less access to NCD care and treatment, are less aware of correct lifestyles to hamper the onset of NCDs and to prevent advanced-stage disease and complications. Some studies show that poor living conditions and primary education are associated with physical inactivity [7] and that there is an inverse social gradient for the feminine obesity [4]. Other studies show the positive impact of education on the consumption of healthy nutrients and on the reduction of individual body-mass index [3]. The majority of chronic diseases affects the overall health of patients by limiting their well-being, the functional status, productivity and health-related quality of life. The main limitation of psychosocial well-being regards the minor involvement in social activities that implicates a reduction of positive reinforcement [5]. A classical approach to investigate the statistical association between chronic diseases (presence or not of chronic diseases) and covariates of interest (SES variables, lifestyle indicators, QoL indicators, cognitive functioning variables) is based on generalized linear models, which make a net classification of variables into covariates and the response. In a multifactorial complex disease system, such that one of NCDs, it should be desirable to analyze the associations between all covariates with all variables being potentially dependent, using Additive Bayesian Networks (ABN) [10]. In order to explain this interrelationship, a data-driven approach using Additive Bayesian Network was applied to find the most probable structure.

2 Material and Methods

2.1 The sample

For the purpose of our study, it was analysed the Survey of Health, Ageing and Retirement in Europe (SHARE) dataset. This dataset is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks on 27 European countries and Israel. In particular, for this analysis we focused on Italian data. The dataset contains, 5288 observations and 17 variables, after having removed missing values. Categorical variables with multiple categories have been transformed into binary, in order to analyze the data with the ABN methodology.

2.1.1 Dataset structure

The variables under study are: gender with two categories "Male " and "Female ", age of the respondent, years of education (yedu), marital status (mstat) with categories "Married" and "Not Married", household total net income (thinc), household net worth (hnetw), current job status (cjs), with categories "Labor force"and "Not labor force", Body Mass Index (BMI) of the respondent, smoking (esmoked), with categories "Yes" and "No", physical inactivity (phinact), with categories "Yes" and "No", number of chronic disease (chronic), US version of self-perceived health (sphus), with categories "Less than good" and "More or equal than good", global activity limitation (gali), a binary variables with categories "Limited" and "Not Limited", score of verbal fluency test (fluency), score regarding "first trial ten words list learning" test (cf008tot), score regarding "delayed trial ten words list learning" test (cf016tot), quality of life in older ages (casp), ranging between 0 to 12. Fluency, cf008tot and cf016tot measure subjective and objective aspects of the respondent's cognitive function, like memory and verbal fluency.

2.2 Statistical methods

Descriptive statistics were carried out. Continuous variables were summarized with mean, median and standard deviation. Categorical variables were analysed with frequencies distributions. Additive Bayesian network (ABN) analysis was used to identify factors associated with the non-communicable disease. Bayesian networks (BNs) belong to the family of probabilistic graphical models (GMs). These models represent a set of variables and their conditional dependencies through a directed acyclic graph (DAG). They are enabled to represent the joint probability distribution (JDP) over a set of random variables. The structure of a DAG is represented by a set of nodes that represent random variables and a set of edges visualized by arrows between nodes that are direct dependencies among variables. Additive Bayesian networks (ABN) are a special type of BN models, where the parameters instead to be based on contingency table, are based on generalized linear model (GLM). Each node in the DAG is the equivalent to the response variable in a GLM. To all DAG was imposed a uniform prior to allow a full data-driven approach. The identification of the best DAG was doing with an exact search method and the identification of the best DAG was based on the marginal log-likelihood. The marginal log-likelihood represents the goodness of fit metric in Bayesian modeling and it includes an implicit penalty for model complexity. The first step to identify the best DAG consists in increasing the maximum number of parents allowed per node (that are the number of allowed covariates in each model) until the goodness of fit remained constant. The model selection procedure started from one to twelve possible parents per node. The best DAG was identified, with eleven maximum number of possible parents per node. The second step regards the model adjustment for over-fitting, that usually is doing when the sample is small and there are many variables. In this step before to generate dataset with MCMC simulation it useful check if the area under the curve of the posterior density integrates to one. In the third step, the marginal posterior log odds ratio was estimated for each parameter from the integration of the posterior distribution with respect to the set of parameters. The maximum likelihood estimates were obtained by the joint posterior distribution. Since ABN methodology allows to evaluate the association between all variables, an arc between two variables in the final ABN model indicates a “direct” relationship, whereas an “indirect” relationship is defined as a relationship between two variables through an intermediate variable. Data were analyzed with R software (version 3.3.2), and the ABN methodology was implemented with “abn” package [11].

3 Results

The exploratory data analysis revealed that there is a strong correlation between cf008tot and cf016tot (p=0.72) and a negative correlation between age and cf008tot (r=-0.39), a positive correlation between chronic and age (r=0.35) and between years of education and cf008tot. All other quantitative variables show a correlation < 0.33. By exploring the scatter-plot of the marginal log-likelihood, the optimal number of parents allowed was found to be six. The optimal DAG (see Fig. 1), showed that the number of chronic disease showed a direct association with BMI (coeff= 0.13), years of education (coeff= -0.073) and age (coeff= 0.12) and an indirect association with gender (coeff= 0.26). Very interestingly, gender is a direct determinant of smoking status. Relating to self-perceived health, the DAG shows direct associations with chronic diseases (coeff= -0.52), cognitive function indicator (cf008 coeff=0.07, cf016tot coeff= 0.29 and fluency coeff=0.04), global activity limitations (coeff= 1.99), quality of life (coeff= 0.12) and with socio-behavioural determinants (marital status coeff=-0.36, working status coeff= -0.65 and physical inactivity coeff= -0.63).

4 Discussion

Additive Bayesian networks analysis of the Share dataset confirms the multidimensional approach of NCD determinants which is suggested by WHO [8]. According to this framework, all determinants are distributed along four successive causation levels, going from physiological influences to social structure, passing through lifestyle and environmental influences. Among physiological and lifestyle influences, our study found out BMI as direct and smoking as indirect determinant of the number of chronic diseases in line with other literature [2, 6]. Within the social structure, another important finding was the role of education and age as direct determinants and of gender as indirect determinant of the number of chronic diseases. Better-educated individuals obtain better health outcomes from a fixed set of inputs because they have the abilities and information to make better choices for their lifestyles. Education indirectly facilitates individual development and interpersonal relationships, enabling people to pursue personal and professional success, which has a positive impact on health [1]. Another finding of our study is that better self-perceived health is associated with lower number of chronic diseases, lower physical limitations and higher quality of life. All these indicators can be considered within a unitary vision to represent well-being of elderly people, as they share a similar distribution by gender and age. In fact, a higher proportion of males yield positive

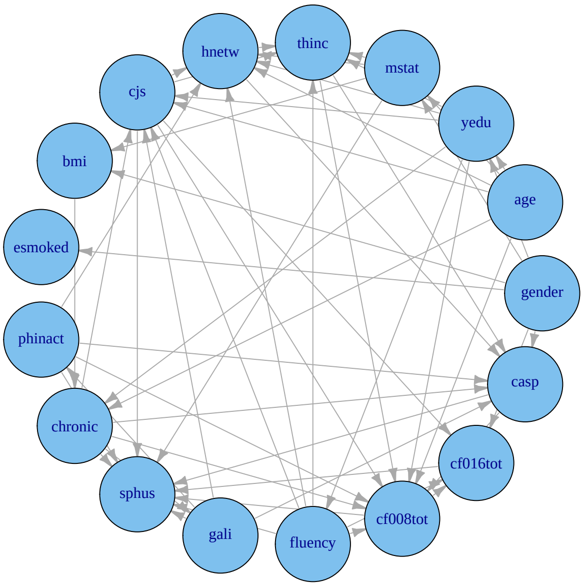


Fig. 1 Optimal ABN model, in terms of marginal likelihood, with a maximum number of six parents per node

outcomes on psychological distress, self-rated health, life satisfaction and chronic diseases and a slightly higher percentage of females display better outcomes regarding their quality of life and BMI. Furthermore, a higher level of well-being appeared to be prevalent among elderly people embedded more frequently in social activity participation [9]. One strength of our study relates to the statistical method used. Additive Bayesian networks supply a general framework to understand the multiple association process that can emerge from the complex interrelationship among health indicators, socio-economic and behavioural determinants. Moreover, through prior elicitation, it should be possible to embed all available information about the statistical association among them. Finally, the data-driven approach could help us to formulate new research hypotheses and to design and construct subsequent theoretical models.

References

1. Di Cesare, M. et al.: Inequalities in non-communicable diseases and effective responses. The Lancet, Elsevier **381**, 585–597 (2013)
2. Giampaoli, S. and Palmieri, L. et al.: Cardiovascular health in Italy. Ten-year surveillance of cardiovascular diseases and risk factors: Osservatorio Epidemiologico Cardiovascolare/Health Examination Survey 1998–2012. Eur J Prev Cardiol, SAGE Publications Sage UK: London, England **22**, 9–37 (2015)
3. Matranga, D. and Tabacchi, G. and Cangialosi, D.: Sedentariness and weight status related to SES and family characteristics in Italian adults: exploring geographic variability through multilevel models. Scandinavian journal of public health, SAGE Publications Sage UK: London, England, (2017)
4. McLaren, L.: Socioeconomic status and obesity. The Oxford handbook of the social science of obesity.
   http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199736362.001.0001/oxfordhb-9780199736362-e-016. Accessed 23 September 2016
5. Megari, K.: Quality of life in chronic disease patients. Health Psychology Research, PAGEPress **1**, (2013)
6. Panico, S. and Palmieri, L. et al.: Preventive potential of body mass reduction to lower cardiovascular risk: the Italian Progetto CUORE study. Prev Med, Elsevier **47**, 53–60 (2008)
7. van Der Berg, J.D. and Bosma, H. et al.: Midlife determinants associated with sedentary behavior in old age.. Med Sci Sports Exerc, NIH Public Access **46**, 1359 (2014)
8. World Health Organization: STEPS: A framework for surveillance. (2003)
   Available at $http : //www.who.int/ncd_surveillance$
9. Vozikaki, M. and Linardakis, M. et al.: Activity participation and well-being among European adults aged 65 years and older. Social Indicators Research, Springer **131**, 769–795 (2017)
10. Lewis, F.I. and McCormick, B.:Revealing the complexity of health determinants in resource-poor settings. American journal of epidemiology **176**, 1051–1059 (2012)
11. Pittavino, M. and Lewis, F. and Furrer, R.: an R package for modelling multivariate data using additive Bayesian networks. The Comprehensive R Archive Network (CRAN), 1–37, (2016)

Using Almost-Dynamic Bayesian Networks to Represent Uncertainty in Complex Epidemiological Models: a Proposal

Utilizzo di una Rete Bayesiana Quasi-Dinamica per Rappresentare l'Incertezza in Modelli Epidemiologici Complessi: una Proposta

Sabina Marchetti

Sommario We introduce a dynamic model to deal with uncertainty in complex epidemiological processes. Our proposal is based on the Dynamic Bayesian Networks formalism, where each node is associated a random variable, whose value specifies the state of an individual from a given population.
Sommario *Si introduce un modello dinamico per gestire l'incertezza nei processi epidemiologici complessi. La proposta presentata si basa sul formalismo delle Reti Bayesiane dinamiche, in cui ciascun nodo corrisponde ad una variable aleatoria, il cui valore definisce lo stato di un individuo in una data popolazione.*

Key words: SIR Model, Dynamic Bayesian Network, Propagation of Uncertainty

1 Introduction

Deterministic epidemiological models for the propagation of an infectious disease on a given population were first introduced by Kermack and McKendrick [5]. We propose a graphical implementation of a simple SIR model, to account for uncertainty in the propagation process. This is carried out via repeated simulations from what we call an *almost-dynamic* Bayesian network, whose pattern of mutual conditional relevances evolves with time.
Basic concepts on the SIR model and Dynamic Bayesian Networks are introduced in Sec. 2.1 and 2.2, respectively. Our proposal is sketched in Sec. 2.3. Some results, remarks and possible extensions are discussed in Sec. 3.

Sabina Marchetti
Dip. Scienze Statistiche, Università Sapienza, P.le A. Moro 5, 00185 Roma (Italy), e-mail: sabina.marchetti@uniroma1.it

2 Methods

2.1 Deterministic Epidemiological Models

Infectious disease models are based on ordinary differential equations (ODEs). Each equation is associated with a single layer of the population, that is thus partitioned into homogeneous compartments. As an example, consider the well-known SIR model [5], where a given population is constituted by three mutually exclusive and exhaustive components: susceptible (S), infectious (I) and recovered (R), to some infectious disease. At any time t, $t \geq 0$, people in S may acquire the disease, according to rate $\lambda(t)$, and enter compartment I, where they spend γ^{-1} units of time on average, before they move to R; see Fig. 1. The rate $\lambda(t)$ is called *force of infection* and is often defined as

$$\lambda(t) = \tau c \frac{I(t)}{N(t)}, \tag{1}$$

where $c > 0$ is the average number of contacts per individual per unit of time, E.g. per day, resulting in an infection according to $\tau \in [0,1]$, *transmissibility* parameter, and $I(t)/N(t)$ is the proportion of infectious individuals in the population, also called the *infection prevalence*, at t. If the population is stratified into k homogeneous classes, that specify different behavior among components, $\lambda(t) \in \mathbb{R}^k$ is derived from the product of τ and contact matrix $C \in \mathbb{R}^{k \times k}$, multiplied by vector $I_k(t)/N_k(t) \in \mathbb{R}^k$, whose elements correspond to the proportion of infectious individuals in each class.

γ is called *recovery rate*. ODEs describe the flows across compartments. At each time step, $S(t), I(t)$ and $R(t)$ report the number of individuals in each homogeneous component of the population, assumed constant, i.e. $\frac{\partial S(t)}{\partial t} + \frac{\partial I(t)}{\partial t} + \frac{\partial R(t)}{\partial t} = 0$, for all $t \geq 0$. No demographic dynamics (births, deaths, ageing of the population, etc.) are considered by a simple SIR model, whose flow diagram is depicted in Fig. 1. Also, by definition individuals are assumed to acquire lifelong immunity to the disease considered, once recovered. Thanks to its simple parametrization, the SIR model was applied in a number of works, E.g., [4]. Particularly, the related set of assumptions it applies to short-termed diseases, where dynamics strictly related to the population may be neglected, as well as total or partial waning of the immunity conferred by an infection.

When an infectious individual is introduced in a fully susceptible population, the number of secondary cases she produces in a unit of time is called *basic reproduction number* [5], denoted as R_0. It may be easily proved that in an SIR model, it corresponds to $R_0 = \tau c \gamma^{-1}$ (see, E.g., [1]).

Uncertainty and sensitivity analysis of deterministic epidemiological models is usually performed following what we call a *second-order* approach: each parameter value is varied within some range according to some distribution of uncertainty, either singularly or simultaneously. Stochastic approaches were also proposed, dating back to the Reed-Frost model, in 1928 (see [2] for a survey), to deal with uncertainty in such propagation processes. In Sec. 2.3, we propose a dynamic network model, that accounts for uncertainty

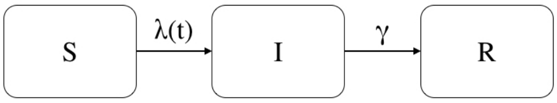


Figura 1 Flow diagram of a simple SIR model.

both in the parameters and in the pattern of contacts of any individual from the population. This way, while homogeneity of each compartment results in the parametrization of a collection of conditional probability tables, corresponding to transition matrices, the topology of the network increases (or decreases) risks at each time step. Our proposal is based on a graphical dynamic implementation of an SIR model. Our proposal is alternative to both existing epidemiological stochastic approaches, as it accounts for individual risks, and to the so-called *individual based* modeling [3], whose assumptions differ from those of an SIR.

2.2 Probabilistic Graphical Models

Probabilistic graphical models (PGMs) are used to represent the joint probability distribution of a (large) collection of random variables $\mathbf{V} = \{X_0, \dots, X_n\}$, by a graph. Bold letters are used to denote sets of random variables. PGMs exploit conditional independence assumptions among pairs of random variables (in a one-to-one correspondence with the nodes of the graph) to reduce inferential complexity.

Nodes are connected by arcs ($-$) in an undirected graph, while edges ($\rightarrow$) are used to induce an ordering among the elements of $\mathbf{V}$ in a directed graph. We write $Adj(X)$ to denote the set of nodes *adjacent* to X in any graph, irrespective of their direction. Let X and Y be any two adjacent nodes in a directed graph, if there is an outgoing edge from X into Y, X is called a *parent* of Y, whereas Y is a *child* of X. $Pa(X)$ and $Ch(X)$ denote, respectively, the parents and children set of node $X \in \mathbf{V}$. Arcs are denoted as $((X,Y))$, while edges read (X,Y); i.e. let E be the set of links in the graph, $((X,Y)) \in E$ implies $X \in Adj(Y)$ and $Y \in Adj(X)$, while $(X,Y) \in E$ implies $X \in Pa(Y)$, $Y \in Ch(X)$, for any pair $X,Y \in \mathbf{V}$. Also, let $d_X = |Pa(X)|$ denote the in-degree of node X, i.e. cardinality of its parents set.

Bayesian networks are PGMs whose graphical component is an acyclic directed graph $\mathscr{G} = (\mathbf{V}, E)$. A Bayesian network is specified by the pair $(\mathscr{G}, P)$, where P is a strictly positive joint probability mass function over $\mathbf{V}$. By the Markov condition, for a given ordering in $\mathbf{V}$, each node is independent of its non-descendants in the graph, given its parents. It follows P may be equivalently represented by a collection of $n+1$ conditional probability tables (CPTs), whose columns correspond to distinct configurations of the parents of a node.

Dynamic Bayesian Networks (DBNs, [6]) are sequences of Bayesian networks, whose structure and/or parametrization change with time. In a DBN, conditioning always extends to each node's previous state, and possibly its parents'. As a result, the joint PMF at time t corresponds:

$$P(X_0^{t+1} = x_0^{t+1}, \dots, X_n^{t+1} = x_n^{t+1}) = \prod_{i=0}^{n} P\left(X^{t+1} = x_i^{t+1} | X^t = x_i^t, Pa^{t+1}(X_i) = pa^{t+1}(X_i), Pa^t(X_i) = pa^t(X_i)\right) \tag{2}$$

where each configuration $(X_0^t = x_0^t, \dots, X_n^t = x_n^t)$ belongs to product sample space $\Omega_{\mathbf{V}} = \times_{i='}^{n} \Omega_{X_i}$, and $Pa^t(X_i) = pa^t(X_i) \in \times_{X_j \in Pa^t(X_i)} \Omega_{X_j}$ is consistent with the first, $t \geq 0$. For a given event $\mathbf{x}^t \in \Omega_{X \in \mathbf{X}}$, we write $P(\mathbf{X}^t = \mathbf{x}^t) = P(\mathbf{x}^t)$ to simplify notation.

In graph theory, a population may be described by a network, whose nodes correspond to units, i.e. individuals, whereas in PGMs, nodes are random variables. In next section, we will introduce a simplified DBN whose nodes represent individuals. Each node X is associated a three-valued random variable, whose states, x_S, x_I and x_R, indicate her location in an SIR model.

2.3 Dynamic Probabilistic Modeling of an SIR Model

Let $\mathscr{B}^t = (\mathscr{G}^t, P^t)$ denote a graphical model at any time $t \geq 0$. In our formalism the graphical component is partially directed: with time, arcs may be changed into edges, and *vice versa*. Although conditioning ought to consider the whole adjacency set of each node, say X, as for undirected networks, relevance is restricted to those in $Pa(X) \subseteq Adj(X)$. Hence, at each t, $\mathscr{G}^t$ may be intended as an acyclic directed graph.

$\mathscr{G}^t = (\mathbf{V}, E^t)$, represents the pattern of contacts among units of a population. The graph may either be a given social network, or be randomly generated according to some known contact matrix C. Let $\mathbf{V}$ be the set of $(n+1)$ nodes (individuals in the network); as already mentioned, each random variable X_i takes values in $\Omega_{X_i} = \{x_{i,S}, x_{i,I}, x_{i,R}\}$, $i = 0, \ldots, n$. State $x_{i,j}$ indicates X_i belongs to compartment j of an SIR structure, $j = S, I, R$. Also, each X_i is assigned label t_i, initialized as $t_i = -\infty$.

In our model, parameters of the model do not vary with time. Yet, if X_i takes value $x_{i,I} \in \Omega_{X_i}$, all incoming arcs are converted into $Ch^t(X_i)$[1] and t_j is updated to t. Let $d_{j,t} = |Pa^t(X_j)|$ be the *infectious-indegree* of node X_j, conditioning involves the subset of adjacent nodes of X_j in $Pa^t(X_j)$, and its corresponding state at $(t-1)$, for any $t \geq 1$.
Based on Eq. (1), we derive the individual FOI $\lambda_{j,t} = \tau d_{j,t-1}$, $j = 0, \ldots, n, t \geq 0$. At each t, $P(X^t | X^{t-1}, Pa^t(X)) = \{P(x^t | x^{t-1}, pa^t(X)) : x^t, x^{t-1} \in \Omega_X, pa^t \in \Omega_{Pa^t(X)}\}$. The CPT of node X_j at t is specified in Table 1, that may be as well intended as a transition matrix, whose columns sum to one.

In details, the first column of the CPT represents the infection process, i.e. people moving from S to I in an SIR model. We assume the infectious-indegree of any node X_j serves as a proxy of the product $cI(t)/N(t)$ from Eq. (1). By definition, *zero-case* $(X = x_I)$, i.e. a single infectious individual in a fully susceptible population, is expected to produce $|Ch^0(X)|$ primary infections, at most. We define the basic reproduction number associated to $\mathscr{B} = \cup_{t=0}^{\infty} \mathscr{B}^t$ as follows:

$$R_0^{\mathscr{B}} = \min\left(|Ch^0(X)|, |Ch^0(X)|\tau\gamma^{-1}\right) \geq 0\,. \tag{3}$$

Let $\tau = 1$, if recovery is fast, i.e. γ is large, $R_0^{\mathscr{B}}$ will likely overestimate the number of secondary cases.

	$x_{j,S}^{t-1}; Pa^t(X_j)$	$x_{j,I}^{t-1}; Pa^t(X_j)$	$x_{j,R}^{t-1}; Pa^t(X_j)$
$x_{j,S}^t; t_j$	$e^{-\lambda_{j,t}}$	0	0
$x_{j,I}^t; t_j$	$1 - e^{-\lambda_{j,t}}$	$e^{-\gamma(t-t_j+\varepsilon)}$	0
$x_{j,R}^t; t_j$	0	$1 - e^{-\gamma(t-t_j+\varepsilon)}$	1

Tabella 1 CPT of node X_j, at time t in an SIR-Bayesian network. $\varepsilon \geq 0$, in our application (see Fig. 2) we set $\varepsilon = 1$.

3 Results and Discussion

As a toy example, we applied our proposal to a population of $(n+1) = 20$ individuals, whose pattern of contacts is depicted in Fig. 2(top-panel). The propagation schema was produced over $M = 1000$ simulations of the model with $\tau = 0.45$ and $\gamma = 0.60$. At each m, introduction of a randomly selected infectious zero-

[1] Arcs only are converted into outgoing edges: if X_i has already incoming edges, those are unchanged.

cases produced on average 3.56 secondary cases.[2] We compared our results with the corresponding SIR model; epidemic curves (and empirical quantile values of uncertainty) are depicted in Fig. 2(top-panel).

Let us state some general remarks. As a first, we stress recovery of some node X_i, at $t \geq 1$, acts as a blocking mechanism, flooring $P(x^t_{j,I}|X^{t-1}, Pa^t(X_j) = \{X_i\}, t_j = \infty)$ to zero.[3] Detection of paths originating from the zero-case, say X_0, that are likely to be blocked by recovery of a single node (or few of them) allows prior identification of critical subjects. Those subjects may be targeted by prevention strategies, tackling a minimal subset approach. Graphical tools may be used to detect *blocked* from *active* paths [7] for the propagation of a disease, such as the Bayes Ball algorithm [8].

In this direction, we argue a general evaluation of the topology of the network would critical in this sense. E.g., for a fixed transmissibility, a sparse graph will be more exposed to the blocking mechanisms mentioned above, compared to a denser one. Also, identification of *cliques*, i.e. maximally connected sets of nodes, may constitute valuable knowledge to policy planning.

As a second remark, repeated sampling allows to evaluate uncertainty in the overall propagation process. Several sampling procedures were proposed in the literature of DBNs, see, E.g., [6]. A efficient naive approach would simply take the so-called *maximum a posteriori* (MAP) configurations from $\Omega_{\mathbf{V}}$, at each time step, to update E^t.

Additionally, other than $R_0^{\mathscr{B}}$, measures on the disease propagation process may be derived from repeated iterations, as well as analytically. Among others, incidence and sero-prevalence of a disease [1]. Again, MAP configurations may be considered as average values prior to simulating.

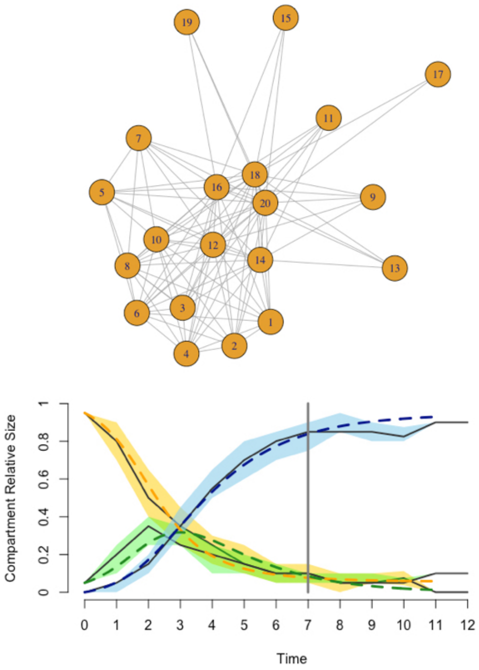


Figura 2 Top-panel: Network generated at random, with $|\mathbf{V}| = 20$, average in-degree 10 and $d = 19$. Bottom-panel: Epidemic curves resulting from $M = 1000$ simulations on the network above, $|\mathbf{V}| = 20$, $\tau = 0.45$, $\gamma = 0.60$. At every simulation, a node is selected at random as zero-case. Epidemic curves describe the relative size of compartments S (yellow), I (green) and R (light blue). Quantile bands and median are compared with the curves resulting from the SIR network, with $c = 0.20$; dashed lines orange, dark green and blue correspond, respectively to compartments S, I and R.

[2] We expected $R_0^{\mathscr{B}} \in [1.33, 10]$, by Eq. (3).

[3] Since $d_{j,t} = 0$.

As a further point, a straightforward extension might assume contacts are characterized by different strengths $w_{i,j} \in [0,1]$, such that $w_{i,j} \to 0$ indicates a almost *vacuos* contacts between nodes X_i and X_j, for any $t \geq 0$. Then, a more general definition of infectious-indegree may be introduced:

$$d_{j,t} = \sum_{X_i \in Pa^t(X_j)} w_{i,j} \mathbb{I}_{X_i^t = x_{i,I}^t} .$$

Finally, suppose we are interested in modeling an SIRS model, where recovered individuals move back to compartment S according to some rate $\phi \geq 0$, that is after ϕ^{-1} units of time in average. It suffices to replace the third column of Table 1 with $\left[e^{-\phi}, 0, 1-e^{-\phi}\right]^T$.

4 Conclusions and Future Work

We proposed an almost-Dynamic Bayesian Network to efficiently deal with uncertainty in the propagation process of a given infectious disease in an SIR model. Particularly, our proposal models uncertainty in the propagation process by i) probabilistic modeling of the transitions across compartments (analogously to a stochastic SIR model), ii) accounting for the dynamic topology of the network (like any individual-based model). We stress our proposal is aimed to provide an intuition of the methodology: extension to models with several layers of complexity is straightforward, the increased complexity being restricted to the preliminary compilation process, i.e. to its parametrization, without affecting the inferential complexity. Future work will consider applications in this direction.

Future research will also consider a thorough approach to uncertainty, by incorporating uncertainty in the parameters, E.g. by means of auxiliary root nodes, to further extend stochastic epidemiological modeling based.

Riferimenti bibliografici

[1] Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[2] Tom Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35, 2010.

[3] Volker Grimm and Steven F Railsback. Individual-based modeling and ecology:(princeton series in theoretical and computational biology). 2005.

[4] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2011.

[5] W Kermack and A McKendrick. A contribution to the mathematical theory of epidemics. In *Proc. Roy. Soc. Lond*, pages 700–721, 1927.

[6] Kevin Patrick Murphy. Dynamic bayesian networks: representation, inference and learning. 2002.

[7] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[8] Ross D Shachter. Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 480–487. Morgan Kaufmann Publishers Inc., 1998.

Educational World

How to improve the Quality Assurance System of the Universities: a study based on compositional analysis

Bertaccini B., Gallo M., Simonacci V., and Menini T.

Abstract The National Agency for the Evaluation of Universities and Research (ANVUR) has for some decades defined the criteria for systematically evaluating student satisfaction. The analysis of these data presents various difficulties both in terms of data collection and analysis. The aim of this work is to propose Candecomp/Parafac for a compositional analysis, which is able to capture the multidimensional aspects of the phenomenon taking into account its ordinal nature and the temporal characteristics of data collection.
Abstract L'Agenzia nazionale per la valutazione delle università della ricerca (ANVUR) definisce da qualche decade i criteri per valutare in modo sistematico la soddisfazione degli studenti. L'analisi di tali dati presenta diverse difficoltia in termini di raccolta dati che di analisi. Scopo del presente lavoro è proporre il Candecomp/Parafac per un'analisi composizionale, il quale, rispettando le caratteristiche temporali della raccolta dei dati e la natura ordinale degli stessi, è in grado di cogliere gli aspetti multidimensionali del fenomeno.

Key words: compositional data, log-ratios, quality, student satisfaction.

Bruno Bertaccini
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence
e-mail: bruno.bertaccini@unifi.it

Michele Gallo
Department of Human and Social Sciences, University of Naples "L'Orientale", Italy.
e-mail: mgallo@unior.it

Violetta Simonacci
Department of Human and Social Sciences, University of Naples "L'Orientale", Italy.
e-mail: vsimonacci@unior.it

Tullio Menini
Department of Human and Social Sciences, University of Naples "L'Orientale", Italy.
e-mail: menini@unior.it

1 Introduction

Since the adoption of Law No. 370/99 in 1999, universities are directly responsible for the system of students' opinion polling on the educational modules undertaken. The main goal is to handle the information collected in a synthetic, efficient and effective way in order to obtain an actual review of the didactical organization and a general improvement of the courses offered by the university. The National Agency for the Evaluation of Universities and Research (ANVUR), established in 2006, defines the criteria to systematically assess student satisfaction on the educational activities as part of the Quality Assurance System of the Universities. Student surveys on the educational training provided, become a necessary step for the accreditation of Universities and single courses of study.

In student satisfaction analysis it is important to keep in mind that the rating expressed by each student does not only reflect the quality of service but also her/his perception, personal experiences, inclinations and socio-cultural background. Each student uses a subjective scale which influences his evaluation of the attributes of education. To address this issue a ratio-based approach such as compositional data analysis can be useful. In order to be coherent, the analysis of this kind of data should also take into account the variability of its many observable attributes, which are generally collected with multiple item questionnaires. Consequently, when student satisfaction is observed across the academic years the use of multilinear techniques as Candecomp/Parafac (CP) is advisable.

Proceeding from these considerations, a compositional analysis by CP is proposed to evaluate student satisfaction. Thus, a short review of the method proposed is given in Sect 2, while in Sect 3 the case study is presented.

2 Compositional analysis: short review

Compositional data were extensively studied in the 1980s by Aitchison, for more details see (1). The turning point of this methodology was the possibility to adopt a transformation in a log-scale to move from a constrained space, defined simplex, to the real unconstrained euclidean space. This important finding contributed towards the use of compositional data in several disciplines. In recent years, a further development of the theoretical framework (namely the principle of working in coordinates (2)) laid the basis for a practical use of compositional data in additional fields of studies as evaluation of academic educational quality (5).
Following (4) for three-way data, where on the first way there are I students, on the second way there are J aspects (variables) and on the third way there are K academic years, the starting three-way array $\underline{\mathbf{V}}$ $(I \times J \times K)$ is transformed in a new array $\underline{\mathbf{Z}}$. Here the frontal slices are obtained as $\mathbf{Z}_k = \mathbf{L}_k \mathbf{P}_J^{\perp}$, where $\mathbf{L}_k$ is the k-th frontal slice of the array $\underline{\mathbf{L}}$ with generic element $log(\dot{v}_{ijk})$. In addition, $\mathbf{P}_J^{\perp} = \left(\mathbf{I} - \mathbf{1}^T\mathbf{1}/J\right)$ represents the idempotent row centering matrix where $\mathbf{I}$ is an identity matrix of order J and $\mathbf{1}$ is a J dimensional vector of 1s. A three-way analysis can now be performed

on $\underline{\mathbf{Z}}$ where the generic row vector $\mathbf{z}_{ik}$ is a centered log-ratio (*clr*) defined as:

$$\mathbf{z}_{ik} = clr\left(\dot{\mathbf{v}}_{ik}\right) = \left[\log\frac{\dot{v}_{i1k}}{g(\dot{\mathbf{v}}_{ik})}, \ldots, \log\frac{\dot{v}_{iJk}}{g(\dot{\mathbf{v}}_{ik})}\right] \text{ with } g(\dot{\mathbf{v}}_{ik}) = \sqrt[J]{\prod_{j=1}^{J} \dot{v}_{ijk}} \quad (1)$$

As for principal component analysis on two-way arrays, the use of *clr*-coefficients allows for a direct application of the CP model on the array $\underline{\mathbf{Z}}$ without further concerns. We thus obtain the three loading matrices $\mathbf{A}$ $(I \times F)$, $\mathbf{B}$ $(J \times F)$, and $\mathbf{C}$ $(K \times F)$, where F is the number of components extracted. This decomposition can be written as follows:

$$\mathbf{Z}_I = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{R}_I \quad (2)$$

$\mathbf{Z}_I$ $(I \times JK)$ and $\mathbf{R}_I$ $(I \times JK)$ identify the array of *clr*-coordinates and the array of residuals respectively, unfolded with respect to the first mode. In other words they are the horizontal concatenation of the $k = 1, \ldots, K$ matrices $\mathbf{Z}_k$ and $\mathbf{R}_k$ of dimension $(I \times J)$, i.e. frontal slices. The symbol $\odot$ identifies the Khatri-Rao product. The parameters contained in the three loading matrices are often estimated in a least squares sense by use of the Alternating Least Squares algorithm, an iterative procedure for which the objective function is:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} = \|\mathbf{Z}_I - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|^2 \quad (3)$$

where $\|\cdot\|$ is the Frobenius norm.

Of course, the CP results are expressed in $clr-$coordinates, thus it is important to keep in mind that they need to be translated back into compositional terms for proper interpretation, see (3) and (6) for more details.

3 Case study

Following the ANVUR criteria, the University of Florence monitors the students' opinions on teaching quality through questionnaires. Questionnaires have the same item structure, however, some items are excluded for students who did not attend courses. In any case, they were electronically administered for each teaching and all responses were kept anonymous. Only data collected from 2012 to 2017 were considered.

In general several univariate statistics are used to analyze the data and they are available at https://valmon.disia.unifi.it/sisvaldidat/unifi. In this work a CP analysis of the compositional structure is proposed in order to investigate the logconstrasts between the different characteristics of service. Proceeding in the fashion provides several advantages compared to others statistical methods: 1) it exposes the preference structure between items and allows to properly visualize it by means of *ad hoc* graphical tools and interpretational rules; 2) it always guarantees a coherent outcome whether the questionnaire is really measuring a specific construct or not; 3) the results are the same even if the number of items considered should change (subcompositional coherence). The full case study will be discussed at conference.

References

[1] Aitchison, J.: The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability, Chapman and Hall Ltd., London (1986) (Reprinted in 2003 with additional material by The Blackburn Press

[2] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal C.: Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35, 279-300 (2003).

[3] Gallo, M: Log-Ratio and Parallel Factor Analysis: An Approach to Analyze Three-Way Compositional Data. Advanced Dynamic Modeling of Economic and Social Systems, 209.

[4] Gallo, M: Tucker3 model for compositional data. Communications in Statistics - Theory and Methods, 44(21), 4441-4453.

[5] Simonacci V., Gallo, M.: Statistical tools for student evaluation of academic educational quality. Quality and Quantity, 51(2), 565-579 (2017).

[6] Simonacci V., Gallo, M.: Detecting public social spending patterns in Italy using a three-way relative variation approach. Social Indicators Research, 1-14 (2018).

Evaluation of students' performance at lower secondary education. An empirical analysis using TIMSS and PISA data.

L'impatto della scuola secondaria di I grado sul rendimento degli studenti. Un'analisi empirica con l'impiego dei dati TIMSS e PISA.

G. Graziosi, T. Agasisti, K. De Witte and F. Pauli

Abstract The present paper aims to investigate the net impact of lower secondary education on the level of literacy in mathematics, using the international assessment of educational systems provided by TIMSS in 2007 and PISA in 2012, across OECD countries and partners. We apply a pseudo–panel approach, linking PISA achievements of 15-year-old students in 2012 with those of the same cohort in the 2007 edition of TIMSS, since *the same generation* of students is taken into account. From this perspective, we are able to assess the cumulative effects of education of students at 4th grade in 2007 on performance at 9th grade in 2012.

Abstract *Il presente lavoro ha lo scopo di individuare l'impatto della scuola secondaria di I grado sul rendimento degli studenti, misurati dalle rilevazioni internazionali TIMSS e PISA, nei paesi appartenenti all'OCSE. L'analisi si basa sulla costruzione di uno pseudo–panel in cui i risultati forniti da PISA nel 2012 vengono collegati ai dati TIMSS raccolti nel 2007, poiché la stessa coorte di studenti è stata sottoposta alle rilevazioni. Da questo punto di vista, siamo in grado di valutare l'effetto cumulativo dei cicli di istruzione, dalla scuola primaria (4 grado di istruzione) alla scuola secondaria di secondo grado (9 grado di istruzione).*

Key words: Evaluation of students' performance, pseudo–panel approach, causal inference.

Grazia Graziosi
DEAMS, University of Trieste (Italy) e-mail: ggraziosi@units.it

Tommaso Agasisti
Politecnico di Milano (Italy) e-mail: tommaso.agasisti@polimi.it

Kristof De Witte
LEER, KU Leuven (Belgium) e-mail: kristof.dewitte@kuleuven.be

Francesco Pauli
DEAMS, University of Trieste (Italy) e-mail: francesco.pauli@deams.units.it

1 Introduction

Lower secondary school is a key stage of the educational path since it gives the best opportunity to strengthen basic skills and to protect students against the risk of lack of competences and educational failure in the transition between primary and upper secondary school. Preteen students go through a complex shift in their social, physical and intellectual development, as they leave childhood behind and prepare for adult responsibilities. These years are a critical point for maturation as children's roles in school and society change [4].

The present paper aims to investigate the net impact of lower secondary education on the level of literacy in mathematics, using the international assessment of educational systems provided by TIMSS[1] in 2007 and PISA[2] in 2012, across OECD countries and partners[3].

2 Methodology

We apply a pseudo–panel approach [3], linking PISA achievements of 15-year-old students in 2012 with those of the same cohort in the 2007 edition of TIMSS, since *the same generation* of students is taken into account. From this perspective, we are able to assess the cumulative effects of education of students at 4th grade in 2007 on their performance at 9th grade in 2012.

Following [1], we employ a two-step procedure:

1. For each subject i, we first estimate the impact of time-invariant observables[4] of 4th-graders, $Z_i^{(4)}$, on the performances of students' in 2007 by a linear model

$$y_{i,2007}^{(4)} = \gamma^T Z_i^{(4)} + \varepsilon_{i,2007} \tag{1}$$

where $y_{i,2007}^{(4)}$ is the score obtained by student i in TIMSS at 4th-grade and γ^T denotes the vector of coefficients attached to the time-invariant individual characteristics at grade 4.

In order to predict the test scores at grade 4 for the 9th-graders in 2012, we substitute the appropriate Z_i values for the time-invariant variables of 9th-graders, K_i, employing the vector of coefficient estimated using TIMSS data

[1] Trends in International Mathematics and Science Study

[2] Programme for International Student Assessment

[3] We collected data from the following countries: Australia, Austria, Czech Republic, Denmark, England, Germany, Hong Kong, Hungary, Italy, Japan, Kazakhstan, Latvia, Lithuania, Netherlands, Norway, New Zealand, Qatar, Slovakia, Russia, Scotland, Singapore, Sweden, Taiwan, Tunisia, United States

[4] I.e., gender, year of birth, family composition and specific items to catch the socioeconomic status of students.

$$\hat{y}^{*}_{i,2007} = \hat{\gamma}^{T} K_i^{(9)} \tag{2}$$

2. We include the predicted score at 4th grade as a proxy of the entry-level of students in secondary schools, with the aim to understand how sociodemographic characteristics affect students achievement at the beginning of upper secondary school. Therefore, we employ a model where we control for the learning gaps by the end of grade 4, including time-variant characteristics, $\hat{\gamma}^{T} K_i^{(9)}$, with time-invariant observables.

$$y^{(9)}_{i,2012} = \alpha \hat{y}^{*}_{i} + \gamma^{T} K_i^{(9)} + \beta^{T} X_i^{(9)} + u_i \tag{3}$$

where $y^{(9)}_{i,2012}$ is the score obtained by student i in PISA at 9th-grade.

Preliminary results for Italy

To verify our approach, we first run the analysis for Italy, since the recent research conducted by [2], shows that the performance of the Italian lower secondary schools is quite worrying. Further, this gap is difficult to fill at the secondary schools level, especially for students enrolled in vocational high school, and this mechanism could be responsible for disparities in both educational prospective and results of students.

Table 1 reports the time-invariant factors observed in both dataset. Unfortunately, TIMSS data at grade 4 does not include information on the educational level of parents, therefore we use the number of books at home and three specific country's item as a proxy of the indicator of sociocultural background of students', since there is a strong association among these covariates and the parents' educational attainment.

Table 1 Descriptives of the Italian sample: time-invariant factors

	TIMSS	**PISA**
N. of Obs.	4,470	31,073
Variables (%)		
Female	51.3	49.1
Born in the country	95.1	92.4
Age of arrival in Italy		
$<$ 1 year old	35.5	29.5
Btw. 1 and 5	41.5	27.2
$>$5 years old	23.0	43.3
Father born in the country	91.4	88.2
Mother born in the country	90.3	86
Books at home		
$0-10$	14.4	11.7
$11-25$	30.8	19.6
$26-100$	30.4	29.8
$101-200$	12.0	17.9
>200	12.4	21.0
Internet Connection	54.5	92.7
Air Conditioning	47.9	55.7
Alarm System	34.2	36.7

Table 2 Estimates of time-invariant observables on test scores at grade 4 (Standard error in brackets).

Variables	**Model 1**	**Model 2 (With fixed effects)**
Sex (Male=1)	-16.747^{***} (2.254)	-13.599^{***} (1.929)
Born in the Country (=1)	26.935^{*} (8.564)	28.811^{***} (7.676)
Age of arrival in Italy (*Ref. Btw. 1 and 5*)		
<1 year old	1.108 (13.282)	1.253 (11.604)
$>$5 years old	32.636^{*} (11.648)	23.026^{*} (9.983)
Father born in the country (=1)	12.031^{*} (4.843)	13.906^{**} (4.256)
Mother born in the country (=1)	1.616 (4.609)	3.971 (4.530)
Books at home (*Ref.* $0-10$)		
$11-25$	21.204^{***} (3.520)	8.970^{**} (3.044)
$26-100$	38.692^{***} (3.557)	20.975^{***} (3.108)
$101-200$	41.623^{***} (4.411)	23.766^{***} (3.865)
>200	36.641^{***} (4.380)	17.232^{***} (3.854)
Internet Connection	8.485 (2.345)	1.623 (2.103)
Air Conditioning	-5.507^{*} (2.300)	-3.350 (2.020)
Alarm System	2.219 (2.446)	.764 (2.153)
Self-perception of ability	No	Yes
% students disadv. SE background	No	Yes
Area of residence fixed effect	No	Yes
School fixed effect	No	Yes
Constant	445.230^{***} (8.588)	429.51^{***} (14.500)
# Observations	4,470	4,470
R-squared	0.062	0.378

*** $p < .001$; ** $p < .01$; *$p < .05$

Table 2 reports the estimates of time-invariant observables, $\hat{\gamma}^T$, on test scores at grade 4 (Model 1). As a robustness check we repeat the estimation including time-vaying individual variables (e.g., the self-perception of ability of students' and the percentage of students' with disadvantaged socioeconomic background), the area of residence and school fixed effects (Model 2). We find that parameters do not change significantly and the R-squared improves considerably. In order to predict the entry level of students in 2012 we only include in equation (2) the significant coefficients estimated in Model 1.

Table 3 shows the estimates of the test scores' determinants at 9th-grade, accounting for the entry level of students in secondary schools, personal and sociodemographic characteristics of students. The number of observations drops from $31,073$ to $20,320$ due to missing values. The indicators of sociocultural background is at the center of our investigation on the influence of family on achievement and educational choices. Thus, we consider the education attainment of parents as a proxy of the amount and quality of family inputs and we find that the lower educational level of parents negatively affects the performance of students. According to the relevant literature, the type of high school impacts on the achievements of students: vocations schools negatively influence students' performance, while lyceum positive affects students' results. We also include the PISA index of Economic, Social and Cultural Status (ESCS) of students finding that better status corresponds to higher performance. The analysis takes into account the confidence of students towards mathematics, measured throughout the mathematics anxiety index where

Table 3 Estimates of the test scores' determinants at grade 9

Variables	**Coeff.**	**Std. error**
Estimated entry level at grade 4	.708***	.0368
Sex (Male=1)	40.712***	1.204
Native students (=1)	7.293**	2.176
Age of arrival in Italy (*Ref. Btw. 1 and 5*)		
$<$ 1 year old	−11.989**	4.055
$>$5 years old	−17.298***	3.368
Father Education (Ref. Higher)		
Lower secondary or less	−4.294∗	1.844
Upper secondary education	2.302	1.844
Mother Education (*Ref. Higher ed.*)		
Lower secondary or less	−6.261***	1.348
Upper secondary education	−5.952***	1.133
Type of high school ((*Ref. Technical*)		
Vocational	−53.322***	1.483
Lyceum	36.276***	1.198
ESCS Index	3.865***	.7618
Math anxiety	−26.501***	.5609
Disciplined clima	8.0357***	.502
Italian Region (*Ref. Abruzzo*)		
Basilicata	−24.500***	2.585
Calabria	−53.510***	2.225
Campania	−41.456***	2.619
Emilia Romagna	19.259***	2.619
FVG	25.373***	2.626
Lazio	−21.866***	2.627
Liguria	-4.397	2.656
Lombaridia	20.816***	2.594
Marche	13.150***	2.623
Molise	−37.849***	2.864
Piemonte	13.994***	2.245
Puglia	−8.240**	2.575
Sardegna	−31.626***	2.722
Sicilia	−42.910***	2.656
Toscana	14.360***	2.698
Trentino Alto Adige	21.388***	2.364
Umbria	1.217	2.671
Veneto	36.949***	2.378
Constant	428.740***	3.509
Observations	20,320	
R-squared	0.4138	

*** $p < .001$; ** $p < .01$; *$p < .05$

higher difficulty corresponds to higher level of anxiety, that negatively affects the results of students. Moreover, the information on disciplinary climate in the classroom, based on five items, reveals that students better learn in disciplined contexts. Finally, we consider the impact of the Italian regions where students are enrolled on their performance. We observe high variability among Italian regions, in line with both national and international assessments: students attending schools located in the northern Italian regions outperform their peers enrolled in the South.

Conclusions

Gender gap: the observed gap at 9th grade has been actually generated during the lower secondary school, with respect to what we observed at 4th grade.
Achievements of immigrant: foreign-origin students lag behind Italian native students, and the gap increases from primary to secondary schools.
Family background: students with a disadvantaged background (i.e., parents with at most lower secondary education completed) score lower than their peers with higher educated parents.
The impact of the high school: students attending vocational institutes show lower performance, while the Lyceum increase the results of students. This results confirms that students' achievements are extremely diversified across types of secondary schools (Bratti et al., 2007).
Economic, social and cultural status of students (ESCS): students with a higher ESCS outperform their peers with lower ESCS.
Students' attitude towards mathematics: the anxiety towards mathematics negatively affects the performance of students.
Disciplinary climate in the classroom: students enrolled in quiet classrooms achieve better performance.
Italian Regions: the analysis highlights the variability existing among Italian regions, when the results of students are considered. According to both the National and International assessments, students in the North of Italy outperform their peers from the South.

References

1. De Simone, G. (2011). Render unto primary the things which are primary's: Inherited and fresh learning divides in Italian lower secondary education. Economics of Education Review, vol. 35, n. C, pp. 12–23.
2. Gavosto A. (2011). Rapporto sulla scuola italiana 2011, Fondazione Giovanni Agnelli, Laterza Editori.
3. Moffitt, R. (1993). Identification and estimation of dynamic models with a time series of repeated cross-sections. Journal of Econometrics, n. 59, pp. 99–124.
4. OECD, (2011). Improving Lower Secondary Schools in Norway. Reviews of National Policies for Education. OECD; Paris.

Testing for the Presence of Scale Drift: An Example

Verifica della Presenza di Scale Drift: un Esempio

Michela Battauz

Abstract The comparability of the scores is a fundamental requirement in testing programs that involve several administrations over time. Differences in test difficulty can be adjusted by employing equating procedures. However, various sources of systematic error can lead to scale drift. Recently, a statistical test for the detection of scale drift under the item response theory framework was proposed. The test is based on the comparison of the equating coefficients that convert the item parameters to the scale of the base form. After briefly explaining the methodology, this paper presents an application to TIMSS achievement data.

Abstract *La comparabilità dei punteggi è un requisito fondamentale nei programmi di valutazione attraverso test somministrati ripetutamente nel tempo. Le differenze nella difficoltà dei test si possono correggere impiegando procedure di equating. Tuttavia, diverse fonti di errore sistematico possono portare a scale drift. Recentemente, è stato proposto un test statistico per rilevare lo scale drift nel contesto della item response theory. Il test si basa sulla comparazione dei coefficienti di equating che convertono i parametri degli item nella scala di riferimento. Dopo una breve spiegazione della metodologia, questo articolo presenta un'applicazione ai dati TIMSS sull'apprendimento.*

Key words: equating, item response theory, scale stability

1 Introduction

Students' achievement level can be monitored on the basis of large scale testing programs. The fundamental requirement to guarantee a fair evaluation is the comparability of the achievement levels over different administrations. Certainly, the

Michela Battauz
University of Udine - Department of Economics and Statistics, via Tomandini 30/A 33100 Udine (Italy), e-mail: michela.battauz@uniud.it

row scores (as for example the number of correct responses) are not directly comparable because they depend on the difficulty of the test form, which can be different across the administrations. Equating is a statistical process that adjusts for differences in difficulty of the forms of a test. The literature proposes various equating methods [8], and this paper focuses on the Item Response Theory (IRT) approach. However, various sources of variability can lead to scale drift [7], causing the scores to be not comparable. A statistical test for the detection of scale drift is proposed in [3]. In this paper, the methodology is briefly explained and illustrated through an application to TIMSS achievement data.

2 Models and Methods

The 2-Parameter Logistic (2PL) model is an IRT model for dichotomous responses. The probability of a correct response to item j is modeled as a function of the ability level, θ, and the item parameters a_j and b_j

$$P(a_j, b_j | \theta) = \frac{\exp\{a_j(\theta - b_j)\}}{1 + \exp\{a_j(\theta - b_j)\}}. \tag{1}$$

The 1-Parameter Logistic (1PL) model is special case that results when the discrimination parameters a_j are equal to one (for a broad review of IRT models see [14]). These models are typically estimated using the marginal maximum likelihood method [4], which assumes a standard normal distribution for θ. For this reason, when the parameters of the model are estimated separately for different groups of subjects, the item parameter estimates are expressed on different measurement scales. The item parameters can be converted from the scale of Form $g-1$ to the scale of Form g using the following equations

$$a_{jg} = \frac{a_{j,g-1}}{A_{g-1,g}}, \qquad b_{jg} = A_{g-1,g}\, b_{j,g-1} + B_{g-1,g}, \tag{2}$$

where $A_{g-1,g}$ and $B_{g-1,g}$ are two unknown constants called equating coefficients. The literature proposes various methods for the estimation of the equating coefficients between two forms with some common items [8]. When two forms can be linked through a chain of forms, it is possible to compute the chain equating coefficients [1]

$$A_p = \prod_{g=2}^{l} A_{g-1,g}, \qquad B_p = \sum_{g=2}^{l} B_{g-1,g}\, A_{g,\dots,l}\,, \tag{3}$$

where $p = \{1, \dots, l\}$ is the path from Form 1 to Form l, and $A_{g,\dots,l} = \prod_{h=g+1}^{l} A_{h-1,h}$ is the coefficient that links Form g to Form l. Each path that links two forms yields a different scale conversion. The differences are due to sampling variability or to systematic error. Since the latter can lead to scale drift, the detection of differences in the scale conversion that can not be attributed to random error, indicates the presence of scale drift. The proposal for the detection of scale drift in [3] is a test with null hypothesis

$$H_0 : \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \cdots = \begin{pmatrix} A_p \\ B_p \end{pmatrix} = \cdots = \begin{pmatrix} A_P \\ B_P \end{pmatrix} \tag{4}$$

and as test statistics

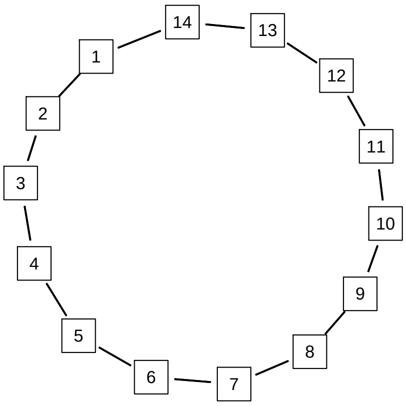
$$W = (\mathbf{C}\hat{\beta})^\top (\mathbf{C}\Sigma\mathbf{C}^\top)^{-1}\mathbf{C}\hat{\beta}, \tag{5}$$

where $\hat{\beta} = (\hat{A}_1, \ldots, \hat{A}_P, \hat{B}_1, \ldots, \hat{B}_P)^\top$, Σ is the covariance matrix of $\hat{\beta}$, $\mathbf{C}$ is a block diagonal matrix composed of two blocks both equal to $(\mathbf{1}_{P-1}, -\mathbf{I}_{P-1})$, $\mathbf{1}_{P-1}$ denotes a vector of ones with dimension $P-1$, and $\mathbf{I}_{P-1}$ denotes the identity matrix with dimension $P-1$. The covariance matrix can be computed using the delta method, considering that the equating coefficients are a function of the item parameter estimates in different administrations. Under the null hypothesis, the test statistic follows asymptotically a Chi-square distribution with $2 \times (P-1)$ degrees of freedom.

3 An Example

To illustrate the application of the procedure we used data collected for TIMSS 2011, considering achievement data in Mathematics of students at the fourth grade in Italy. Students were administered one of 14 forms (booklets). These forms present items in common as shown in Figure 1. Only dichotomous items were considered for this analysis. The total number of examinees is 3992, distributed quite uniformly between the different forms. The number of items for each form ranges between 20 and 27, while the number of common items ranges between 8 and 14.

Fig. 1 Linkage plan of the example.



The 2PL model was fit to the data of each form separately. All analyses were performed using the R statistical software [13]. The mirt package [5] was used for the estimation of the IRT models, while the equateIRT [2] package was used for the estimation of the equating coefficients.

The direct equating coefficients between forms with common items were computed using the Haebara method, and the chain equating coefficients to convert the item parameters from the scale of Form 8 to the scale of Form 1 are reported in Table 1. The two paths that connect these forms present quite different equating coefficients. Anyway, the test, also reported in the table, indicates that the differences are not statistically significant at the 0.05 level.

Table 1 Estimates of chain equating coefficients (standard errors) and scale drift test.

Path	A_p	B_p
$p = \{8, 7, 6, 5, 4, 3, 2, 1\}$	1.57 (0.36)	-0.43 (0.24)
$p = \{8, 9, 10, 11, 12, 13, 14, 1\}$	0.89 (0.19)	0.02 (0.16)
$W = 4.73$, df = 2, p-value = 0.094		

4 Discussion and Conclusions

The proposal of this paper constitutes a novel approach in the literature concerned with the detection of scale drift. While traditional methods compare the scores resulting from different administrations [9, 10, 11, 12], the approach followed here is based on the comparison of the equating coefficients. This new approach permits to formulate a statistical test for the detection of scale drift, thus allowing to take into account the presence of random error.

If the test indicates that the scale conversions deriving from different paths are different, it is then necessary to investigate which items are responsible of the drift. This can be performed using tests for the detection of differential item functioning between pairs of forms [6]. After removing these items, the test for the detection of scale drift can be performed again to verify if the scale conversions can be considered equal.

References

1. Battauz, M.: IRT test equating in complex linkage plans. Psychometrika. **78**, 464–480 (2013)
2. Battauz, M.: equateIRT: An R package for IRT test equating. Journal of Statistical Software.**68**, 1–22 (2015)

3. Battauz, M.: A test for the detection of scale drift. Working paper n. 7/2017, Department of Economics and Statistics, University of Udine (2017)
4. Bock, R. D., Aitkin, M.: Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika. **46**, 443–459 (1981)
5. Chalmers, R.: mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software. **48**, 1–29 (2012)
6. Donoghue, J. R., Isham, S. P.: A comparison of procedures to detect item parameter drift. Applied Psychological Measurement. **22**, 33–51 (1998)
7. Haberman, S., Dorans, N. J.: Scale consistency, drift, stability: Definitions, distinctions and principles. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education. San Diego, CA (2009)
8. Kolen, M. J., Brennan, R. L.: Test Equating, Scaling, and Linking. Springer, New York (2014)
9. Lee, Y.-H., Haberman, S. J.: Harmonic regression and scale stability. Psychometrika. **78**, 815–829 (2013)
10. Lee, Y.-H., von Davier, A. A.: Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. Psychometrika. **78**, 557–575 (2013)
11. Li, D., Jiang, Y., von Davier, A. A.: The accuracy and consistency of a series of IRT true score equatings. Journal of Educational Measurement. **49**, 167–189 (2012)
12. Puhan, G.: Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. Applied Measurement in Education. **22**, 79–103 (2009)
13. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2017)
14. van der Linden, W. J.: Handbook of Item Response Theory, Volume One: Models. Chapman & Hall/CRC, Boca Raton (2016)

The evaluation of Formative Tutoring at the University of Padova*

La valutazione del tutorato formativo all'Università di Padova

Renata Clerici, Lorenza Da Re, Anna Giraldo, Silvia Meggiolaro

Abstract The Programme of Formative Tutoring, an integrated tutoring model to contrast drop-out and empower university students, has been experimented in Academic Year 2016-2017 in eight First Cycle Degree Courses at the University of Padova. A propensity score matching procedure has been used to build a suitable control group to evaluate the effectiveness of the tutoring. The data used to match students comes from university administrative archives and from the questionnaire submitted to all freshmen. Results show that students that attended the tutoring activities during their first year of university performed better in term of university outcomes and number of credits achieved. Results are slightly different according to the area (scientific or humanities) of the Degree Course.

Abstract *Il Programma di Tutorato Formativo, un modello di tutorato integrato per contrastare il drop-out e favorire l'empowerment gli studenti universitari, è stato sperimentato nell'Anno Accademico 2016-2017 in otto corsi di laurea triennale dell'Università di Padova. Per valutare l'efficacia del tutorato è stato costruito un appropriato gruppo di controllo utilizzando il propensity score matching. I dati utilizzati per abbinare gli studenti provengono da archivi amministrativi universitari e dal questionario compilato da tutte le matricole. I risultati mostrano che gli studenti che hanno frequentato le attività di tutorato durante il loro primo anno di università hanno conseguito risultati migliori in termini di esiti universitari e numero di crediti raggiunti. I risultati sono leggermente diversi a seconda dell'area (scientifica o umanistica) del corso.*

Key words: effectiveness, university outcomes, propensity score matching

[1] Renata Clerici, Department of Statistical Sciences, renata.clerici@unipd.it
Lorenza Da Re, Department of Philosopy, Sociology, Education and Applied Psychology, lorenza.dare@unipd.it
Anna Giraldo, Department of Statistical Sciences, anna.giraldo @unipd.it
Silvia Meggiolaro, Department of Statistical Sciences, silvia.meggiolaro@unipd.it

* This work has been founded by the University of Padova grant: PRAT2015-CPDA157974

1 The Formative Tutoring

The Programme of Formative Tutoring (Da Re et al. 2017), an integrated tutoring model to contrast drop-out and empower university students, has been extensively experimented in Academic Year 2016-2017 in eight First Cycle Degree Courses at the University of Padova. The considered Degree Courses belong to two macro scientific-didactic areas: Scientific and Humanities.

The Formative Tutoring (FT) has been offered to all first year students of the eight Degree Courses, about 1770 students (1135 in the Scientific area and 635 in the Humanities area). The program has been promoted in different forms to all students: flyers and posters were distributed in places frequented by the students during the period of completion of the enrolment procedures, e-mail was sent to all the enrolled students, presentation has been made during the days of welcome to the freshmen and during the lectures of the first semester.

The Program has foreseen some training and informative meetings divided in:

- Service tutoring: the University Student Services have met the students orienting them to the use of their resources and proposals;
- Tutoring and Peer tutoring: the students, divided into small groups, worked on transversal skills (the method and the study skills, the participation in university life, the knowledge of the academic context, the reflection about their academic and professional expectations, the ability to evaluate and self evaluate, the development of problem solving strategies, the ability to work in a group, the ability to make informed choices, etc.), with the support and coordination of a professor of the Degree Course (Academic Tutor) or of a student in the years following the first (Student Tutor). The total number of meetings for Degree Course ranged from a minimum of 10 to a maximum of 16 meetings.

The implementation of the activities of the FT, designed and prepared with the active collaboration of the Student Tutors, depends on the different scientific didactic contexts, the internal organization of each Degree Courses and its specific needs.

2 Participation to the Formative tutoring

Formative tutoring was offered to all 1770 freshman, but the actual participation was on a voluntary basis. About 42% of the students (750 students out of 1770) have been involved at least for one meeting in the FT. The students who have attended it regularly (at least one third of the activities) have been 218, around 12%, with some differences between the two macro scientific-didactic areas. From now on, we will call these students “the participants”.

Administrative data and data from a questionnaire given to all students at the moment of enrolment, can shed some light on the characteristics of the participants. A logistic regression model has been estimate to study the variables that affect participation to the FT (see Table 1).

The first result is that characteristics of the students that affect participation are different for the two macro areas, indicating different profiles of participants. Regarding the Scientific macro area, participants are characterized by having obtained higher marks at the final high school exam and have used more information sources to choose the degree course. Also practical aspects matters: students temporarily transferred in Padova have a greater propensity to participate regularly with respects to commuters. Also for the Humanities macro area a high mark at the final exam of high school increases participation, but differently from the Scientific macro area coming from a Technical Institute has a negative effect on participation. The most important factor affecting participation is the declared intention of the students to dedicate time to the study and to attend the lessons.

Table 1: *Logistic regression to model the participation to the Formative tutoring vs. non participation*

Variables	*Scientific macro area*	*Humanities macro area*
Secondary school final score	0.028***	0.04**
Enrolment after leaving secondary school (ref: Not immediately after)		
Immediately after	0.461	0.378
Secondary school (ref: High school)		
Polytechnic		-0.656**
Vocational school		-0.168
Degree course (ref: Sciences)		
Engineering	0.133	
Place of residence (ref: Live-in students)		
Commuting students	-0.731***	
Resident students	-0.924**	
Number of sources used to get information to choose the Degree Course	0.410***	
Friends as source of information to choose the Degree Course	-0.928***	
Declared intentions of attendance and commitment in the first year (ref: maxima)		
Substantial attendance and commitment		-0.823**
Low attendance but substantial commitment		-2.71***
Low attendance and commitment		-2.379***
Minimum attendance and commitment		-1.749***

*** p<0.01; ** p <0.05

In general, for both areas, there is no effect on the probability to participate of variables regarding the socio-cultural characteristics of the students' family of origin and variable related to the intentions to work during the Degree Course.

3 Evaluation of the Formative tutoring

To evaluate the effectiveness of the FT we would like to compare university outcomes of participants and non participants to the FT. But since in the previous section we have seen that participants and non participants are different in term of observed characteristics, the differences in outcomes, if any, could be attributed not to the FT but to the differences between the two groups of students. Considering the FT as a treatment (Rubin 1974, Martini and Sisti, 2009) we need to find a group of non treated students, the non-participants, comparable to the treated, the participants.

3.1 Method and results

To a build a comparable comparison group we use a matching procedure. The rationale is that each treated unit will be matched to a comparison with the same observed characteristics. In this way differences between the groups of treated and non treated with regards to university outcomes, could be attributed only to the treatment (the FT). To apply matching the underlining assumption is that all the reasons why people participate to FT are known and observed via the variables considered, in other word it's satisfied the independence of individual characteristics with respect to the presence or absence of treatment. Given that the variables considered (listed below) range from socio-economic characteristics to variables related to study motivation, we may think the unobservable part of the selection process has a negligible effect. The assumption that could be debatable is the existence of a single version of the treatment since each Degree Course organise the FT according to their needs (see section 1). Nevertheless we can considerer the FT as a whole, since the set of intervention are designed and performed with the same goals and only slightly adapted to the single context.

As regard the application of the methodology, given the high number of variable affecting the participation to the FT an exact matching in not feasible. A propensity score matching is then performed (Rosembaum and Rubin, 1983, Thoemmes and Kim 2011). The propensity score, the probability to participate to the FT conditional to a number of observed variables, has been estimated on participant and non participant with a logistic regression. The data used to match students comes from university administrative archives and from the questionnaire submitted to all freshmen. In particular, the variables used for estimating the propensity score are: gender, age, type of high school and final grade, enrolment immediately after graduation or not, score obtained in the university admission test, Degree Course, cultural and professional level of the family of origin, working status, commuting, motivation for university choice.

A nearest neighbour propensity score matching on participants and non-participants, separately for the two macro areas, has been performed. The analysis of the histograms of the estimated propensity score for participant and non participant shown a quite good common support.

The outcome variables are the distribution of university outcomes (drop-out, course change, regular students and delay) and the number of credits attained at the end of the first year. The results are shown in Tables 2 and 3.

Table 2: *Scientific macro area: university outcome on matched participants and non-participants*

Outcome	*Non-participants*	*Participants*
Drop-out	19.2	7.4
Change of Degree Course	14.9	4.3
Regular Student	55.3	74.5
Delay in the study	10.6	13.8
Number of credits	30.4*	40.7*

* Difference between the two groups significantly different from zero

Table 3: *Humanities macro area: university outcome on matched participants and non-participants*

Outcome	*Non-participants*	*Participants*
Drop-out	17.8	8.9
Change of Degree Course	5.4	3.6
Regular Student	67.9	87.5
Delay in the study	8.9	0.0
Number of credits	38.6*	60.2*

* Difference between the two groups significantly different from zero

Results show that students that attended the tutoring activities during their first year of university performed better in term of university outcomes and number of credits achieved. Results are slightly different according to the macro area (Scientific or Humanities) of the Degree Course.

For the Scientific macro area one the most relevant effect is on the outcomes: the participants have much lower rates of drop-out and change of Degree Course and a higher percentage of regularity of the studies (74.5% vs 55.3). The effect on the number of credits is present but in comparison with the other macro area it is lower (+10.3 credits). For the Humanities macro area the most relevant effect is on performance: the participants achieved a significant higher number of credits (+21.6). The differences in outcomes between participants and non- participants are, compared to the Scientific macro area lower, although still of interest.

References

1. Da Re, L. Clerici, R., Álvarez Pérez, P.R.: The formative tutoring programme in preventing university drop-outs and improving students' academic performance. The case study of the University of Padova (Italy), Italian Journal of Sociology of Education, 9 (3): 156-175 (2017)
2. Martini, A., Sisti M.: Valutare il successo delle politiche pubbliche. Il Mulino, Bologna (2009)
3. Rosembaum, P., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects, Biometrika, 70 (1): 41-55 (1983)
4. Rubin D.: Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, J. Educ. Psychol., 66 (5): 688–701 (1974)
5. Thoemmes, F.J., Kim, E. S.: A Systematic Review of Propensity Score Methods in the Social Sciences, Multivariate Behav Res, 46 (1), 90 -118 (2011)

Benefits of the Erasmus mobility experience: a discrete latent variable analysis

Benefici dell'esperienza Erasmus: un'analisi a variabili latenti discrete

Silvia Bacci, Valeria Caviezel and Anna Maria Falzoni

Abstract Internationalization of higher education has become a priority in the European education policy. For this reason, research in this area is expanding with the aim of understanding motivations and potential benefits of international mobility. In such a context, an online survey addressed to about 1,600 students with Erasmus mobility experiences was conducted by the University of Bergamo (IT). Two latent traits, that is, the impact of the Erasmus experience on the student's abilities and the student's satisfaction for this experience, are analyzed through a two-dimensional latent class Item Response Theory model under a concomitant variable approach. The twofold issue concerning the choice of the optimal number of latent classes as well as the selection of significant covariates is specially addressed.

Abstract *L'internazionalizzazione della formazione universitaria rappresenta una priorità della politica europea dell'istruzione. Per questo motivo, la ricerca in questo ambito è in espansione, al fine di comprendere le motivazioni e i potenziali benefici della mobilità internazionale. L'Università di Bergamo ha condotto un'indagine online rivolta a circa 1,600 studenti con esperienza di studio all'estero. Due variabili latenti, cioè l'impatto dell'esperienza Erasmus sulle abilità degli studenti e il grado di soddisfazione per l'esperienza vissuta, sono oggetto di analisi tramite un modello Item Response Theory bidimensionale a classi latenti con approccio di variabili concomitanti. Particolare attenzione è posta al duplice problema della scelta del numero di classi latenti e della selezione delle covariate significative.*

Silvia Bacci
Dipartimento di Economia, Università di Perugia, Via A. Pascoli 20, 06123 Perugia e-mail: silvia.bacci@unipg.it

Valeria Caviezel
Dipartimento di Scienze Aziendali, Economiche e Metodi Quantitativi, Università di Bergamo, Via dei Caniana 2, 24127 Bergamo e-mail: valeria.caviezel@unibg.it

Anna Maria Falzoni
Dipartimento di Scienze Aziendali, Economiche e Metodi Quantitativi, Università di Bergamo, Via dei Caniana 2, 24127 Bergamo e-mail: anna-maria.falzoni@unibg.it

Key words: Concomitant variable approach, Credit mobility, Graded Response Model, Item Response Theory, Latent class model

1 Introduction

Internationalization of higher education has become a priority in the European education policy. According to the strategic objectives of Europe 2020, "*EU average of at least 20% of higher education graduates should have had a period of higher education-related study or training abroad, representing a minimum of 15 ECTS credits or lasting a minimum of three months*" (EU Council of Ministers of Education, November 29, 2011). Since it began in 1987/1988, the world's most successful student mobility programme, the Erasmus programme, has provided over three million European students with the opportunity to go abroad and study at a higher education institution or train in a company [5].

In the last years, a number of studies has analysed motivations and potential benefits of international mobility. Research findings seem to show that studying abroad positively affect students personality development and improve their language competencies and soft skills (such as problem-solving and decision-making skills). Relatedly, mobility students seem to have better labour market prospects, in particular abroad ([3], [4]).

To analyse students' motivations and the fullfilment of expectations regarding the study international experience, we asked 1,576 students, enrolled in the University of Bergamo with a credit mobility experience during the a.y. from 2008/09 to 2014/15, to answer a questionnaire. Administrative information at enrolment (e.g., age, gender, field of study, etc.) is also available for each student.

In this contribution we focus on two main aspects related to the Erasmus experience: the impact of the Erasmus experience on the student's skills and the student's fulfilling the expectations for this experience. Both these elements represent latent variables, which are measured through a set of polytmously-scored items. Our aim consists in measuring these latent variables, providing evidence of one or more problematic items. Moreover, we intend to detect individual characteristics that significantly explain the level of latent variables. For these aims, we formulate a bidimensional Latent Class Item Response Theory (LC-IRT) model ([1], [2]) with a concomitant variable approach [6].

2 Erasmus mobility data: description

To assess the international experience of the students of the University of Bergamo an ad hoc questionnaire was prepared, which was organised in three sections: Decision to study abroad, International experience and Coming back. Student's individual characteristics were also collected, such as gender, parents' level of education, parents' employment status, previous international experiences of him/herself or family, student's current employment status. The survey involved all the 1,576 students, which spent one/two semesters abroad for an Erasmus or Extra EU program from a.y. 2008/09 to a.y. 2014/15; the response rate was 48.6% (766 students).

The sample includes students from all the five fields of study of the University of Bergamo: Foreign Languages (45.7%), Economics (28.6%), Human and Social Sciences (13%), Engineering (9.6%), and Law (3.1%). They are enrolled in a bachelor degree (64%), a master degree (33.8%) and in a five-years degree of study (2.2%). As far as international experience is concerned, about half of the students spent the fall semester abroad (48.2%) and the 20.4% of students the spring semester; for just less than a third of students (31.4%) the experience lasted for the whole academic year. The preferred destination was Spain (28.1%), followed by Germany (17.7%), United Kingdom (15.8%), and France (13.7%); other European countries were overall chosen by the 18.7% of students, whereas the Extra-UE destinations were USA (2.7%), China (2.1%) and Australia (1.2%).

3 The statistical model

Given a set of J ordered polytomous items for the measurement of some latent traits about the Erasmus experience of n students, let $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_s)'$ be the vector of latent variables that drive the response process and let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_s)'$ denote one of its possible realizations. Vector $\boldsymbol{\Theta}$ is assumed to have a discrete distribution with k support points, denoted by $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_k$, and mass probabilities (or weights) $\pi_{i1}, \dots, \pi_{ik}$ $(i = 1, \dots, n)$. From an interpretative point of view, each support point detects a group (latent class) of individuals that share a common level of the latent traits, whereas the corresponding mass probability denotes the probability of belonging to each latent class according to the individual characteristics.The mass probabilities are assumed to depend on a set of p individual characteristics $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})'$ $(i = 1, \dots, n)$, that is, $\pi_{iu} = p(\boldsymbol{\Theta} = \boldsymbol{\xi}_u | \mathbf{X}_i = \mathbf{x}_i)$ $(u = 1, \dots, k)$, through a multinomial logit model

$$\log \frac{\pi_{iu}}{\pi_{i1}} = \delta_{0u} + \mathbf{x}_i' \boldsymbol{\delta}_{1u}, \quad i = 1, \dots, n;\ u = 2, \dots, k,$$

where δ_{0u} and $\boldsymbol{\delta}_{1u}$ denote, respectively, the class-specific intercept term and the class-specific vector of regression coefficients.

As concerns the response process, the probability that student i $(i = 1, \dots, n)$ answers category y $(y = 1, \dots, l_j - 1)$ to item j $(j = 1, \dots, J)$ is modelled through a multidimensional Latent Class Graded Response Model (LC-GRM; see [1] and [2] and references therein for details on the model specification and the estimation):

$$\log \frac{p(Y_{ij} \geq y | \boldsymbol{\Theta} = \boldsymbol{\theta})}{p(Y_{ij} < y | \boldsymbol{\Theta} = \boldsymbol{\theta})} = \gamma_j \sum_{d=1}^{s} (z_{jd} \theta_d - \beta_{jy}), j = 1, \dots, J;\ y = 1, \dots, l_j - 1,$$

with z_{jd} dummy variable equal to 1 if item j measures latent trait θ_d and 0 otherwise, and γ_j and β_{jy} item discriminating and difficulty parameters, respectively, having the usual interpretation as in the item response theory context.

The manifest distribution of $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iJ})'$ follows as

$$p(\mathbf{Y}_i = \mathbf{y}_i) = \sum_{u=1}^{k} p(\mathbf{Y}_i = \mathbf{y}_i | \boldsymbol{\Theta} = \boldsymbol{\theta}_u)\pi_{iu}$$

where, in virtue of the local independence assumption,

$$p(\mathbf{Y}_i = \mathbf{y}_i | \boldsymbol{\Theta} = \boldsymbol{\theta}_u) = \prod_{j=1}^{J} p(Y_{ij} = y | \boldsymbol{\Theta} = \boldsymbol{\theta}).$$

In what follow we assume $s = 2$, indicating with Θ_1 the impact of the Erasmus experience on the student's skills and with Θ_2 the fulfilling the expectations of the Erasmus experience. In addition, we consider $J = 12$ items with three ordered response categories ($l_j = 0, 1, 2$) denoting increasing levels of positive impact or satisfaction; Θ_1 is measured by 7 items and Θ_2 by the remaining 5 items.

A relevant aspect with the model at issue and, more in general, with the class of multidimensional LC-IRT models, to which the LC-GRM belongs, is represented by the choice of the number k of latent classes. The main stream of the literature agrees on using the Bayesian Information Criterion (BIC) to select the value of k. In practice, a common approach consists in estimating the model for increasing values of k, being constant all the other elements of the model, and selecting that value of k corresponding to the minimum BIC. However, it has also to be taken into account that, in the presence of several covariates, the likelihood of the model and the value of BIC are affected by the selection process of covariates. In other words, if one or more covariates are eliminated from the model because of their not significance, the sequence of BIC values would be likely different with consequences on the optimal k. In addition, strategy based on the minimization of BIC index often leads to numerous and small latent classes that are not simple to interpret.

To solve the two above problems, we here propose an iterative procedure consisting in the following steps:

Step 1: estimate the multidimensional LC-GRM without covariates and for increasing values of k until the first increasing BIC is obtained. Select that value of k corresponding to the first relative decrease of BIC smaller than a given (small) threshold α, say $\alpha = 0.01$;

Step 2: given k selected at Step 1, estimate the multidimensional LC-GRM with all the plausible covariates of interest. If all the covariates are significant at a given level (e.g., 5%) stop, otherwise go on with the next step;

Step 3: estimate the multidimensional LC-GRM with covariates selected at Step 2 and for increasing values of k until the first increasing BIC is obtained. Select the value of k as in Step 1. If all the covariates are significant stop, otherwise repeat Step 3.

The procedure finishes when there are not any more changes in the value of k and in the set of significant covariates.

4 Application to Erasmus mobility data: main results

According to the iterative procedure described above, we selected a LC-GRM with $k = 3$ components. As shown in Table 1, class 1 collects students that benefited just a little from the international experience (average weight equal to 21.0%), whereas students with significant advantage from this type of experience belong to class 3 (average weight equal to 33.1%). The remaining part of students (45.9%) is allocated in class 2, showing intermediate levels on both the latent variables.

Table 1 Estimates of support points $\hat{\xi}_{ud}$ and averages of weights $\hat{\pi}_{iu}$ $(d = 1,2;\ u = 1,2,3)$

Latent variable d	Class $u = 1$	Class $u = 2$	Class $u = 3$
Impact on skills (Θ_1, $d = 1$)	-0.951	-0.092	1.043
Satisfaction (Θ_2, $d = 2$)	-0.956	-0.084	1.039
Avg. weights	0.210	0.459	0.331

In Table 2 the distribution of the statistically significant covariates is shown. Covariates related to motivations to study abroad allow us for a clear characterisation of the three classes: improving own *curriculum studiorum* (CV), improving foreign language competencies, studying the culture of a foreign country, curiosity about new challenges, and to enhance future employability characterise class 2 and, mainly, class 3, according to an increasing trend. Instead, the impact on the latent class membership of the use of interactive teaching methods is less clear. It is worth to be noted that the host country as well as many other individual characteristics are not enclosed in the subset of significant covariates.

To conclude, Table 3 shows the item parameter estimates. The most problematic aspects related to the impact of international experience are represented by the learning ability and by the motivation to study: in both cases, the difficulty parameters $\hat{\beta}_{j1}$ and $\hat{\beta}_{j2}$ $(j = 2,7)$ are definitely greater with respect to the other items; on the opposite, the adaptability $(j = 5)$ is not a problem at all. As far as the fulfilling of the expectations, the benefits of the international experience for the Italian labour market $(j = 8)$ is perceived as the most critical aspect, whereas the students' expectations in terms of personality development $(j = 10)$ are usually completely satisfied.

Table 2 Distribution of students in the latent classes, for each significant covariate (proportions)

Covariate	Class $u = 1$	Class $u = 2$	Class $u = 3$
Motivations			
Improve CV	0.349	0.453	0.640
Improve foreign language skills	0.671	0.897	0.955
Improve host country knowledge	0.221	0.441	0.702
Curiosity about new challenges	0.732	0.915	0.955
Enhance future employability	0.121	0.306	0.566
Use of interactive teaching methods			
less than in Italy	0.013	0.006	0.025
the same as in Italy	0.182	0.062	0.091
more than in Italy	0.805	0.932	0.884

Table 3 Estimates of item difficulty parameters $\hat{\beta}_{jy}$ and item discriminating parameters $\hat{\gamma}_j$ ($j = 1,\dots,12; y = 1,2$)

Item	$\hat{\beta}_{j1}$	$\hat{\beta}_{j2}$	$\hat{\gamma}_j$
Communication skills ($j = 1$)	0.000	3.497	1.000
Learning ability ($j = 2$)	2.533	6.074	0.789
Foreign language skills ($j = 3$)	-0.757	2.864	1.005
Team working skills ($j = 4$)	0.946	5.012	0.646
Adaptability ($j = 5$)	-1.402	2.365	0.928
Problem solving skills ($j = 6$)	0.254	4.211	0.830
Motivation to study ($j = 7$)	2.704	6.612	0.541
Enhance future employability in Italy ($j = 8$)	0.000	2.196	1.000
Enhance future employability abroad ($j = 9$)	-0.514	1.404	1.518
Personality development ($j = 10$)	-1.256	-0.118	3.325
Foreign language skills ($j = 11$)	-0.633	0.303	3.770
Ability to interact with foreign people ($j = 12$)	-0.530	0.419	3.672

References

1. Bacci, S., Bartolucci, F., Gnaldi, M.: A class of multidimensional latent class IRT models for ordinal polytomous item responses. Commun Stat Theory Methods **43**, 787 – 800 (2014)
2. Bartolucci, F., Bacci, S., Gnaldi, M.: Statistical analysis of questionnaires: A unified approach based on R and Stata. Chapman & Hall/CRC Press, Boca Raton (2015)
3. CHE Consult, Brussels Education Services, Centrum fur Hochschulentwicklung, Compostela Group of Universities, Erasmus Student Network: The Erasmus impact study. European Commission, Brussels (2014)
4. CHE Consult GmbH and CHE Consult Prague s.r.o.: The Erasmus impact study. Regional Analysis. European Commission, Brussels (2016)
5. European Commission: Erasmus - Facts, Figures & Trends. Publications Office of the European Union, Luxembourg (2015)
6. Formann, A. K.: Mixture analysis of multivariate categorical data with covariates and missing entries. Comput Stat Data Anal **51**, 5236 – 5246 (2007)

University choice and the attractiveness of the study area. Insights from an analysis based on generalized mixed-effect models

Silvia Columbu, Mariano Porcu and Isabella Sulis

Abstract In this work we investigate upon the determinants of students' choices to attend bachelor degree studies outside the region of residence using information provided by the Italian National Student Archive (NSA) on a cohort of students enrolled for the first time at the university in a.y. 2014/15. The aim of the analysis is twofold: (i) to suggest value-added measures of university reputation and (ii) to assess and split the role played by the field of study in determining the power to attract students from other regions. To do this, a three-level logistic regression model to deal with cross-classified units has been adopted for modelling the probability that freshmen in a given university are mover students (instead that stayer) as a function of their socio-demographic characteristics, territorial area information and other sources of heterogeneity which concern both the field of specialization and the university reputation.

Abstract In questo lavoro, considerando i dati dell'Anagrafe Nazionale Studenti (ANS) per la coorte di immatricolati nell'a.a. 2014/15, si studiano i fattori che spingono gli studenti italiani a lasciare la propria regione di residenza per intraprendere gli studi universitari. L'analisi riportata si pone due obiettivi principali: (i) suggerire delle misure di valore aggiunto della reputazione delle università e (ii) individuare l'influenza dell'indirizzo di studi nella capacità di attrarre studenti da altre regioni. Si considera un modello di regressione logistica a tre livelli, che parametrizza una struttura cross-classified, per modellizzare la probabilità che uno studente al primo anno sia fuori sede in funzione delle sue caratteristiche socio-demografiche, delle informazioni territoriali, e di altre fonti di eterogeneità relative all'area di specializzazione scelta e alla reputazione dell'università.

Key words: multilevel models, value-added, student mobility, university reputation

Silvia Columbu, Mariano Porcu and Isabella Sulis
Università degli Studi di Cagliari, e-mail: isulis@unica.it

1 Introduction

In the last years there has been an increasing request to evaluate the effectiveness of the tertiary education institutions; these evaluations have been used to support and enhance universities and curricula which satisfy specific quality standards. The development of the assessment policies in tertiary education has determined a huge contraction of curricula supplied in many universities and a redefinition of the shape of the educational offer. Among the indicators used by the government to determine the share of the yearly public financial provisions transferred to public universities are considered, also, the job placement of graduates and the universities' capability to attract students from other territories. However, students' mobility choices are related to students' socio-economic conditions and to the economic peculiarities of both the geographical areas of students' provenance and the place where the universities are located, as well as to the university reputation and effectiveness [6] [2] [8]. Recent empirical studies on student mobility in Italy, highlight the presence of a North-South divide in the way university students make their choices [7][8]. Specifically, the distribution of university students, with respect to the region where they reside (origin) and the region where they attend the university studies (destination), shows that about 1 out of 4 freshmen students in Italy move from the South to the North or Center regions, while very few move from North-Center to South.

Many authors have addressed phenomenon of intra-and international students mobility. In an economic perspective, Dotti et al. (2013) [6] state that the migration of students in Italy can be seen as a reaction to low employment rates in the provinces of origin, and is directed mainly to the provinces with better job opportunities. D'Agostino *et al.* (2016) [5] discuss the geographical dimension of students mobility in the Italian case. Enea (2016) [7] analyses the student Italian mobility case in the transition from 1st level degree to master, highlighting that students with the best educational background have the high risk to move to another university for their master studies and that, among these, about 75% chooses a northern University. Suhonen (2014) [11] proves that in the Finnish university system the distance effect is highly conditioned to the choice of specific fields of study. The geographical dimension has also been investigated by Cattaneo et al (2016) [3], who focus on the role played by transport accessibility in determining students' choices and prove that the easiness to move is a factor of attractiveness of Italian universities. There is evidence that the migration phenomenon can be linked to the quality of university research and teaching in the Italian context ([4][2](2016) [10]). In an international framework, Beine et al. (2014) [1] identified that there exists a network effect, that contributes to spread the reputation of a university and attract new students. Giambona, Porcu & Sulis (2016) [8] highlight that the determinants of students' flows among geographical areas seem more related to the opportunities offered by the destination areas than to the characteristics of their universities in terms of facilities and services.

Moving from this framework, we define *mover* those students who enroll in universities that are not located in the region where they reside and *stayer* those who come from the same region. Since the raw rate of *movers* enrolled in a university does not allow to determine which differences in university attractiveness are related to students' characteristics and/or socio-economic background of the territorial areas of the universities and which differences are related to the vocation of the universities in terms of their education offering and/or to better education opportunities and higher level of competencies, we advance a micro analysis of the determinants of students' university choices.

The aim of the analysis is to assess and split the roles played by the field of study and by the university in determining the power to attract students from other regions. To do this, a three-level logistic regression model has been adopted for modelling the probability that freshmen in a given university are *mover* students (instead that *stayer*) as a function of their socio-demographic characteristics, territorial area information and other sources of heterogeneity of the universities; this last one concern both the field of specialization and the university reputation.

2 Data description

We use the data from Italian National Students Administrative Archive (*NSA*) provided by the Ministry of Education and Research (*MIUR*) from a cohort of students' enrolled for the first time, at a public or private university located in the national territory, in the a.y. 2014/15. We consider only records of students enrolled in traditional Bachelor or in Single-Cycle (Bachelors+Masters) Degree Programs, excluding those related to e-learning degree programs (i.e. the 10 Italian universities that offer only e-learning degree programs) or to foreign students. The analysis has been bounded to 229,813 observations, belonging to 80 (61 public and 29 private) universities that provide traditional tertiary education programs in Italy. The *NSA* collects information on students' socio-demographic characteristics, information on previous studies , university curricula (i.e. degree program and university where they enroll) and proficiency in the university studies. According to their place of origin, students are classified in 106 provinces and 20 regions. Only 99 out of 106 provinces host at least one university.

In Italy, we observe at least one university per region, thus we could advance the hypothesis that all students considered in the analysis are *free movers* [8]; namely, they all had the opportunity to attend their university studies in the region where they reside but decided to move to another. Overall, 56496 *mover* students have been identified among the subset of 229,813 students. Descriptive statistics clearly show that the propensity to be a *mover* varies across universities, the kind of degree program attended, the macro-area and the region of origin and destination. The research combines personal details on students, provided by the *NSA*, with socio-economic indicators of the provinces where the universities are located, provided by the Italian National Institute of Statistics (*ISTAT*), and other information on degree programs and universities effectiveness and reputation, provided by the surveys on Graduates' Profile and Graduates' Employment condition (carried out by the *AlmaLaurea Consortium*). Therefore, the following covariates have been considered to study the determinants of students' mobility choices: a) gender (GENDER), age at enrollment (AGE), the region of residence (REGION) and the information related to high school background- such as final grade obtained at the end of high school (GRADE) and the kind of high school attained (HIGHSCHOOL) - at student level; b) at province level the youth unemployment rate (25-34 years old) (YUNEMPLOYMENT) and a normalized indicator of gross value added per capita (NGVA) have bee used to contextualize differences in the socio-economic conditions between students' provinces of residence; moreover, the information on the number of different degree programs (NCOURSES) supplied by the universities located in the province of residence has been used as a proxy of the dimension of the tertiary education supply; c) at degree program level, (DPL) the information on graduates' rate of employment one year after graduation (EmployAfterGrad-DPL), on their average wage (WageAfterGrad-DPL) and on their self-reported perception on several aspects of the universities studies (Satisfaction -DPL), have been considered to contextualize differences in degree programs effectiveness and reputation. The information on the percentage of students with both parents graduated (GradParents -DPL) has been adopted to take into account the heterogeneity of socio-economic conditions of students belonging to different degree programs. Finally (d) at the university level the same indicators used at degree program level and described at point (c) (EmployAfterGrad-UL; Satisfaction -UL; GradParents- UL) have been adopted to contextualize differences in the characteristics of the universities. The macro-area where the university is located has been used to assess the effect of geographical components.

3 Modeling approach

Let us indicate with Y_{ijg} an indicator variable which assumes value 1 if student i ($i = 1,\dots,n$) from degree program j ($j = 1,\dots,J$) of the university g ($g = 1,\dots,G$) resides in the region where the university is located (is a *mover*) and 0 if she/he resides in the same region (is a *stayer*). The probability to be a mover is modeled using a logistic function as follows

$$logit[\gamma_{ijg}] = \tau + \boldsymbol{X}_{ijg}^T\beta + \boldsymbol{Z}_j^T\gamma + \boldsymbol{U}_g^T\delta + \theta_j + \boldsymbol{\lambda}_g^T\boldsymbol{v}_g \qquad (1)$$

where $\gamma_{ijg} = \pi_{ijg}$ and $\mathbf{X}_{ijg}$ is a vector of individual covariates (Level-1), $\mathbf{Z}_j$ is a vector of covariates at degree program level j (Level-2), and $\mathbf{U}_g$ is a vector of covariates at university level (Level-3). $\boldsymbol{\theta}_j \sim N(0,\sigma_\theta^2)$ is a random term shared by observations related to the same degree program, whereas $\boldsymbol{v}_g \sim MVN(0,\Sigma_v)$ enables to account for heteroskedasticity between universities due to geographical components (e.g between macro-area propensity to attract students). The presence of both random components enables to split the between university variability from the between degree program variability in the propensity to attract students. $\boldsymbol{\lambda}_g$ is a vector which enables to identify the macro-area where the university is located (e.g considering three geographical macro areas, $\boldsymbol{\lambda}_g$ for student i assumes value 1 only for the macro-area where the university is located). The expected posterior predictions of the random terms ($\hat{\theta}_j$ and $\hat{v}_g$) enable to make inference about degree programs (i.e. degree program-university combination) and universities attractiveness. The cross-classified model described on (1) (where students are clustered in degree programs (Level-2) that belong to different universities (Level-3)) has been made hierarchical by considering at Level-2 degree program-university combinations. Thus, the variability in the propensity to attract students is split in two components: the between-university variability and the between-degree program within-university variability, allowing the divergences in $\hat{\theta}_{jg}$ to capture the gap in the attractiveness of degree program j belonging to the university g from the overall university attractiveness parameter $\hat{v}_g$. The model has been estimated with the runmlwin routine which calls MLwiN scripts from Stata by adopting Monte Carlo Markov Chain algorithm (Leckie & Charlton (2013)) [9].

4 Model Results

A model building strategy has been carried out to select relevant predictors at each level of analysis. Starting from the null model, relevant assumptions for the definition of the model structure have been tested. The main findings suggest that the variance of the random terms across geographical areas is homoskedastic. Thus the random component $\boldsymbol{v}_g$ has been considered univariate. Results depicted in Table 1 show that `AGE` and `HIGHSCHOOL` attended influence the probability to be a *mover*. Namely, it is higher for older students who attended a classical or scientific lyceum and come from southern regions (with the exception of Aosta Valley) such as Basilicata, Calabria, and Apulia, while it is lower for younger students who attended a professional school and come from northern and center regions such as Emilia Romagna, Tuscany and Lombardy. *Movers* come mainly from provinces with disadvantage economic conditions, as it is shown by the coefficient of both socio-economic indicators `YUNEMPLOYMENT` and `NGVA` addressed to depict the socio-economic framework. The dimension of the tertiary education supply in the origin provinces does not seem to be a detterent to students' propensity to move. Looking at the geographical macro-areas where the universities are located, we can observe that *movers* prefer to study in universities of the north or center of Italy. The employment rate one year after the graduation in a given university seems to be a pull factor of mobility. The characteristics of the degree program highlight that the propensity to be a *mover* is higher in the degree programs attended by students coming from more educated families (`GradParents -DPL`), and lower in those where the perceived student satisfaction towards the university studies is higher (`Satisfaction-DPL`). The variances of the random terms show that about 70% of the variability in the propensity to be a *mover* is explained at University level, whereas differences between degree programs within each university, even if significant, have a lower effect ($\sigma_v^2 = 2.84$ vs $\sigma_\theta^2 = 0.335$).

Table 1 Multilevel logistic regression estimates

Fixed Effects	Coeff.	Std. Error	p-value	Fixed Effects	Coeff.	Std. Error	p-value
Students' characteristics				**Characteristics of the Territorial areas of origin**			
GENDER (Baseline=Male)				REGION (Baseline=ABRUZZ0)			
Female	0.015	0.0151		BASILICATA	5.833	0.0973	***
AGE	0.037	0.0019	***	CALABRIA	4.417	0.0948	***
HIGHSCHOOL(Baseline= Other Lyceum)				AOSTA VALLEY	3.427	0.155	***
Classic Lyceum	0.305	0.0244	***	APULIA	2.747	0.0598	***
Foreign School	0.072	0.0486		SICILY	2.273	0.0705	***
Professional School	-0.141	0.0345	***	MOLISE	1.933	0.0802	***
Scientific Lyceum	0.125	0.0205	***	SARDINIA	1.882	0.0931	***
Technical School	-0.102	0.0244	***	CAMPANIA	0.972	0.0565	***
				LIGURIA	0.660	0.0647	***
GRADE	0.006	0.0006	***	UMBRIA	0.228	0.0693	***
University characteristics				TRENTINO ALTO ADIGE	-0.137	0.069	**
MACRO AREA(Baseline=CENTER)				PIEDMONT	-0.429	0.0555	***
NORTH	0.206	0.117	*	MARCHE	-0.576	0.0514	***
SOUTH	-2.525	0.230	***	FRIULI VENEZIA GIULIA	-0.681	0.0699	***
EmployAfterGrad-UL	0.017	0.002	***	VENETO	-0.767	0.0545	***
Degree Program characteristics				LAZIO	-1.156	0.0463	***
GradParents-DPL	0.010	0.004	**	LOMBARDY	-1.548	0.0488	***
Satisfaction-DPL	-0.020	0.003	***	TUSCANY	-1.637	0.0564	***
EmployAfterGrad-DPL	0.0000766	0.001		EMILIA ROMAGNA	-2.172	0.0525	***
WageAfterGrad-DPL	-0.0000307	0.000		YUNEMPLOYMENT	0.011	0.0023	***
				NGVA	-0.024	0.000728	***
				NCOURSES	-.0000584	0.0005	
Random effects				**Goodness of fit**			
	Sd	SE	p-value				
University	2.840	0.516	***	DEVIANCE	159711.78		
Degree (by university)	0.335	0.018	***	DIC	160820.64		

*** $p < 0.01$, **$p < 0.05$, *$p < 0.1$

References

1. Beine, M., Noe, R., Ragot, L.:Determinants of the international mobility of students. Economics of Education Review, 41, 40-54 (2014)
2. Bratti, M. and Verzillo, S., The 'Gravity' of Quality: Research Quality and Universities' Attractiveness in Italy. IZA Discussion Paper No. 11026 (2017)
3. Cattaneo, M., Malighetti, P., Paleari, S., Redondi, R.:The role of the air transport service in interregional long-distance students' mobility in Italy. Transportation Research Part A, 93, 66-82 (2016)
4. Ciriaci, D.: Does University Quality Influence the Interregional Mobility of Students and Graduates? The Case of Italy, Regional Studies, 48(10), 1592-1608 (2014)
5. D'agostino A., Ghellini G., Longobardi S. : University mobility at enrollment: geographical disparities in Italy. Atti della XLVIII Riunione Scientifica SIS 2016, Salerno 8-10 giugno (2016)
6. Dotti, N.F., Fratesi, U., Lenzi, C. and Percoco, M. :Local Labour Markets and the Interregional Mobility of Italian University Students, Spatial Economic Analysis. Taylor and Francis Journals, 8 (4), 443-68 (2013)
7. Enea, M.: From South to North? Mobility of Southern Italian students at the transition from the first to the second level university degree. Paper presented at the 48th Scientific meeting of the Italian Statistical Society (2016)
8. Giambona, F., Porcu, M., Sulis, I.: Students Mobility: Assessing the Determinants of Attractiveness Across Competing Territorial Areas. Soc Indic Res, 133(3) , 1105-1132 (2017)
9. Leckie, G. and Charlton, C.: runmlwin - A Program to Run the MLwiN Multilevel Modelling Software from within Stata. Journal of Statistical Software, 52 (11),1:40 (2013)
10. Pigini, C., and Staffolani, S. Beyond participation: Do the cost and quality of higher education shape the enrolment composition? The case of Italy. Higher Education, 71(1), 119142 (2016)
11. Suhonen, T.: Field-of-Study Choice in Higher Education: Does Distance Matter?, Spatial Economic Analysis, 9(4), 355-375 (2014)

Environment

The climate funds for energy sustainability: a counterfactual analysis

I Fondi per il clima a sostegno della sostenibilita' energetica: una analisi di impatto

Alfonso Carfora and Giuseppe Scandurra

Abstract In this paper we analyze the effectiveness of climate funds to combat climate change and promote mitigation and adaptation policies. We analyse the funds received by the recipient through a counterfactual analysis. The results show that the policy contributed to the decreasing of greenhouse gas emissions and promoted the change in generation energy systems supporting the replacement of fossil sources with renewable sources.
Abstract Si vuole analizzare l'efficacia dei fondi per il clima nella lotta al cambiamento climatico e per la promozione di politiche di mitigazione e adattamento. Sono analizzati i fondi ricevuti dai Paesi beneficiari attraverso un'analisi controfattuale. I risultati mostrano che questa politica ha contribuito alla riduzione delle emissioni di gas serra ed ha promosso il cambiamento nei sistemi di generazione elettrica favorendo la sostituzione delle fonti fossili con fonti rinnovabili.

Key words: climate finance, propensity score matching

1 Introduction

During the 15th Conference of the Parties (COP15) held in December 2009 in Copenhagen, developed countries pledged to provide new and additional resources to combat climate change, approaching USD 30 billion for the 2010-2012 period, with balanced allocation between mitigation and adaptation

Alfonso Carfora
Department of Management and Quantitative Sciences, via Generale Parisi, 13 Napoli
e-mail: alfonso.carfora@uniparthenope.it

Giuseppe Scandurra
Department of Management and Quantitative Sciences, via Generale Parisi, 13 Napoli
e-mail: giuseppe.scandurra@uniparthenope.it

strategies. This collective commitment is known as fast-start finance and prefigures the institution of Green Climate Fund (GCF) established by the 194 countries that are members of the United Nations Framework Convention on Climate Change (UNFCCC) in 2010, to support a paradigm shift in the global response to climate change. Through the GCF mechanism, donor governments distribute funds to recipient developing countries to finance low-emissions and climate-resilient projects and programs in these countries. As they are proliferating, the challenges of coordinating funds and the monitoring of recipient countries emissions became an important matter to assess their effectiveness. In this paper, we want to evaluate the impact of the climate funds distributed by donor countries on environmental and economic factors analyzing the flow of funds among countries and conducting a counterfactual analysis. To achieve our aims, we employ propensity score matching (PSM) analysis on a large dataset of 149 countries.PSM is a statistical method which make the construction of a probabilistic match among units that have participated in a treatment (treated) and units that have not participated (untreated), utilizing characteristics that are common to both groups [1]. The remainder of the paper is organized as follows: Section 2 describes the data. Section 3 reports the empirical results while in Section 4 we discuss the results. Finally Section 5 offers some concluding remarks.

2 Data

To assess the impact of fast-start finance, we use the AidData Research Release 2.1 database that is based on the Credit Report System database, managed by the OECD's Development Assistance Committee (DAC).We consider the funds for energy generation and supply by renewable sources and the flows of funds targeted at biosphere protection using a dataset of 149 countries. It considers the totality of countries eligible to receive funds according to the OECD's Official Development Assistance (ODA) list. Dataset includes countries that have received funds in 2010 (treated –- 83 countries) and those that did not receive funds (untreated –- 66 countries). Explanatory variables can be grouped as target and control indicators.
Among the target variables, we include i) the share of renewable energy in the total energy generated (shren), ii) GDP per capita (gdp), iii) CO_2 per capita emissions (CO_2) and iv) the share of fossil energy in the total energy generated (shfoss).

In the group of control variables, we consider those typically indicated by the previous literature [2] as key factors that drive countries toward increasing generation from renewable energy sources: electricity consumption, the oil supply, energy intensity, the female population and the population growth rate.

3 Empirical Results

The coefficients of the propensity score probit model are reported in Table 1.

Table 1 Coefficients and goodness of fit statistics of the probit propensity score model

Variables	Coefficients
Intercept	1.5527
Electricity Consumption	0.3752^a
Oil Supply	-0.0003^c
Energy Intensity	-0.4111^c
Female	0.0207
Population Growth	0.8419^b
Electricity Consumption Squared	0.0225
Population Growth Squared	-0.2019^b
Loglikelhood	-73.7389
Pseudo R^2	0.3185

Significance: ^{a} 0.01; ^{b} 0.05; ^{c} 0.1

Climate funds are more attractive for countries characterized by increasing population growth rates, even though at decreasing marginal rate (because the second order coefficient is negative), and high levels of energy consumption. There is no empirical evidence that climate funds are attractive to countries where the female population composition is higher, probably because in the developing countries the women feeling for environmental issues are less consolidated. By contrast, oil-exporting countries and those that are more oriented toward the use of traditional energy sources (high energy intensity) prove to be more resistant to these types of policies in support of renewable energy generation because they imply structural changes in their industrial structures and economic systems that are generally well-developed. The results of the probit model confirm several consolidated issues and they represent an important starting point for the next step of the work, which concentrates on the analysis of impact of the funds on countries that have obtained them. Moreover, the matching performed using the fitted values of the model (the propensity scores) ensures that the similarities between matched countries are respected: the average values of the control variables of the untreated countries are not significantly different from those of the countries to which they have been matched (Table 2 column 2). The matching is obtained using the nearest neighbor (1) algorithm that provides a one-to-one matching setting the caliper threshold equal ton 0.25 [3].

Table 2 Tests of balance: similarities of means of the control variables before and after matching

Variable	Before Matching	After Matching
	Electricity Consumption	
Mean Treated	2.2136	2.2136
Mean Untreated	0.7167	2.2912
p-value	0.0001	0.7083
	Oil Supply	
Mean Treated	407.1200	407.1200
Mean Untreated	535.0500	220.1200
p-value	0.6155	0.1324
	Energy Intensity	
Mean Treated	8.4311	8.4311
Mean Untreated	8.5444	8.6231
p-value	0.4435	0.1874
	Female	
Mean Treated	50.2560	50.2560
Mean Untrated	49.0910	50.3640
p-value	0.0697	0.4943
	Population Growth	
Mean Treated	1.6457	1.6457
Mean Untreated	1.7591	1.7338
p-value	0.6837	0.5587
	Electricity Consumption Squared	
Mean Treated	9.6207	9.6207
Mean Untreated	5.8147	8.6408
p-value	0.0162	0.3964
	Population Growth Squared	
Mean Treated	3.6393	3.6393
Mean Untreated	7.3523	4.0908
p-value	0.0709	0.3631

4 Discussion

The treatment effect on treated (ATT) represents a comparison between the observed values and the expected values of the target variables for the treated countries if they had not participated in the treatment. Countries that have received funds, in fact, are similar, in terms of the control variables, to the countries that have not received funds to which they have been matched. However, they are different in terms of the target variables, and the basic hypothesis is that this difference is due to the treatment. Table 3 reports the values of the estimated ATT.

Table 3 Average treatment effects on treated

Variables	ATT
Shren	0.1872^b
Shfoss	-0.1670^c
CO_2	-2.8205^b
GDP	$1,344.3^b$

Significance: ^{a} 0.01; ^{b} 0.05; ^{c} 0.1

In terms of CO_2 emissions , without the funds, there would have been no differences between treated countries and their similar matched countries. Instead, the significant reduction of about 2.8 metric tons in the CO_2 per capita emissions of treated countries is a result that is in line with the global climate finance architecture. This result suggests that climate finance mechanisms, in fact, are useful to enhance the efforts to reduce emissions. Focusing on per capita GDP, we observe another difference between treated and untreated countries. On average, the GDP of countries that have received funds increases approximately 1,340 USD with respect to that of the counterfactual part. This result confirms those of several recent studies on the positive effects of climate financing on the economies of developing countries [4] with renewable energy being a crucial component for the economic growth of developing countries [5]. Observing the estimation results, we note that treated countries have, on average, a share of energy produced by renewable sources (shren) that is significantly higher with respect to their similar untreated countries by approximately 19%. Complementarity, the share of energy produced by fossil fuel (shfoss) is significantly lower, on average, by approximately 17% with respect to the counterfactual part of countries. This important result suggests that climate finance can help countries increase investments in RES generation and can substitute for fossil power generation. Moreover, this result indicates that the climate funds help to change the electricity basket generation, increasing the share of RES generation in place of fossil fuel generation .

5 Conclusion

The results obtained in this paper provide clear indications on the effectiveness of climate funds in promoting the green growth. The results show that funds have been devoted to enhance energy efficiency and sustainability: the recipient countries, in fact, reduced their GHG emissions respect to their similar counterparts. The factors that explain this empirical result are the positive consequence of the policies implemented in the last years in which the need to reach the targets imposed by the climate finance leaded them toward an increasing attention for environmental issues. Moreover, the results show a decrease in the shares of electricity generated by fossil fuels and an increase in RES generation. In particular in these countries, we observe that the decrease of the generation of energy by fossil sources is balanced by the increase of the generation of RES. The increase in GDP per capita occurs in the recipient countries, with respect to the counterfactual part. Climate funds can be considered helpful instruments to promote the path towards a sustainable energy system, based on a high share of RES generation, for developing countries. In order to ensure the efficient funding allocation, policy makers have to regularly monitor the achieved results of financed projects. Our results could support the Ad Hoc Working Group on the Paris Agreement (APA) in monitoring the progresses made to reach the goals of the climate funds. Moreover, the findings provide a starting point to plan environmental policies to be undertaken in preparation to the full implementation of the Paris Agreement. The analysis carried out shows that the beneficiary countries have increased the share of electricity from renewable sources and have reduced the share of electricity from fossil ones, finding it more useful and advantageous to replace them with renewable ones. However, funding should be better targeted.

References

1. Rosenbaum, P. R., Rubinm D. B.: The Central Role of the propensity score in observational studies for causal effects, Biometrika 70, 41-–55 (1983).
2. Marques, A.C., Fuinhas, J. A., Pires Manso, J. R.: A quantile approach to identify factors promoting renewable energy in European Countries. Environmental and Resources Economics 49, 351–366 (2011).
3. Austin, P.C.: An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies, Multivariate Behavioral Research, 46, 399-–424 (2011).
4. Jakob, M., Steckel, J.C., Flachsland, C., Baumstark, L.: Climate Finance for Developing Country Mitigation: Blessing or Curse?, Climate and Development, 7, 1–15 (2015) .
5. Saidi, K., Ben Mbarek, M.: Nuclear energy, renewable energy, CO2 emissions and economic growth for nine developed countries: Evidence from panel Granger causality tests, Progress in Nuclear Energy 88, 364–374 (2016).

Exploratory GIS Analysis via Spatially Weighted Regression Trees

Analisi esplorativa di dati GIS mediante alberi di regressione pesati spazialmente

Carmela Iorio, Giuseppe Pandolfo, Michele Staiano, and Roberta Siciliano

Abstract The challenging goal of the *Moving Towards Adaptive Governance in Complexity: Informing Nexus Security* (MAGIC) project is to quantitatively enlighten the nexus among energy, food, water and land use toward informed governance inside EU aimed at long term environmental feasibility, economic viability and social desirability. Within the framework of recursive partitioning algorithms by tree-based methods, this paper provides an application on a real Geographic Information System (GIS) dataset regarding the irrigation communities in Almeria, Spain. We propose to build an explorative regression tree – spatially weighted – aiming at classifying the specific consumption of water (per hectare) of the farming communities based on either water management areas and different mix of sources for irrigation water (surface, groundwater, waste water, desalination).

Abstract *Il progetto Moving Towards Adaptive Governance in Complexity: Informing Nexus Security (MAGIC) mira ad illustrare quantitativamente il nesso tra consumi di energia, cibo ed acqua e l'uso del suolo nell'Unione Europea, con lo scopo di sostenerne i processi decisionali orientati alla compatibilità ambientale, fattibilità economica e desiderabilità sociale di lungo periodo. Nel presente lavoro è descritta un'applicazione a dati reali, raccolti in un sistema informativo geografico (GIS) relativo alle comunità irrigue dell'Almeria, inquadrata nell'ottica dei metodi*

Carmela Iorio
Department of Industrial Engineering, University of Naples Federico II, Napoli e-mail: carmela.iorio@unina.it

Giuseppe Pandolfo
Department of Industrial Engineering, University of Naples Federico II, Napoli e-mail: giuseppe.pandolfo@unina.it

Michele Staiano
Department of Industrial Engineering, University of Naples Federico II, Napoli e-mail: michele.staiano@unina.it

Roberta Siciliano
Department of Industrial Engineering, University of Naples Federico II, Napoli e-mail: roberta@unina.it

di partizione ricorsiva su alberi. La proposta consiste nella costruzione di un albero di regressione esplorativo opportunamente pesato, per classificare il consumo specifico di acqua (per ettaro di superficie irrigua) nelle diverse comunit agricole della regione in base alle aree di appartenenza ed al differente mix di fonti.

Key words: Regression trees, Supervised statistical learning, Spatial methods

1 Motivation

The MAGIC project aims at suitably tackling the nexus among energy, food and water to assess the sustainability as a complex predicate. It features a novel perspective rooted in bioeconomics toward the accounting of technical and environmental resources required to assure the living standards of our societies. In order to inform and steer the processes of decision and policy making inside UE, quantitative contents of narratives about the various theme related to the project – water, energy, food accounts along with labour and land use entanglement, their nexus and scenarios for innovations in the current state of affairs – are to be produced by means of a rigorous and transparent approach to official data, specific domain knowledge and agreed models. Therefore, statistical learning is envisioned as a key tool to lift data to information and up to knowledge, as needed to cope with the complex predicament of sustainability.

Among a set of pilot case studies approached during the first year of the project, some real datasets were exploited as test bench for the fruitful interaction of domain experts and statisticians.

2 A short introduction to exploratory regression trees

Classification And Regression Trees (CART) developed by [1] is a milestone in the evolution and spreading of the tree-based methodology. Recursive partitioning tree procedures adopt a supervised approach: the response variable (of numerical or categorical type) drives the learning process on the basis of a set of predictors (of numerical or categorical type). For this reason CART methodology can be viewed as precursor of supervised statistical learning introduced by [10] and outlined by [5] . Tree-based methods have been proposed for both prediction and exploratory purposes. In the exploratory context, binary segmentation can be understood as a recursive partitioning of objects into two subgroups due to some splitting variables derived from available predictors such to obtain internally homogeneous and externally heterogeneous subgroups with respect to a target or response variable [6]. The final result is a binary tree visualizing the dependence relationship between the response variable and the predictors.

In regression trees, the splitting criterion at each non terminal node can be to maximize the decrease of impurity of the response variable within the two sub-nodes, where the impurity is a measure of variation for numerical responses. Nodes are declared to be terminal on the basis of a stopping rule. Terminal nodes include disjoint and homogeneous subgroups of objects, defining a partition of the starting group of objects at the root node with respect to the response variable. The tree with only one split at the root node is called stump. It describes the best partition of the objects into just two subgroups such to minimize the internal variation of the target variable within the two nodes. The quality of any tree can be measured by an overall impurity measure of the tree. In regression trees, this is calculated by the sum of the variation measures of the target variable within its terminal nodes. Typically, the overall impurity of any tree is compared with respect to the impurity at the root node: the ratio of the two quantities provides a relative cost reduction measure named deviance. By definition of splitting criterion, the deviance reduces as the number of terminal nodes or tree size increases. For a more detailed literature study one can refer to [7, 8, 9, 3, 2] and to the references therein.

3 A Real Problem

The analysis has been performed on a data set collected in 2013 and related to the all the irrigation communities in Almeria which are grouped geographically by 18 water management areas and technically by 38 distinct water sources patterns. Raw GIS data matrix consists of 376 instances, designated as irrigation communities. As the entire population is surveyed, the analysis is only exploratory, not confirmatory nor predictive. Exploratory trees belong to data mining methods where also visualization helps the analyst to better understand the phenomena [4]. We built an exploratory regression tree spatially weighted aiming at classifying the specific consumption of water per hectare of the irrigation communities base on either water management area (coded as in Table 1), and sources of water used (surface, groundwater, waste water, desalination), grouped in 8 different profiles described in Table 2. An expanded regression tree with the purest terminal nodes, accordingly to an initial stopping rule, has 48 leaves. Since this structure is rather complex to be interpreted, a simpler structure of the regression tree can be identified. One can fix

Table 1: Water Management Area (WMA) codes

WMA	**Code**	**WMA**	**Code**	**WMA**	**Code**
Alpujarra	a	Campo de Tabernas	g	Medio Almanzora	m
Alto Almanzora	b	Comarca de Guadix	h	Medio Andarax	n
Alto Andarax	c	El Saltador	i	Nacimiento	o
Bajo Almanzora	d	Higueral de Tjola	j	Poniente	p
Bajo Andarax	e	Los Guiraos	k	Riegos de Pulp	q
Campo de Njar	f	Los Vlez	l	Z.R. Cuevas de Almanzora	r

Table 2: Profiles codes

Profile	**Code**
80-99% surface, remaining groundwater	1
surface water only	2
groundwater only	3
60-80% surface, remaining groundwater	4
40-60% surface, remaining groundwater	5
10-39% surface, remaining groundwater	6
1-10% surface, remaining groundwater	7
remaining profiles (reused and desalted)	8

a certain threshold value for the overall impurity, then select among the sub-trees not exceeding the fixed threshold value the one with minimum number of leaves. Specifically, Fig. 1 shows that the deviance reduction is marginal for the trees with more than 9 leaves. Consequently, we chose the regression tree with 9 terminal nodes as final tree. We named it "intermediate regression tree" and its representation is displayed in Fig. 2. The labels set in Fig. 2 at any internal node indicate the splitting variables with those categories inducing the objects to the left sub-node; at the leaves of the tree are reported the mean values of the target variable (cubic meters of water consumed in one year per hectar of farming surface). Fig. 2 points out that the mean response values increase from the left-hand side to the right-hand side. This clearly conveys to the domain expert the information that the irrigation communities belonging to the leftmost leaves consume less water per hectare than those belonging to the rightmost terminal nodes. By watching the mean fitted values of the rightmost terminal leaves it can state that the WMA belonging to both Campo de Nijar and Poniente are the more water intensive consumption areas. Specifically, the irrigation communities belonging to Poniente has the highest mean value of water use per hectare. The simplest result of analysis reduces the tree to a stump with the only splitting rule due to WMA. Thus, the stump partitions the irrigation communities into two regions of predictor space. The resulting partition singles out the most water intensive management areas. On the right-hand side, there are the management areas with lower values of water consumed per hectare (mean value equals to 3377); on the left-hand side, there are the remaining management areas, Campo de Njar and Poniente, with higher values of water consumed per hectare (mean value equals to 6199). Thus, the stump reveals that the most important factor associated with water consumption per hectare is due the water management areas. Fig. 3 displays a geographic visualization of the stump (the map was produced elaborating the R output by open source GIS software Q-GIS and includes a satellite imagery title obtained by BING).

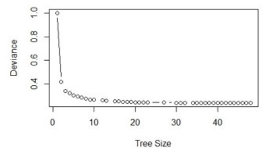


Fig. 1: The overall deviance reduction of the regression tree with increasing the number of leaves

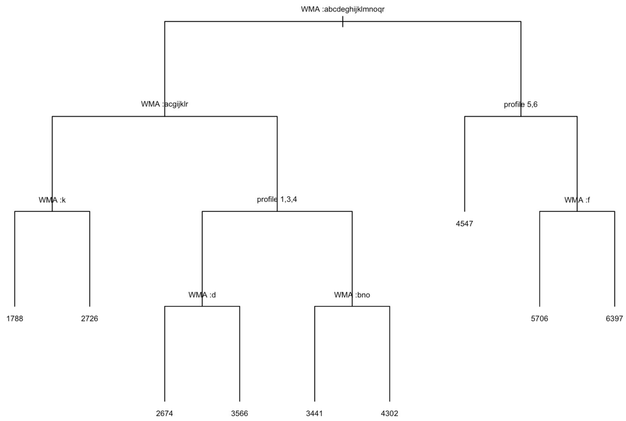


Fig. 2: The intermediate regression tree for Almeria data set.

4 Final remarks

In the framework of MAGIC project, this paper was designed to deal with a real problem of statistical analysis. We analyzed the water consumption by irrigation communities in Almeria. The regression tree is a device simple to be presented and understood; nevertheless, it required a careful tailoring of the standard approach (being the observations to be spatially weighted) and a trade off between easiness and richness of the representation should be agreed in order to choose the number of leaves. The "intermediate regression tree" enables to highlight some key features in the data set. Indeed, the leaves resulted ordered from left to right hand side of the tree with increasing levels of average water consumption per hectare: this classification is a good starting point for deepening the analysis. The water consumption per hectare in some management areas is quite more spread than in some others; this clearly depend on the types of crops and cropping methods along with the irrigation means and water sources, so grafting current data to other information could offer a

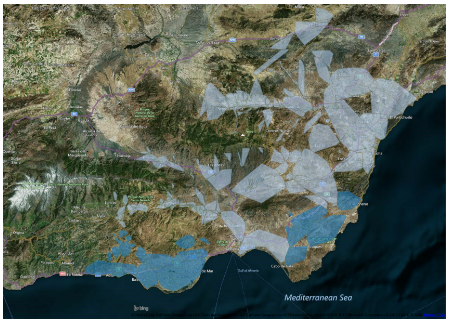


Fig. 3: GIS visualization of the STUMP (the blue areas mark the WMAs with higher consumption of water per hectare Campo de Nijar and Poniente compared to the gray ones).

better insight. The role of profiles of sources varies in different sets of water management areas, so we could technically say that the two predictors interact: this is the clue for a further investigation. A richer GIS approach to the problem could be beneficial for the analysis, so collecting more layers of data and comparing spatially any pattern could strenghten the knowledge discovery process.

Acknowledgements MAGIC project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 689669.
The present work reflects only the authors' view and the Funding Agency can not be held responsible for any use that may be made of the information it contains.

References

1. Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J.: Classification and regression trees. CRC press (1984).
2. D'Ambrosio, A., Aria, M., Iorio, C., Siciliano, R.: Regression trees for multivalued numerical response variables. Expert Systems with Applications **69**, 21–28 (2017)
3. D'Ambrosio, A., Heiser, W.J.: A recursive partitioning method for the prediction of preference rankings based upon kemeny distances. Psychometrika **81**(3), 774–794 (2016)
4. Fayyad, U.M., Wierse, A., Grinstein, G.G.: Information visualization in data mining and knowledge discovery. Morgan Kaufmann (2002).
5. Friedman, J., Hastie, T., & Tibshirani, R.: The elements of statistical learning. New York: Springer series in statistics (2001).
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media (2009)
7. Mola, F., Siciliano, R.: A two-stage predictive splitting algorithm in binary segmentation. In: Computational statistics, pp. 179-184. Springer (1992)
8. Mola, F., Siciliano, R.: A fast splitting procedure for classification trees. Statistics and Computing **7**(3), 209–216 (1997)
9. Siciliano, R., Mola, F.: Multivariate data analysis and modeling through classification and regression trees. Computational Statistics & Data Analysis **32**(3), 285–301 (2000)
10. Vapnik, V.N.: The nature of statistical learning theory (1995).

A functional regression control chart for profile monitoring

Carta di controllo mediante regressione funzionale per il monitoraggio di profili

Fabio Centofanti, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo and Simone Vantini

Abstract In many applications, profile monitoring techniques are needed when the quality characteristic under control can be modeled as a function. Moreover, measures of other functional covariates are often available together with the functional quality characteristic. To combine the information coming from all the measures attainable, a new functional control chart is proposed for profile monitoring. It relies on the residuals of a function-on-function linear regression of the quality characteristic on the functional covariates. The effectiveness of the proposed monitoring scheme is illustrated on a real-case study about the monitoring of CO_2 emissions from a Ro-Pax ship owned by the shipping company *Grimaldi Group*.

Abstract *In molte applicazioni è necessario utilizzare metodi di controllo statistico di processo a caratteristiche di qualità modellabili come funzioni. Inoltre, è spesso possibile integrare nello schema di monitoraggio anche la disponibilità di osservazioni di covariate funzionali ad esse correlate. In questo lavoro, viene presentata una carta di controllo funzionale basata sui residui di un modello di regressione lineare funzionale della caratteristica di qualità sulle covariate ad essa associate. Le potenzialità dell'approccio proposto, vengono illustrate mediante un caso studio sul monitoraggio delle emissioni di CO_2 di una nave da carico e passeggeri.*

Key words: statistical process control, profile monitoring, functional data analysis, functional linear regression

Fabio Centofanti, Antonio Lepore and Biagio Palumbo
Department of Industrial Engineering, University of Naples Federico II, P.le V. Tecchio 80, 80125, Naples, Italy
e-mail: fabio.centofanti@unina.it, antonio.lepore@unina.it, biagio.palumbo@unina.it

Alessandra Menafoglio and Simone Vantini
MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy e-mail: alessandra.menafoglio@polimi.it, simone.vantini@polimi.it

1 Introduction

New statistical process control (SPC) methods have to be developed in order to handle complex data, whose collection and storage is nowadays facilitated by modern data acquisition technologies. In many practical situations the quality characteristic of the process can be modelled as a function defined on a compact domain. Data of such kind are the foundation of a rapidly expanding area of statistics referred to as *functional data analysis* (FDA [14, 4]). SPC methods that allow monitoring and controlling such processes are better known as *profile monitoring* techniques [12].

Measures of functional covariates are often available together with the functional quality characteristic. In this communication we build a novel control charts scheme [11] which fully exploit the additional information content of those functional covariates to improve the profile monitoring of the quality characteristic. The proposed chart is referred to as *functional regression control chart* (FRCC) and provides an extension of the *regression* (or *cause-selecting*) *control chart*, which arises in the multivariate context [10, 7, 8, 16, 15]. The effectiveness of the proposed control chart is demonstrated by means of a real-case study dealing with monitoring CO_2 emissions and identifying (negative) shifts after a specific energy efficiency initiative (EEI) that has been performed on a Ro-pax ship owned by the shipping company *Grimaldi Group*.

2 The proposed control chart

Let $Y(t)$ represent the functional quality characteristic (hereinafter referred to as *response variable* or simply *response*) and $\boldsymbol{X}(t) = (X_1(t), \dots, X_p(t))^\top$ be the functional covariates (hereinafter referred to as *predictor variables* or simply *predictors*) with $t \in \mathscr{S}$, a compact set in $\mathbb{R}$, which are assumed to be elements of $L_2(\mathscr{S})$ (the Hilbert space of square integrable functions defined on the closed interval $\mathscr{S}$). In this setting, we model the predictors as influencing the response according to the following multivariate functional linear regression model

$$Y^Z(t) = \int_{\mathscr{S}} (\boldsymbol{\beta}(s,t))^\top \boldsymbol{X}^Z(s)\,ds + \varepsilon(t), \tag{1}$$

with $Y^Z(t)$ and $\boldsymbol{X}^Z(t) = \left(X_1^Z(t), \dots, X_p^Z(t)\right)^\top$ the point-wise standardized response and predictor variables and $\boldsymbol{\beta}(s,t) = (\beta_1(s,t), \dots, \beta_p(s,t))^\top$ the vector of the regression functional parameters. Given n iid observations $(Y_i(t), \boldsymbol{X}_i(t))$, $i = 1, \dots, n$ (i.e., Phase I dataset), an estimator $\hat{\boldsymbol{\beta}}(s,t)$ of the coefficent vector $\boldsymbol{\beta}(s,t)$ is obtained by considering the truncated versions of the univariate and multivariate Karhunen-Loève's basis expansions [6] of $Y^Z(t)$ and $\boldsymbol{X}^Z(t)$, respectively. The rationale behind the FRCC is monitoring the functional residual

$$e(t) = Y^Z(t) - \hat{Y}^Z(t), \tag{2}$$

where

$$\hat{Y}^{Z}(t)=\int_{\mathscr{S}}\left(\hat{\boldsymbol{\beta}}(s,t)\right)^{\top}\boldsymbol{X}^{Z}(s)\,ds \tag{3}$$

is the prediction of $Y^{Z}(t)$ given $\boldsymbol{X}^{Z}(t)$ based on the linear model (1). For this purpose, we use the approach adopted in [5, 17, 13] where the residuals are decomposed via principal component analysis and the corresponding coefficients are monitored by means of the Hotelling's T^2 and the squared prediction error (SPE) control charts. However, other expansion techniques can be used as well. The control limits of the T^2 and SPE control charts are calculated as the percentiles of their empirical distributions obtained from the Phase I sample based on an overall Type I error α. This phase, along with the estimation of the sample version of Equation (1), will be referred to as Phase I.

For a new observation $(\boldsymbol{X}^{*}(t), Y^{*}(t))$, the residual and the associated T^2 and SPE statistics can be calculated. An alarm is issued if at least one of the latter two statistics violates the control limits (Phase II).

3 A real-case study

In this section an application to a real-case study of the FRCC is illustrated. In particular, real data are collected from a Ro-Pax ship owned by the Italian shipping company *Grimaldi Group* from December 2014 to October 2017. The *CO_2 emissions* per each voyage are considered as the response variable, whereas, the *sailing time*, the *speed over ground* and the *longitudinal* and *transverse wind components* are assumed as the predictors (further information on the variables can be found in [1, 3, 9]).

During February 2016 a specific EEI was performed that plausibly produced a negative shift in the response mean [3]. In light of this, observations collected before the considered EEI are used as the Phase I sample, whereas the remaining observations pertain to the Phase II. The overall Type I error α is set to 0.0027, as commonly done to set three-sigma limits on classical Shewhart control charts. To evaluate the FRCC performance, two competitor profile monitoring schemes proposed in [2] and [13], respectively, are considered as well. The first consists of monitoring the scores along the first principal components of the response by means of the Hotelling's T^2 and SPE control charts (hereinafter referred to as as RESP control chart). The second competitor is the index-based (INBA) control charting, which monitors the area under the response curve.

The performance of the considered control charts are evaluated by means of the *average run length* (ARL) [11] after the EEI, given that the in-control ARL is equal to $1/\alpha = 370$. The estimated ARLs, denoted with $\widehat{\text{ARLs}}$, are reported in Table 1. They clearly point out that the proposed control chart outperforms the competitors. Indeed, the $\widehat{\text{ARL}}$ achieved by the proposed FRCC is markedly lower than those of the RESP and INBA control charts.

Table 1 $\widehat{ARL}$s for the FRCC, RESP control chart and INBA control chart

	$\widehat{ARL}$
FRCC	18.60
RESP	34.09
INBA	50

4 Conclusion

A regression control chart is proposed to monitor a functional quality characteristic when observations of other functional covariates are available. It consists of monitoring the functional residuals coming from a function-on-function linear regression of the response, instead of the response. An application of the proposed control chart to a real-case study aiming to monitor the CO_2 emissions of a Ro-pax ship demonstrates that it outperforms two other popular alternatives proposed in the literature in identifying a shift on the response. However, further investigations should be done to asses the FRCC performance over different scenarios.

References

1. Bocchetti, D., Lepore, A., Palumbo, B., Vitiello, L.: A statistical approach to ship fuel consumption monitoring. Journal of Ship Research **59**(3), 162–171 (2015)
2. Colosimo, B.M., Pacella, M.: On the use of principal component analysis to identify systematic patterns in roundness profiles. Quality and reliability engineering international **23**(6), 707–725 (2007)
3. Erto, P., Lepore, A., Palumbo, B., Vitiello, L.: A procedure for predicting and controlling the ship fuel consumption: Its implementation and test. Quality and Reliability Engineering International **31**(7), 1177–1184 (2015)
4. Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer Science & Business Media (2006)
5. Grasso, M., Menafoglio, A., Colosimo, B.M., Secchi, P.: Using curve-registration information for profile monitoring. Journal of Quality Technology **48**(2), 99 (2016)
6. Happ, C., Greven, S.: Multivariate functional principal component analysis for data observed on different (dimensional) domains. Journal of the American Statistical Association (2016)
7. Hawkins, D.M.: Multivariate quality control based on regression-adiusted variables. Technometrics **33**(1), 61–75 (1991)
8. Hawkins, D.M.: Regression adjustment for variables in multivariate quality control. Journal of Quality Technology **25**, 170–182 (1993)
9. Lepore, A., Palumbo, B., Capezza, C.: Monitoring ship performance via multi-way partial least-squares analysis of functional data. In: SIS2017 Statistical Conference—Statistics and Data Science: new challenges, new generations. University of Florence, Italy (2017)
10. Mandel, B.: The regression control chart. Journal of Quality Technology **1**(1), 1–9 (1969)
11. Montgomery, D.C.: Introduction to statistical quality control. John Wiley & Sons (2007)
12. Noorossana, R., Saghaei, A., Amiri, A.: Statistical analysis of profile monitoring, vol. 865. John Wiley & Sons (2012)

13. Pini, A., Vantini, S., Colosimo, B.M., Grasso, M.: Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. Journal of the Royal Statistical Society: Series C (Applied Statistics) (2017)
14. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer Series in Statistics. Springer (2005)
15. Shu, L., Tsung, F., Tsui, K.L.: Run-length performance of regression control charts with estimated parameters. Journal of Quality Technology **36**(3), 280–292 (2004)
16. Wade, M.R., Woodall, W.H.: A review and analysis of cause-selecting control charts. Journal of Quality Technology **25**, 161–169 (1993)
17. Woodall, W.H., Spitzner, D.J., Montgomery, D.C., Gupta, S.: Using control charts to monitor process and product quality profiles. Journal of Quality Technology **36**(3), 309 (2004)

Understanding pro-environmental travel behaviours in Western Europe

Analisi delle scelte di trasporto pro-ambientali nell'Europa Occidentale

Gennaro Punzo, Rosalia Castellano, and Demetrio Panarello[1]

Abstract This study aims at understanding, from a gender perspective, the reasons behind citizens' choice of using public transport, and whether this choice is driven by pro-environmental behaviour. Using Eurobarometer data (2013), we perform ordered logistic regressions comparatively for Germany, Italy and the Netherlands. Financial, political and environmental factors are shown to have significant roles in shaping travel behaviours, with interesting gender and cross-country differences.

Abstract *Questo studio esplora, in una prospettiva di genere, i fattori che indirizzano i cittadini alla preferenza del trasporto pubblico, con l'intento di indagare se tale scelta sia anche dettata da atteggiamenti pro-ambientali. L'analisi è eseguita su dati Eurobarometro (2013) in un'ottica di comparazione internazionale tra Germania, Italia e Olanda. I modelli di regressione logistica ordinale consentono una valutazione del ruolo che le componenti finanziarie, politiche e ambientali svolgono nella definizione dei comportamenti di viaggio, evidenziando tratti distintivi per genere e Paese.*

Keywords: Sustainable transport, Cross-country comparison, Gender-based analysis

1 Introduction

Dealing with personal transportation is one of the major environmental challenges in current years when the car is the leading travel mode. In Europe, car use growth accelerated sharply during the last decades [1], until exceeding 80% of inland

[1] Gennaro Punzo and Demetrio Panarello, University of Naples Parthenope, Department of Economic and Legal Studies, email: gennaro.punzo@uniparthenope.it; demetrio.panarello@uniparthenope.it

Rosalia Castellano, University of Naples Parthenope, Department of Management and Quantitative Studies, email: lia.castellano@uniparthenope.it

passenger transport at the beginning of the 2000s. Since then, the share of passenger transport by car was within the range 83% to 83.7% up to 2015 – against 7.7% for trains and 9.2% for motor coaches, buses and trolley buses (Eurostat database) – and it is expected to keep its dominant position in the future.

Traffic that is not properly managed generates serious consequences for human health and large scale social and environmental problems with detrimental effects on climate changes, as well as negative emotional reactions because of noise and pollution in urban areas [5,7]. This is why many world's countries are making significant efforts to develop pro-environmental behaviours (PEBs) associated with transportation to meet the environmental challenge in the path to sustainability – e.g., Kyoto Protocol (1997), White Paper "European Transport Policy" (2001).

A widespread literature investigates the complexity of PEBs at the individual and social level by multiple perspectives. All these conceptual models, together with the recent approaches on cultural participation and civic virtue [6], have proved to be complementary as a way of providing policy makers with solid grounds to better recognise socio-economic behaviours and to design interventions for environmental sustainability [8]. Specific travel behaviours, which are usually distinctly gender-related, should constitute as small a burden on the environment as possible. For example, measures aimed at taming the car (i.e., by providing safe and inexpensive alternatives to private motor vehicles such as walking, cycling, and public transport) could be viable ways to create more sustainable transport systems [1,7].

By considering the choice of transport mode as a pro-environmental proxy of sustainability [1,8], this paper performs a comparative investigation in three Western European countries (Germany, Italy and the Netherlands) of the reasons behind citizens' choice of using public transport (PT) rather than private transportation modes, and whether this choice is driven by sustainable behaviour. Indeed, all these countries are leaders in sustainable mobility in some respects (see [4]). The research hypothesis assumes that the impact of the variety of objective and subjective factors on the use of public service differ across countries, whereby the differentials among countries should also be analysed from a gender perspective.

2 Methodology and data

One of the main elements of novelty of this work is that it links research on the driving forces of PT usage across countries to the assessment of PEBs from a gender perspective. Methodologically, we estimate ordered logistic regressions by country and gender on a set of covariates, which are grouped into three dimensions (travel mode choice, own financial perception, politics and environment) supposed to influence the frequency of use of PT (y_i, manifest variable), controlling for a range of personal socio-demographic characteristics.

An underlying decisional process – based on a comparison among the utilities of different travel modes, leading to the choice of using PT – is expected. Therefore, a continuous unobservable propensity (y_i^*, latent variable) would cross thresholds (τ) that differentiate adjacent levels of the observed ordered y_i's. It follows that a cumulative model for the ordered logit model is equivalent to a system composed of a set of thresholds τ_m and a linear regression model for an underlying continuous response. The latent variable, which represents the relative advantage of using PT, is supposed to be linearly related to the observed x's through the structural model:

$$y_i^* = x_i\beta + \varepsilon_i \qquad (1)$$

where β is the vector of coefficients and ε_i is the error term with mean zero and standard deviation $\pi/\sqrt{3}$.

The manifest ordinal variable (y_i), which measures PT use frequency through a 7-point Likert scale, is related to y_i^* according to the following model:

$$y_i = m \quad if \quad \tau_{m-1} \le y_i^* < \tau_m \qquad for \quad m = 1 \quad to \quad J \qquad (2)$$

where m identifies the seven levels (J=7) of the manifest variable (*never*, *less than twice a month*, *two/three times a month*, *about once a week*, *two/three times a week*, *once a day*, *several times a day*) and τ the estimated thresholds on a latent variable used to differentiate the levels of PT usage.

The probability that the individual i will select the alternative m is:

$$p_{im} = P(y_i = m|x_i) = F(\tau_m - x_i\beta) - F(\tau_{m-1} - x_i\beta) \qquad (3)$$

where F is the logistic cumulative density function (c.d.f.).

We use Eurobarometer data on the attitudes of Europeans towards urban mobility [3]. This survey, carried out in the EU-28 between May and June 2013, consisted in conducting face-to-face interviews in people's homes to look at their transport habits, experiences, opinions and expectations. In all three countries, it is possible to drive with supervision at age 17, and without supervision at 18. This is why we include only individuals aged 18 and above (until 79), as the use of PT before that age is driven by necessity rather than personal choice[1].

3 Main results and discussion

Tables 1 and 2 show our main results for men and women, respectively[2]. A brief focus on the outcomes from each country follows.

Table 1: Ordered logit estimation – Men (2013)

Variables	***Coefficients***		
	Germany	***Italy***	***Netherlands***
Socio-demographic			
Marital status (1: *married*)	-0.200	-0.356	-0.467^{*}
Age at completion of education (1: *<16 years*)	-0.340	-0.367	-0.987^{**}
Age (1: *61-79 years*)	-0.472^{**}	-0.801^{***}	-0.631^{***}
Job level (1: *professional/manager*)	-0.206	0.326	0.273

[1] The total number of observations is 3,329. Specifically, there are 1,391 individuals from Germany, 1,001 from Italy and 937 from the Netherlands. Post-stratification weights, based on a comparison for each sample with the respective universe, are properly accounted for.

[2] Standard errors are not included for sake of brevity; though, they are available upon request.

Community (1: *city*; 0: *town/rural*)	1.166***	0.955***	0.544**
Children (1: *aged < 10*)	-0.088	-0.086	-0.126
Travel mode choice$^{\#}$			
Car use frequency: *every day*	-2.355***	-0.574	-2.128***
Car use frequency: *every week*	-1.183***	0.772	-0.841*
Bicycle use frequency: *every day*	0.193	1.197***	1.305***
Bicycle use frequency: *every week*	0.825***	0.941***	0.811***
Pedestrian paths use frequency: *every day*	1.167***	0.855***	0.612**
Pedestrian paths use frequency: *every week*	0.741**	0.408	0.200
Urban travel frequency: *every day*	0.883***	0.540	0.618**
Urban travel frequency: *every week*	0.589**	0.490	0.217
Own financial perception			
Cost of living (1: *good*)	0.295	-0.070	0.347
Cost of energy (1: *good*)	0.318	0.457	-0.075
Household financial situation (1: *good*)	-0.143	0.389*	0.611
Car ownership (1: *yes*)	-0.554*	-1.115***	0.328
Apartment/house ownership (1: *yes*)	-0.248	0.011	-0.001
Politics and Environment			
Road congestion issues (1: *important*)	-0.107	-0.344	-0.490**
Air pollution issues (1: *important*)	-0.051	-0.236	-0.513**
Travel cost issues (1: *important*)	-0.313	0.310	0.606**
Lower public transport prices (1: *important*)	0.657***	-0.215	-0.281
EU policies against poverty (1: *important*)	-0.650***	-0.187	-0.312
Higher public transport quality (1: *important*)	0.296*	0.012	0.599***
Better walking facilities (1: *important*)	-0.127	-0.001	0.441*
Better cycling facilities (1: *important*)	0.179	-0.590**	0.058
Access time restrictions (1: *important*)	-0.587**	0.209	-0.098
Car sharing incentives (1: *important*)	0.422**	-0.486*	0.203
Urban traffic responsibles (1: *citizens*)	0.089	-0.156	-0.503**

*Notes: *p<0.10; **p<0.05; ***p<0.01* $^{\#}$*Reference: less than once a week*

Table 2: Ordered logit estimation – Women (2013)

Variables	***Coefficients***		
	Germany	***Italy***	***Netherlands***
Socio-demographic			
Marital status (1: *married*)	-0.331	0.251	-0.880***
Age at completion of education (1: *<16 years*)	-0.298	-1.012***	-0.638*
Age (1: *61-79 years*)	-0.384*	-0.280	-0.306
Job level (1: *professional/manager*)	0.316	0.509*	-0.003
Community (1: *city*; 0: *town/rural*)	1.636***	0.390*	0.554**
Children (1: *aged < 10*)	-0.219	-0.243*	-0.150
Travel mode choice$^{\#}$			
Car use frequency: *every day*	-2.181***	-1.523***	-1.530***
Car use frequency: *every week*	-0.885***	-0.313	-0.822**
Bicycle use frequency: *every day*	-0.336	0.289	0.522*
Bicycle use frequency: *every week*	0.360*	0.187	0.122
Pedestrian paths use frequency: *every day*	0.982**	0.979***	0.283
Pedestrian paths use frequency: *every week*	0.763	0.968***	0.309
Urban travel frequency: *every day*	0.747***	0.802***	0.799***

Urban travel frequency: *every week*	0.285	1.039***	0.527**
Own financial perception			
Cost of living (1: *good*)	0.284	0.690**	0.354
Cost of energy (1: *good*)	0.391**	-0.076	0.537**
Household financial situation (1: *good*)	-0.094	0.214	-0.497*
Car ownership (1: *yes*)	-0.604**	-0.684**	-0.379
Apartment/house ownership (1: *yes*)	-0.162	-0.714***	0.005
Politics and Environment			
Road congestion issues (1: *important*)	0.315	0.011	0.525**
Air pollution issues (1: *important*)	-0.334	-0.300	-0.297
Travel cost issues (1: *important*)	-0.123	0.349	-0.250
Lower public transport prices (1: *important*)	0.188	0.064	0.207
EU policies against poverty (1: *important*)	-0.261	-0.100	-0.389*
Higher public transport quality (1: *important*)	0.154	0.468**	0.399**
Better walking facilities (1: *important*)	0.233	-0.079	0.614**
Better cycling facilities (1: *important*)	-0.152	0.054	-0.047
Access time restrictions (1: *important*)	0.097	-0.020	0.021
Car sharing incentives (1: *important*)	-0.144	-0.062	0.171
Urban traffic responsibles (1: *citizens*)	0.499***	-0.406**	-0.075

*Notes: *p<0.10; **p<0.05; ***p<0.01* $^{\#}$*Reference: less than once a week*

Results show that the variety of factors have different roles in shaping travel behaviours in the three countries and for each gender; nevertheless, a cross-national analysis of sex differences in PEBs may be drawn.

A high education level has a positive effect on PT usage for Dutch people; in Italy, women who are well-educated and have got a high-level job are about to use PT with a greater intensity than women either with a low education and low-level job, students or unemployed. This states the key role of education in environmental matters: well-educated people are, in general, more aware and concerned and, thus, more willing to engage in concrete environmental protection actions. Due to the lower availability of childcare services in Italy, which leads to major challenges in urban travelling [2], Italian women with kids appear less prone to use PT. Living in a big city has a positive influence on the probability of PT usage, as populous urban areas are usually characterised by shorter distances and more developed PT infrastructures. In each country, elderly male citizens are less likely to use PT than their younger counterparts, showing to be more addicted to private vehicles than the new generations, and confirming that the rate of people using multiple modes of transportation, rather than just the car, declines in later life stages.

Men, in general, appear to be more prone to active and multimodal transport compared to women, especially combining bike and PT use. Among women, Dutch are the only ones combining PT use and cycling in daily life, apparently due to a stronger bicycle culture and better cycling facilities in the country. Using pedestrian paths on a daily basis also seems to be positively related to PT use. Frequent car users, on the other hand, are less likely to use PT. These outcomes confirm that public and active transport can be integrated, while a car-based mobility style does not give room for other means of transportation.

Italians with a high socioeconomic status, proxied by home ownership and perceived household financial situation, are less likely to travel by PT if women, and more likely if men, maybe because women are usually travelling shorter distances to work than men, which does not make PT an attractive (e.g., fast, cheap) alternative to the car. People who own a personal car, as expected, are less likely to use PT, but not in the Netherlands, which appears to be more open to alternative transportation modes compared to the other countries.

Dutch people who evaluate road congestion as an important urban problem act differently depending on their gender: while men who do so are significantly less likely to use PT services, women are more likely to use them. It looks like Dutch women care about urban problems more and try to solve them with their behaviour. This is furtherly explained by the fact that Dutch men considering air pollution a relevant urban problem are less likely to use PT as a sustainable solution. Dutch men who are concerned about travel costs, though, are more likely to move by PT. Even if costs are indeed a sensitive issue for mobility, only German male PT users seem to mind about ticket prices. Most PT users demand for a higher PT quality, tending to favour the development of policies that improve conditions for the transport modes they habitually use. In Italy, men who think that enhanced cycling facilities could be appropriate measures to improve urban travelling apparently use PT services with a lower frequency than men who do not, showing not to have a multimodal lifestyle. German and Italian women thinking that urban traffic responsibility is a matter of citizens themselves are more likely to use PT, and the same applies to Dutch men, showing concern for the quality of urban environment and trying to make it better.

All these outcomes indicate that policies aimed at designing pro-environmental transport systems need to be adapted based on local context and gender.

References

1. Buehler, R., Pucher, J., Gerike, R., Götschi, T.: Reducing car dependence in the heart of Europe: lessons from Germany, Austria, and Switzerland. Transp Rev 37(1), 4-28 (2017)
2. Castellano, R., Punzo, G., Rocca, A.: The generational perspective of gender gap in wages and education in southern Europe. Rev Soc Econ (2018) doi: 10.1080/00346764.2017.1423512
3. European Commission: Attitudes of Europeans towards urban mobility. Special Eurobarometer 406 (2013) http://ec.europa.eu/public_opinion/archives/ebs/ebs_406_en.pdf
4. European Commission: State of the art on alternative fuels transport systems in the European Union. Final Report. Expert Group on Future Transport Fuels. Brussels (2015)
5. Khreis, H., May, A.D., Nieuwenhuijsen, M. J.: Health impacts of urban transport policy measures: a guidance note for practice. J Transp Health 6, 209-227 (2017)
6. Owen, A.L., Videras, J.: Civic cooperation, pro-environment attitudes, and behavioral intentions. Ecol Econ 58(4), 814-829 (2006)
7. Redman, L., Friman, M., Gärling, T., Hartig, T.: Quality attributes of public transport that attract car users: A research review. Transp Policy 25, 119-127 (2013)
8. Steg, L., Vlek, C.: Encouraging pro-environmental behaviour: An integrative review and research agenda. J Environ Psychol 29(3), 309-317 (2009)

Family & Economic issues

Measuring Economic Uncertainty: Longitudinal Evidence Using a Latent Transition Model

Misurare l'Incertezza Economica: risultati longitudinali usando un modello di transizione a classi latenti

Francesca Giambona, Laura Grassini and Daniele Vignoli

Abstract Economic uncertainty has become an increasing important factor in explaining socio-economic household behaviour, especially for its implications in household choices and demographic dynamics. Accordingly, an operational definition and a measure of economic uncertainty are needed. To this aim, the main purpose of this contribution is to measure economic uncertainty of Italian families by using the 2008-2011 and 2012-2016 longitudinal Italian SILC (Statistics on Income and Living Conditions) data in a latent transition analysis (LTA) approach. LTA is applied, in order to: *i*) to classify Italian households into homogenous classes characterized by different levels of economic uncertainty, and *ii)* to assess whether changes in latent class membership occurred in the time span selected.

***Abstract** L'incertezza economica è un fattore di rilievo nello spiegare le dinamiche socio-economiche delle famiglie. Al fine di analizzare l'effetto della incertezza economica sulle scelte familiari è necessario, per prima cosa, fornire una definizione (operativa) e una misura dell'incertezza economica. Il presente lavoro ha lo scopo di misurare l'incertezza economica delle famiglie italiane, utilizzando i dati longitudinali dell'indagine sul reddito e sulle condizioni di vita 2008-2011 e 2012-2016. L'applicazione di un modello di transizione a classi latenti (LTA), permetterà di: i) classificare le famiglie in classi omogenee caratterizzate da diversi livelli di incertezza economica, e ii) valutare se si sono verificati cambiamenti nell'appartenenza a una classe latente nell'arco di tempo selezionato.*

Keywords: economic uncertainty, latent transition analysis, childbearing

[1] Francesca Giambona, Dipartimento di Statistica, Informatica e Applicazioni, Università di Firenze; email: giambona@disia.unifi.it

Laura Grassini, Dipartimento di Statistica, Informatica e Applicazioni, Università di Firenze; email: grassini@disia.unifi.it

Daniele Vignoli, Dipartimento di Statistica, Informatica e Applicazioni, Università di Firenze; email: vignoli@disia.unifi.it

1 Introduction

Uncertainty is one of the key aspects of the globalizing world, of our “risk society”, and economic uncertainty has become a notable factor in explaining socio-economic behaviours, especially for its implications in family choices and demographic dynamics (Scherer, 2009; Blossfeld & Hofmeister, 2006; Blossfeld, Mills, & Bernardi, 2006). Recently, a number of studies have focused the attention on the impact of economic uncertainty on childbearing decisions (Kind and Kleibrink, 2013; Kreyenfeld, 2010, 2015; Kreyenfeld et al., 2012; Özcan, Mayer, & Luedicke, 2010; Pailhé & Solaz, 2012), also for Italy (e.g., Barbieri et al., 2015; Vignoli et al., 2012). Despite the timely and key relevance of the notion of “Economic Uncertainty” for contemporary societies, however, scholars have been so far rather imprecise regarding its definition and operationalization.

Economic uncertainty is a relative concept – relative to expectations that economic prospects may be stable at a given level. Nonetheless, economic uncertainty is, by its very nature, non-observable. Hence, in demographic and sociological research, it is customarily operationalized as an individual risk factor, mainly related to the labour market (e.g. unemployment, short-term contract jobs, underemployment, involuntary part-time, or a combination of these; Mills and Blossfeld 2013; Kreyenfeld, Andersson, Pailhé 2012). Such an approach, however, focuses solely on the objective side of uncertainty.

A significant downside economic risk – i.e. a hazard or danger – looming in the individuals’ economic future, which they are unable to adequately insure against or avoid or ignore (Osberg & Sharpe 2014). Recently, various empirical concepts have been proposed to distinguish the objective aspects of economic insecurity (Ranci et al., 2017) from subjective feelings such as a sense of uncertainty (Mau, Mewes, & Schöneck, 2012). Previous research has largely failed to recognize that individuals, depending on the extent to which they feel and tolerate uncertainty, might differ with respect to how they react and take decisions in uncertain economic situations (Bernardi et al. 2009; Kreyenfeld 2010). In all, although demographic actors routinely face various types of economic uncertainty, the effects of economic uncertainties are often unclear and hard to assess, in part due to the absence of valid measures.

In this paper, we aim to measure economic uncertainty by acknowledging both its subjective and objective sides. To this end, we propose a synthetic measure of economic uncertainty that acknowledges the multidimensional and latent nature of the concept, through a latent transition analysis (LTA) (Collins and Lanza, 2010).

We apply LTA to the longitudinal data from Italian section of the EU-SILC (Statistics on Income and Living Conditions), related to 2008-2011 and 2012-2016 waves in order to compare empirical findings during and after the Great Recession.

2 Methods

In analysing latent variable models, latent transition analysis (LTA) and latent class analysis (LCA) are related methods. LCA is a statistical method used to group individuals (cases, units) into classes (categories) of an unobserved (latent) variable. It is a statistical procedure for identifying class membership probabilities among statistical units (e.g., individuals, families, and so on), using the responses provided to some chosen set of observed variables. In LCA, the class membership probabilities (i.e., the probability that an individual belongs to a certain class) and the item response probabilities conditional upon class membership (i.e., the probability for an individual to provide a certain response to a specific item given that she/he has been classified in a specific latent class) are estimated and, according to the item response probabilities, observations are grouped (clustered) into classes (Collins and Lanza, 2010; Magidson & Vermunt 2000; Lazarsfeld, 1950). A questionable issue is about the number of classes selected. Theoretically and conceptually, classes may be identified according with some *a priori* research assumption, then statistical criteria can be used to confirm theoretical expectations. Criteria for assessing the number of classes suggest several statistical methods. For instance, Nylund, Asparouhov and Muthén (2007) indicate the Bayesian Information Criteria (BIC) as the best one, so that the number of classes is selected by minimizing the BIC value.

Usually, when longitudinal data are to be analyzed, research questions deal not only with latent class membership but also with changes over time. LTA is a type of latent Markov model and is a variation of the LCA designed to model not only the latent class membership, but also the transitions over time in latent class membership. In LCA, latent classes represent stable sets of characteristics or states of behavior, whilst in LTA, individuals may change latent classes over time. Thus, in this framework, the term “latent statuses” is used instead of “latent classes”, as subgroup membership is not assumed stable over time.

Three sets of parameters are estimated in LTA. First, latent status membership probabilities are estimated for each time. Second, transition probabilities reflect the probability of transitioning from a particular latent status at time t to another latent status at time t+1. Third, a set of item-response probabilities reflects the correspondence between the observed indicators of the latent variable at each time period and latent status membership, in much the same way that factor loadings link observed indicators to latent variables in factor analysis. That is, in addition to the number of classes and the size of classes being subject to change, it is interesting to locate the households that are stayers (in the same class at each wave) and those who are movers.

Specifically, we want to distinguish between households that move to a more positive class (minor uncertainty) and those who move to a less positive class.

3 Data and some first results

The European Union survey named Statistics on Income and Living Conditions (EU-SILC), started in 2003, is aimed at gathering comparable cross-sectional and longitudinal individual level data on income, poverty, social exclusion and living conditions in 27 participating countries. Here, we consider the longitudinal the SILC section for 2008-2011 and 2012-2015 Italian waves. For each household we select some items related to economic uncertainty. We selected a set of objective and subjective items of the EU-SILC questionnaire related to the following dimensions: *financial strain* (arrears on utility bills, on mortgage or rental payments, capacity to afford a meal with meat/fish every second day, ability to make ends meet); *housing* (tenure status); *goods possession* telephone, computer, dishwasher, car and TV colour); *occupational status* (temporary/not temporary job).

We recall that the item-response probabilities in LTA play the same role as in LCA; namely, they represent the basis for assigning labels to latent statuses. BIC value and substantial interpretation of classes, suggest a partition into 5 classes (latent statuses).

Table 1 displays the percentage of households classified in each class for the first time. Latent variable is specified to be ordinal, thus state 1 groups households with lower values of economic uncertainty up to state 5, which groups households with the higher level of economic uncertainty. In state 1 households are owner of their home, have durables goods, have a permanent job, don't have arrears in utility bills or on mortgage or rental payments, make easily ends meet and are able to afford a meal with meat/fish every second day.

Table 1: latent states composition

	Latent State (*)				
	1	**2**	**3**	**4**	**5**
2008-2011	0,205	0,336	0,280	0,139	0,041
2012-2015	0,146	0,221	0,396	0,189	0,048

(*) 1-5: from positive to negative status

In Table 2 the transition probability matrix shows that in the period of the Great Recession (2008-2011), the probability of remaining in the same class is lower than in the post-recession period (2012-2015). Households in better states (1,2) are more "stable" in 2012-2015 respect to 2008-2011 (95% in 2008-2011, 99% in 2012-2015). This is more evident for households in 2008-2011 in intermediate position (state 3) that "move" to better states (state 2 in particular). This happens also for states 4 and 5 for both time periods, but especially in 2008-2011. Mainly, households in better states (1,2) tend to be more stable than the others; whilst households in worse states tend to improve their position, in particular households experimented the Great Recession.

Table 2: transition probability matrix

	State[-1]						**State[-1]**				
	1	2	3	4	5		1	2	3	4	5

State						State					
1	0,942	0,001	0,003	0,005	0,000	1	0,992	0,001	0,015	0,000	0,000
2	0,032	0,957	0,054	0,005	0,001	2	0,000	0,997	0,006	0,008	0,000
3	0,024	0,021	0,894	0,153	0,088	3	0,004	0,001	0,978	0,042	0,018
4	0,002	0,022	0,029	0,835	0,078	4	0,003	0,000	0,000	0,914	0,105
5	0,000	0,000	0,020	0,001	0,833	5	0,000	0,001	0,001	0,036	0,877
			2008-2011						**2012-2015**		

Acknowledgements: ERC Grant EU-FER “Economic Uncertainty and Fertility in Europe”, grant agreement n° 725961 (PI: Daniele Vignoli)

References

1. Barbieri, P., Bozzon, R., Scherer, S., Grotti, R., Lugo, M.: The Rise of a Latin Model? Family and fertility consequences of employment instability in Italy and Spain. European Societies, 17(4), 423–446 (2015).
2. Bernardi L., Klarner A., von der Lippe H.: Job Insecurity and the Timing of Parenthood: A Comparison between Eastern and Western Germany. European Journal of Population 24: 287–313 (2009).
3. Blossfeld, H.P., Hofmeister, H.: Globalization, Uncertainty & Women’s Careers: An International Comparison. Cheltenham, UK: Edward Elgar Publishing (2006)
4. Blossfeld, H.P., Mills, M., Bernardi, F.: Globalization, Uncertainty and Men’s Careers: An International Comparison. Cheltenham, UK: Edward Elgar Publishing (2006)
5. Collins, L.M., Lanza, S.T.: Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences. New York, NY: John Wiley (2010)
6. Kreyenfeld, M.: Economic Uncertainty and Fertility. KZfSS Kölner Zeitschrift Für Soziologie Und Sozialpsychologie, 67(1), 59–80 (2015)
7. Kreyenfeld, M., Andersson, G.: Socioeconomic differences in the unemployment and fertility nexus: Evidence from Denmark and Germany. Advances in Life Course Research, 21, 59–73 (2014)
8. Kreyenfeld, M., Andersson, G., Pailhe, A.: Economic uncertainty and family dynamics in Europe: Introduction. Demographic Research, 27(28), 835–852 (2012)
9. Lazarsfeld, P.F.: The Logical and Mathematical Foundation of Latent StructureAnalysis.” Pp. 361–412 in Measurement and Prediction, edited by S. A. Stouffer, et al. Princeton, NJ: Princeton University Press (1950)
10. Magidson, J., Vermunt, J. K.: Latent class analysis. Cambridge, MA:Cambridge University Press (2000).
11. Mau, S., Mewes, J., Schöneck, N.M.: What determines subjective socio-economic insecurity? Context and class in comparative perspective’, Socio-Economic Review, 10(4), 655-682 (2012)
12. Mills, M., Blossfeld, H.P.: Globalization, uncertainty and changes in early life courses. In H.-P. Blossfeld, E. Klijzing, M. Mills, & K. Kurz (Eds.), Globalization, Uncertainty and Youth in Society (pp. 1–24). London, UK and New York, US: Routledge (2006)
13. Mills, M., Blossfeld, H. P.: The Second Demographic Transition meets globalisation: a comprehensive theory to understand changes in family formation in an era of rising uncertainty. In A. Evans & J. Baxter (Eds.), Negotiating the life course. Stability & change in life pathways (pp. 9–33). New York, US: Springer (2013)

14. Nylund, K., Asparouhov, T., Muthén, B.O.: Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling, 14, 535-569 (2007)
15. Osberg, L. Sharpe, A.: Measuring economic insecurity in rich and poor nations", Review of Income and Wealth, 60, S1, pp. 53-76 (2014)
16. Ranci C., Parma A., Bernardi L., Beckfiled J.: The rise of economic insecurity in the EU: concepts and measures, Lives WP 2017/62 (2017)
17. Vignoli, D., Drefahl, S., De Santis, G.: Whose job instability affects the likelihood of becoming a parent in Italy? A tale of two partners. Demographic Research, 26(2), 41–62 (2012)
18. Scherer, S.: The social consequences of insecure jobs. Social Indicators Research 93(3):527–547 (2009)
19. Kind, M. Kleibrink J.: Sooner or Later –Economic Insecurity and the Timing of Frst Birth", Ruhr Economic Papers 422, Ruhr-University Bochum (2013).
20. Özcan, B., Mayer, K.U., Luedicke, J.: The impact of unemployment on the transition to parenthood. Demographic Research 23:807–846. doi:10.4054/DemRes.2010.23.29 (2010)
21. Pailhé, A., Solaz, A.: The influence of employment uncertainty on childbearing in France: A tempo or quantum effect? Demographic Research, 26(article 1), 1–40 (2012)

Intentions to leave Italy or to stay among foreigners: some determinants of migration projects

Le intenzioni degli stranieri di restare o lasciare l'Italia: alcune determinanti dei progetti migratori

Ginevra Di Giorgio, Francesca Dota, Paola Muccitelli and Daniele Spizzichino[1]

Abstract The topic of migration intentions and return migration has been analyzed by different theoretical approaches (Stark (1996); Dustmann (2003)) but still the process of migration decision-making and behaviour is not fully understood (Arango (2000)). On basis of Istat survey "Social Condition and Integration of Foreign Citizens" (SCIF, 2011-2012), this paper aims to explore the migration project of foreign citizens resident in Italy. Using a binomial logistic regression model, has been studied the association of different determinants (socio-demographic characteristics, migration background, personal relationships in Italy, social integration in host country) with the intention of staying permanently in Italy.

Abstract *Il tema delle intenzioni migratorie e delle migrazioni di ritorno è stato analizzato seguendo diversi approcci teorici (Stark (1996); Dustmann (2003)). L'affermarsi di letture diverse e in parte contrastanti rimanda alla complessità di fattori che entrano in gioco nel definire e orientare le intenzioni migratorie (Arango (2000)). A partire dai dati dell'indagine Istat "Condizione e Integrazione Sociale dei Cittadini Stranieri" (SCIF, 2011-2012), questo lavoro si propone di esplorare le intenzioni migratorie degli stranieri che vivono in Italia attraverso un modello di regressione logistica binomiale considerando alcune determinanti (socio-demografiche, del background migratorio, le relazioni interpersonali e alcuni indicatori di integrazione sociale).*

[1] Ginevra Di Giorgio, Istat, digiorgio@istat.it;
Francesca Dota, Istat, dota@istat.it;
Paola Muccitelli, Istat, muccitel@istat.it;
Daniele Spizzichino, Istat, daspizzi@istat.it.

Key words: foreign citizens; migration projects; return behaviour.

1 Background

The issue of return migrations and intention of stay in host country has been analyzed by different theoretical approaches (Stark (1996)); Constant, Massey (2002)). The neoclassical theory associates the decision to return to the country of origin to the lack of integration process in the receiving country, by considering migration as an exclusively individual experience. Instead, the new economic theory of migration considers the return migration as a factor of success of migration experience in which the family plays an important role (Dustmann (2003)). Moreover, the social capital theory (Coleman (1988)) and network analysis approach take into account the role of social networks in return home decision. Therefore no theory is enough to explain factors associated to the intention of return home or stay in host country. However each one refers to the complexity of factors correlated to migrants' decision or intentions.

Since some decades Italy has been a country of immigration with a rising number of migrants. Moreover in the last few years the share of long-term residents has increased (Conti, Strozza (2012)). However, the last census recorded about one million foreigners less then the population register. Likely they have left Italy to re-emigrate to the country of origin or to a third country (onward migration) (Blangiardo (2012)). According to recent studies, based on regional data, the percentage of migrants intending to leave Italy has considerably increased (Terzera (2015)). Therefore, a focus on migrants' intentions seems likely current. The intentions of mobility may vary according to the family and employment conditions or legal status in Italy (Toma, Castagnone (2015); Barbiano di Belgioso (2016)).

2 Research hypotheses and goals

According to the recent debate in the literature, the intention to remain in Italy is not considered only as a positive outcome of the migration project. Indeed, the study aims to explore migrant's intentions (to stay in Italy or to return to the country of origin), taking into account the possible determinants that play an important role to define the migration project.

This study contributes to highlight the importance of socio-demographic characteristics (gender, age, citizenship, education level and employment status), as well as the individual migratory background (migratory intentions at the start, age at the arrival in Italy, years of stay). Moreover it is considered as relevant the family situation in migration (presence of family members in the country of origin or in the receiving country, mixed family, etc.). In fact, the migratory project, regardless of the direction of intentions (stay or return to the country of origin), can represent the outcome of a family choice. The survey data allow a multidimensional analysis of the factors associated to the intentions of individuals living in the family. Similar importance is given to the relational context and to the contacts maintained with the country of origin (periodic return to country of origin) and to some aspects related to social integration in the receiving country (perception of discrimination, knowledge of the Italian language).

3 Data and methods

The analysis of migratory projects was carried out on the data of the sample survey on “Social Condition and Integration of Foreign Citizens” (SCIF), launched for the first time by Istat (Italian National Institute of Statistics) between 2011 and 2012, in partnership with other Italian Institutions[1.]

The SCIF sample survey aims to provide a collection of information about many features of socio-economic inclusion of migrants in Italy in order to explore living condition of resident foreigners. There is a focus on several aspects: family composition, education, migratory path, employment status, discrimination, health conditions and accessibility of health services, migrant integration, citizen's security and victimization, housing conditions. The survey data provides a framework on characteristics, behaviors, attitudes and opinions of the foreign citizens in Italy, as a integration and complement to administrative sources, currently produced by Istat.

SCIF sample survey’s target population is foreign population resident in Italy. The sampling unit is the households resident in Italy with at least one component with foreign citizenship. The family is defined as a group of people living together and related by marriage, kinship, affinity, adoption, protection or affection. In addition to members of the family, people without family ties with residence in Italy have been also interviewed. In the SCIF survey, foreign citizens are identified by citizenship, not by the place of birth.

All the members of 9,553 families were interviewed using CAPI (Computer Assisted Personal Interviewing) techniques. In terms of individuals with foreign citizenship, the survey involved 20.379 people.

The choice of the survey is given by the opportunity to study not only the orientation of migratory projects, but also to investigate the relationship between these and the individual, family and relational sphere of the migrant. To verify the empirical evidence that emerged in the literature, the different association of the factors influencing the migratory pathways was evaluated through logistic regression. The use of this model makes possible to estimate the probability to want to stay permanently in Italy as opposed to the desire to return to one's country of origin (β>0

[1] This experience was co-funded in collaboration with the Ministry of Interior for a project funded by the European Fund for the Integration of Third-Country Nationals.

means an higher propensity to settle down in Italy). The sample used to estimate the model is made up of around 16,000 foreign citizens aged over 15 years.

4 Results

Over 70% of all interviewed migrants had planned to stay in Italy: what are individual, familiar and migratory experience determinants associated to the intentions to stay in Italy?

Results of logistic model confirm the importance of some determinants related to migration background, social integration and socio-demographic characteristics on migration project.

By taking into account migration background emerges how foreigners arrived in Italy with the intention to settle down permanently are likely to confirm same intention at the interviewing (β=3.18). As well as, migrants long-stayers in Italy are more likely to declare to stay compared to short-stayers (β=0.72). Moreover migrants who have never came back to the country of origin (β=0.21) unlike people who came back at least once.

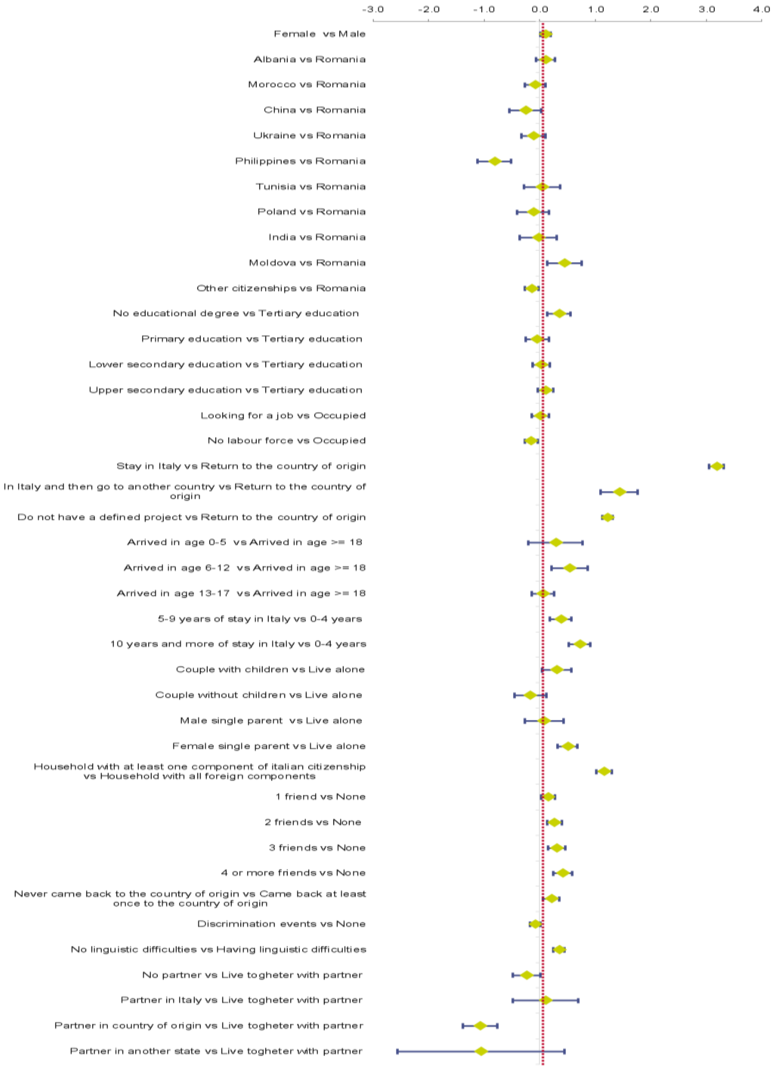
Migrants with a good proficiency of Italian language have higher probability to plan to stay in Italy too (β=0.35). On the economic side, intention to leave Italy is related to less participation to labour market (no labour force β=-0.16), a condition that should makes more difficult socio-economic inclusion.

The intention to stay is higher among foreigners who have settled down in Italy, with own family and a wide friendship network. Compared to those who live alone, couples with children (β=0.31) and single-parent families of mothers (β=0.5) have a higher association with the propensity to stay in Italy. As expected, among people who live with other Italians, the probability of stay in Italy is higher than those who live in families of foreigners only (β=1.16). On the contrary, people with partner in country of origin declare more propensity to go back (β=-1.07).

The citizenship and the ethnic group pattern of social settlement in Italy have an impact on intentions of migrants to stay or come back to the country of origin. Compared to Romanians, Moldovans are more likely to declare to stay, whereas Filipinos have lower propensity to settle down in Italy. Citizenship of origin seems to play a fundamental role to define intentions of migrants as well as years of stay in host country. Therefore the interaction effect between years of stay and citizenship has been taken into account. Filipinos, regardless of during of stay, are always oriented to go back to the country of origin. Instead, Moldovans have higher probability to want to stay in Italy among long stayers (β=0.46 for 5-9 years of stay, β=0.52 for over 10 years of stay). Only among long-stayers, Chinese and Tunisians are more likely to declare the propensity to return to their homeland.

5 Intentions to leave Italy or to stay among foreigners: some determinants of migration projects

Figure 1: Binomial logistic model. Probability to stay in Italy vs to return to country of origin. β value and confidence intervals.



Source: SCIF, 2011-2012.

References

1. Arango J.: Explaining migration: a critical view, International Social Science Journal, vol. 52, (pg. 283-296), 2000.
2. Barbiano Di Belgiojoso E.: “Intention on desired lenght of stay among immigrants in Italy, Genus, 72:1, (2016).
3. Blangiardo G.C.: Un esercito di 1,3 milioni di fantasmi sparito dal censimento, Il Sole 24 Ore. Available via:
   http://www.ilsole24ore.it. Cited 7 May 2012,
4. Coleman J.: Social Capital in the Creation of Human Capital, American Journal of Sociology, suppl. (94), pp. 95-120, (1988).
5. Conti C., Strozza S.: Should I Stay or Should I go? L’immigrazione non comunitaria in Italia, Available via:
   http://www.neodemos.info, Cited 27 March 2012.
6. Constant A., Massey D.S.: Return Migration by German Guestworkers: Neoclassical versus New Economic Theories, International migration, 40 (4), (2002).
7. Dustmann C.: Children and Return Migration, Journal of Population Economics, vol. 16, pp. 815-830, (2003).
8. Stark O.: On the Microeconomics of Return Migration, Occasional Papers no. 1/1996 27/03/2012, University of Vienna, Center for International and Interdisciplinary Studies, (1996).
9. Terzera L.:Famiglie e progetti di mobilità, in G.C. Blangiardo (Ed.) L’Immigrazione straniera in Lombardia XIV Indagine regionale, Rapporto 2014, (2015).
10. Toma S., Castagnone E.: What drives onward mobility within Europe? The case of Senegalese migration between France, Italy and Spain, Population E, 70(1), pp. 65-96, (2015).

Wages differentials in association with individuals, enterprises and territorial characteristics

L'associazione tra differenziali retributivi e caratteristiche individuali, di impresa e territoriali

S. De Santis, C. Freguja, A. Masi, N. Pannuzi, F. G. Truglia[1]

Abstract The availability of new databases integrated by ISTAT allows to draw a territorial map of the private sector employment at the census level, with reference to wage levels and characteristics of individuals and enterprises. The purpose of this paper is to provide a representation of wages and of the most significant variables, referring to individuals, enterprises and territory, mainly contributing to determine wages. After having identify the main factors, it is possible to study the joint distribution of enterprise and worker main characteristics on different areas of the Country, revealing, sometimes unexpected, profiles, which may also help the implementation of local policies. The analysis is conducted on private sector employees, representing three-quarters of the employees and more than half of job positions, and it is focused on unilocalised enterprises with less than 50 employees and with dependent workers, amounting to 1 million 404 thousands enterprises and 7 million 847 thousand job positions.

[1] Stefano De Santis, ISTAT, sdesantis@istat.it

Cristina Freguja, ISTAT, freguja@istat.it

Alessandra Masi, ISTAT, masi@istat.it

Nicoletta Pannuzi, ISTAT, pannuzi@istat.it

Francesco Giovanni Truglia, ISTAT, truglia@istat.it.

The work described in this paper is carried out as part of the Istat project "Indicators for inequality analysis: integration of sources and measurement innovations for a multidimensional phenomenon", with the support of a Scientific Commission of experts from different institutions.

Abstract *La disponibilità di nuove basi dati integrate dall'ISTAT consente di tracciare una mappa territoriale dell'occupazione del settore privato a livello censuario, con riferimento ai livelli salariali e alle caratteristiche dei singoli lavoratori e delle imprese. Lo scopo di questo articolo è fornire una rappresentazione dei salari e delle variabili, riferite a individui, imprese e territorio, che principalmente contribuiscono a determinare la distribuzione dei salari. Dopo aver identificato i principali fattori, è possibile analizzare la distribuzione congiunta delle caratteristiche principali dell'impresa e dei lavoratori in diverse aree del Paese, rivelando profili, a volte inattesi, che possono anche aiutare l'attuazione delle politiche locali. L'analisi è condotta sui dipendenti del settore privato, che rappresentano i tre quarti dei dipendenti e più della metà delle posizioni lavorative, e si focalizza sulle imprese unilocalizzate con meno di 50 addetti e con dipendenti, pari ai 1 milione 404 mila imprese e 7 milioni 847 mila posizioni lavorative.*

Key words: wage differentials, municipalities spatial analysis, inequalities.

1. Premise

Wages play a key role in determining labour supply; at the same time, labour costs are fundamental drivers for business location choices, along with a mix of factors affecting productivity levels in host areas (infrastructures, human capital, legality, business environment, orientation of productive specialization, etc.).

The analysis of wages differentials is notoriously complex, both because multiple factors act simultaneously, and because the available information often is quantitatively and qualitatively inadequate (Porcari et al., 2007).

This paper aims to provide a detailed and comprehensive representation of wages and of the most significant variables, referring to individuals, enterprises and territory, mainly contributing to determine wages distribution.

The analysis is conducted only on private sector employees: they represent three-quarters of the total number of employees (77%) and more than half of job positions (56%); however, it should be pointed out that in the short term the analysis can be extended to self-employed and public servants.

The traditional information asset, represented by ISTAT's social and economic surveys, is today enhanced with an important piece of information coming from the integration of both administrative and statistical data sources. In particular, all administrative sources on labour market can be linked, exploiting the fact that each source is LEED (Linked Employer-Employed Data) and allows to study the joint distribution of enterprise and worker main characteristics (ISTAT, 2016).

The information allows the definition of new aggregates in a statistically consistent way and great detail. It is always possible to add information for unplanned dimensions of analysis (e.g. information collected by survey at both employer and employee levels) and to deepen relationships or association structures.

For this reason, the data bases - obtained from administrative data suitably corrected and integrated - are "statistical products" themselves.

2. Data information

This paper counts on the availability of statistical registers obtained from multiple sources, conveniently integrated at micro-data level: the ASIA statistical register and the "information system on employment"; the RACLI thematic register, extension and part of the aforementioned "information system on employment"; the Extended register Frame-SBS[2].

The first register presents a three-level information structure: enterprise, employee, and employment relationship. At the enterprise level, the statistical register of active enterprises integrates information from public and private administrative sources, and statistical sources; it represents the main source for business demography, the basis for all the ISTAT surveys on enterprises (it identifies the reference population for sampling designs and weighting systems) and it is also used for National Accounts estimates.

At the levels of employee and its employment relationship, the ASIA-Employment archive - derived from the employment database (DB-Occupazione) and annually updated after the updating of the active enterprises register - allows to build the different occupational profiles and to produce wide information on main demographic characteristics of employees and employment relationships.

RACLI is a thematic register on the labour market and an extension of the information system on employment, with reference to wages and inputs of labour. It represents the main innovation within the structural statistics on labour costs and it is mainly based on social security sources.

Finally, Extended Frame-SBS is the register for annual economic data on all active enterprises, based on administrative data integrated with the main surveys on enterprises. This database provides a detailed and multidimensional mapping of enterprises for structural and dynamic analysis of production equipment.

3. Preliminary results

In 2014, the Italian private enterprises are 4 million and 359 thousand, for a total of 4 million 721 thousand local units and 16 million 189 thousand employees. 78% of these enterprises work in the services sector and hires 67% of the total employees. The Italian enterprises system is characterizes by having more small or micro enterprises than the EU average: 95.4% have less than 10 employees and, all together, they employ 46.3% of the total private dependent labour force.

Job positions in the active enterprises correspond to: i) 11.3 million of dependent workers, ii) almost 5 million of independent, iii) 345 thousand of external and iv) 175 thousand of temporary workers. Among the dependent workers, 3 out of 4 occupy a full-time position and 9 out of 10 have a permanent contract.

[2] Data are available for 2011-2014 years and, referring to each individual job position, can be aggregated, for example, at the municipal level, allowing statistically significant estimates.

Our analysis is focused on unilocalised enterprises of the private sector with less than 50 employees and with dependent workers, amounting to 1 million 404 thousands enterprises[3]. They represent about one third of private enterprises: the enterprises with dependent employees amount to 1 million 463 thousands and, among them, 1 million 404 thousands have local units only within the municipality of registration[4]. In these companies, the job positions are 7 million 847 thousands.

The wages map (Figure 1- first map[5]) highlights the existence of several extended areas with wages average levels higher than the national mean (HH clusters). It follows the evidence that enterprises with higher value added are also those paying with higher wage levels and that their territorial distribution appears to be somewhat outlined (Cardinaleschi et al., 2015).

More than a fifth (21.9%) of the total municipalities (7,856[6]) belongs to HH extended areas and they are mainly concentrated in the North of Italy: in Lombardia, Emilia Romagna, Veneto and Bolzano they represent about half of the region municipalities (respectively 56.7%, 55.4%, 49.3% and 46.6%). On the other hand, less than a fifth (18.5%) of the municipalities belongs to LL areas, but the share reaches 79% in Calabria, 61% in Campania and represents almost half of the municipalities in Sicilia (47.8%), Puglia (46.3%) and Basilicata (45.4%). Quite small is the number of municipalities representing a point of discontinuity, because they have values of wages higher (1.4%) or lower (1.7%) than the contiguous municipalities; the share of isolated municipalities is only 0.2%.

The wages map is quite similar to the labour cost map (Fig.1-third map). The areas with higher values (19.6% of the municipalities) are mainly in Lombardia (53.3% of the region municipalities), Emilia-Romagna (48.2%) and, above all, Bolzano (92.2%). In the Centre-South there are almost exclusively large areas with labour costs lower than the average (19.5% of the municipalities): Calabria, Campania, Puglia and Sicilia present the highest share of municipalities belonging to the LL cluster (73.1%, 66.2%, 58.8% and 56%).

Also the second map (added value per employee) show a similar profile, even if the LL areas seem to be less wide than in the other two maps. It highlights the existence of several extended areas with high value added per employee (cluster HH),

[3] Waiting for the availability of the Frame territorial extension (Barbieri et al. 2017), the analysis is restricted to unilocalised active enterprises with less than 50 employees and with dependent employee. They represent about 33% of the total: on 4.264 thousands active enterprises, 4.240 thousands have less than 50 employees; they decreases to 1.463 thousands if only enterprises with dependent employees are considered; finally, 1.406 are unilocalised (not having local units outside the municipality of registration).

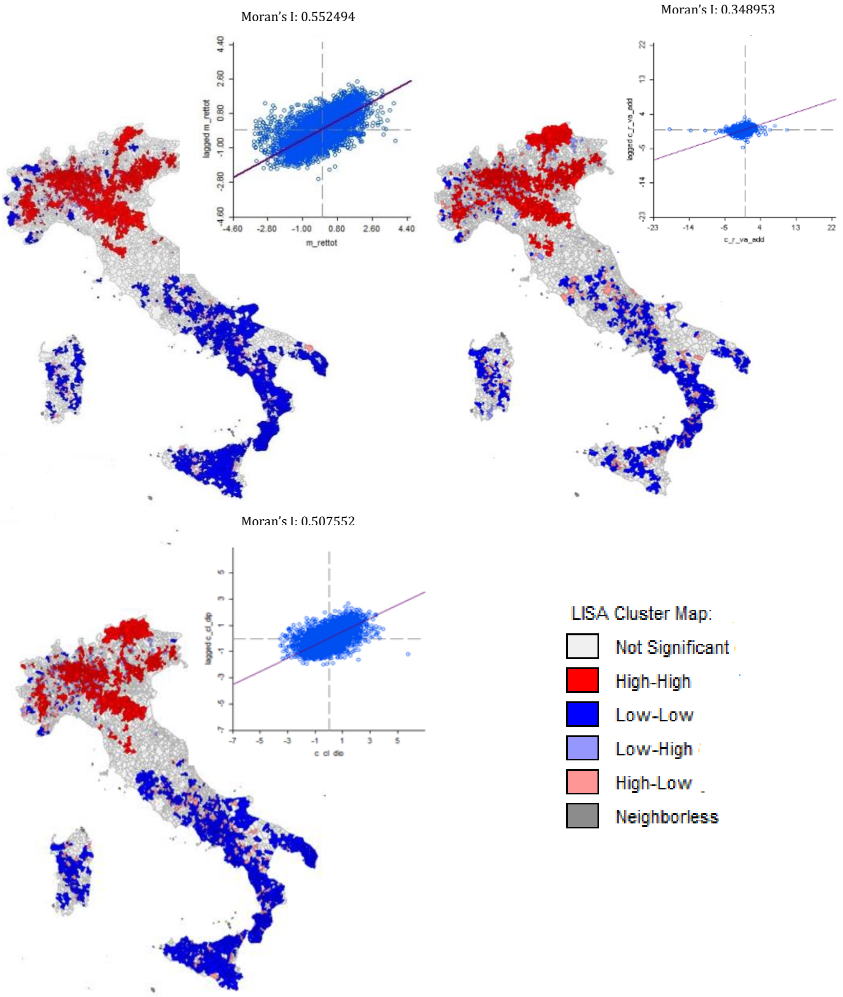
[4] It refers to the legal address municipality.

[5] By the Moran index, the municipalities can be clustered as follows (in addition to the cluster of municipalities not giving a significant contribution to global autocorrelation and to the cluster of municipalities not confining with other municipalities - minor islands): 1. High-High (HH)- contiguous units with values of the variable *x* (wage level) higher than the mean; 2. Low-Low (LL) - contiguous units with values of the variable *x* lower than the mean; 3. High-Low (HL) - units with values of the variable *x* higher than their contiguous units; 4. Low-High (LH) - units with values of the variable *x* lower than their contiguous units. The first two clusters are composed of homogeneous units (referring to the variable of interest), while the last two clusters concern abnormal cases or *enclaves* (Anselin, 2002).

[6] The municipalities' number is lower than the total because in some municipalities there are not active enterprises; moreover other 20 municipalities are excluded from the first map because they represent outliers in consequence of their negative values for the value added per employee.

that contain about a fifth (19.5%) of the municipalities. These areas represent almost half of the region municipalities in Lombardia (49.1%), Emilia Romagna (44.3%) and Veneto (40.6%) and the share rises to 92.2% in Bolzano.

Figure 1: High/low wages (first map), value added per employee (second map) and labour cost (third map). Years 2014



On the other hand, the extended areas with value added per employee lower than the average include over one-tenth of the total municipality (12.2%). They represent about one-fifth of Lazio's municipalities, a fourth in Campania, Puglia, Sicilia and Sardegna and more than 40% in Molise; the share reaches 60.7% in Calabria.

Small is the number of municipalities representing a point of discontinuity: 1.6% those having a values added per employee higher than the contiguous municipalities and 2.2% those having a lower value; the share of isolated municipalities stays at 0.2%. Obviously, for all the considered maps the extended areas with values higher/lower than the average do not necessarily present the same levels.

This kind of analysis can be useful in policies implementation. Just to give an example, it can support the European Structural and Investment Funds design, taking into consideration Classification of municipalities according to degree of remoteness (De Santis et al., 2017). Less than one fourth of the analysed municipalities (22.3%) are in Peripheral or Ultra-peripheral areas [7]", but the share decreases to 2.7%, 4.1% and 4.5% among the HH cluster, considering the wages, the cost of labour and value added per employee respectively. For the LL municipalities, the share reaches 47.6%, 40.7% and 48.2%.

After having individuated areas with wages, labour cost and productivity levels higher or lower than the mean, it is possible to study the enterprises and their employees, contributing to clarify the mechanisms underlying the relationship between local wage differentials, salary rigidity and economic performance.

In order to identify the factors that mainly contribute to determine wage levels in Italy, a regression model can be estimated. Exploiting the information asset coming from the integration of different sources, it is possible to study the joint distribution of enterprise and worker main characteristics on different areas of the Country.

Such a detailed view of the territory and of the distribution of different segments of workers reveals particular, sometimes unexpected, profiles, supporting the definition and the implementation of local policies that are recognized as a crucial option within the broader context of regional policies.

Some references

1. Anselin, L.: Under the Hood. Issues in the Specification and Interpretation of Spatial Regression Models, Regional Economics Application Laboratory (REAL),University of Illinois (2002)
2. Barbieri, G.A., Faramondi, A., Truglia, F.G.: La stima del valore aggiunto a livello territoriale fine: nuovi sviluppi nell'ambito delle statistiche strutturali, Nuova Serie W.P. n. 16, Certet (2017)
3. Cardinaleschi, S., De Santis, S., Shenkel, M., Truglia, F.G.: Un'approccio geostatistico all'analisi dei dati di bilancio – un'analisi panel-spaziale dell'efficienza delle imprese italiane. Arcavarcata di Rende (2015)
4. De Santis, S., Freguia, C., Masi, A., Pannuzi, N., Truglia, F.G.:Wages Differentials and Their Determinants: a Spatial Analysis. Paper presented at the International Conference on Inequality, Trends in inequality: social, economic and political issues. 2-4th November 2017, Bologna (2017).
5. ISTAT: I differenziali retributivi nel settore privato. Anno 2014. Statistiche report, 30/12 dicembre (2016)
6. Porcari, S., Devicienti F. (eds.): Differenziali salariali regionali e performance economica "Monografie sul Mercato del lavoro e le politiche per l'impiego", n. 5 ISFOL – RP(MDL) (2007)

[7] http://www.agenziacoesione.gov.it/it/arint/Cosa_sono/index.html

The Transition to Motherhood among British Young Women: Does housing tenure play a role?

Diventare madri fra le giovani donne bruitanniche: la proprietà dell'abitazione gioca un ruolo?

Valentina Tocchioni, Ann Berrington, Daniele Vignoli and Agnese Vitali[1]

Abstract A positive link between homeownership and fertility is usually presumed. Nevertheless, couples' preferences to become homeowners before having their first child has been undermined by the dramatic changes in the UK housing market over recent decades, causing a marked fall in homeownership rates among young adults. Using prospective longitudinal data from the British Household Panel Survey (1991-2008) and the United Kingdom Household Longitudinal Survey (2009-2016), and applying multilevel discrete-time event-history techniques, we investigate whether and how the link between housing tenure and motherhood has changed over recent decades in Britain, and whether the link is moderated by local area characteristics including housing markets.

Abstract *Secondo la letteratura, la relazione sussistente fra la proprietà dell'abitazione e la fecondità è positiva. Tuttavia, la tendenziale preferenza delle coppie a divenire proprietarie di abitazione prima di diventare genitori è stata minata dai drammatici cambiamenti intercorsi nel mercato immobiliare britannico negli ultimi decenni, determinando una sostanziale caduta dei tassi di proprietà delle case tra i giovani adulti. Usando i dati longitudinali del British Household Panel Survey (1991-2008) e del United Kingdom Household Longitudinal Survey (2009-2016), e applicando modelli di sopravvivenza multilivello (tempo discreto), l'obiettivo è di valutare se e come la relazione sussistente tra il possesso dell'abitazione e la nascita del primo figlio sia cambiata negli ultimi decenni in Gran Bretagna, e se questa sia moderata dalle caratteristiche del mercato immobiliare a livello locale.*

[1] Valentina Tocchioni, University of Florence; email: v.tocchioni@disia.unifi.it
Ann Berrington, University of Southampton; email: A.Berrington@soton.ac.uk
Daniele Vignoli, University of Florence; email: vignoli@disia.unifi.it
Agnese Vitali, University of Southampton; email: A.Vitali@soton.ac.uk

Key words: Britain, fertility, housing tenure, event-history analysis, panel data, multilevel models

1 Introduction

Housing markets affect fertility both directly and indirectly [7]. Direct links from housing to childbearing include the preference to be a homeowner before having children [5, 6, 9]. Even more, housing markets act on fertility through their impact on the ability of young adults to become residentially independent of the parental home. In western countries including the UK, family formation whilst co-residing in the parental home is unusual – family formation usually either coincides with, or follows residential independence from parents. Thus parent adult-child co-residence *delays* partnership formation and hence first births [3, 8].

Among young people, rates of homeownership in the UK have plummeted over the past 25 years (from 67% of 25-34 year olds in 1991 to 36% in 2014) [10]. This is a result of a number of factors not least the increase in house prices over the period, coupled with stagnation or decline in wages and security of employment [2]. The increased uncertainty in young people's lives thus makes private renting the only affordable solution for many young adults who decide to live independently [12]. Unlike in other European countries where private rented accommodation is highly regulated, tenants in the UK have very few rights, dealing with an expensive and very insecure dwelling. Private rented accommodation does not provide the same level of security of tenure as compared to either homeownership or social renting. Everything else being equal we might therefore expect childbearing to be postponed whilst in private rented accommodation, until security of tenure, either within owner occupation, or the social rented sector can be met. However, as the possibility of homeownership continues to recede for most young adults more are still renting at ages when entry into parenthood traditionally occurs.

Housing affordability and availability vary geographically due to local variations in housing costs, average wages and in the stock of social housing available in an area [11]. It is thus important to consider regional differences in entry into parenthood, and the potential role of local housing markets in influencing fertility behaviour.

To sum up, in the UK homeownership is increasingly unaffordable and many young adults are living in privately rented accommodation at later ages, and at stages in their life course when family formation typically takes place. We therefore ask the question as to whether the positive link between homeownership and fertility remains in recent years, which is an issue that needs still to be addressed in Britain. Indeed as noted previously for France, the cost of homeownership might compete with the costs of childbearing and childrearing, favouring childbearing among renters [4]. More specifically, we address the following questions: *Has the probability of becoming a mother whilst in private rented accommodation as opposed to owner occupier or social rented increased since 1991? Is this increase explained by the* *socio-economic and demographic characteristics of women in the different housing tenures? To what extent does the relationship between housing tenure and the progression to the first birth differ according to the local house prices?*

The paper follows with the description of the data and methods used for the analyses. Then, main results are presented, according to the three research questions. A concluding section closes the paper.

2 Data and methods

The study is based on a sample of women generated from the eighteen waves of the British Household Panel Survey (BHPS), in combination with data from the first seven waves of Understanding Society, the UK Household Longitudinal Study (UKHLS). In total our time period spans 1991-2016. We consider all women aged 18-42, who are living independently of the parental home and who were interviewed at least for two consecutive waves.

Multilevel discrete-time event-history analysis [1] is applied to exploit the transition to the first child's conception resulting in a live birth, taking account of the hierarchical structure of the data with women clustered within local authority districts (LADs). Woman's age is the baseline hazard. Observation starts at the first interview living independently after 18, and is censored either when the woman has a first conception, at last wave, when the woman retreats from the survey, when she returns to the parental home, or when she reaches age 44, whichever occurs first. Observations are weighted in order to take into account the changing population target over time. The key explanatory variable is current housing tenure, distinguishing among homeowners, private renters, and social renters.

3 Results

Table 1 shows the predicted probability from the hazard models of conceiving a first child according to housing tenure and calendar period. The predicted probability of conceiving a first child is significantly higher for women who are homeowners compared to those who are private renters until 2012; then, for the years 2013-2016 no significant difference emerges between homeowners and private renters. At first sight, the positive link between homeownership and fertility - which emerges clearly during the 1990 and 2000 - has changed over the last decade, with an increased proportion of private renters who have a first child in the last years.

Table 1: Predicted probabilities and confidence intervals of conceiving a first child according to housing tenure and calendar period. 1991-2016.

Housing tenure #Calendar period	Pred. Prob.	Confidence interval
owner#1991-1999	0.083	[0.072; 0.093]
owner#2000-2007	0.066	[0.056; 0.075]
owner#2008-2012	0.084	[0.073; 0.095]
owner#2013-2016	0.057	[0.046; 0.068]
social rent#1991-1999	0.049	[0.023; 0.075]
social rent#2000-2007	0.070	[0.037; 0.104]
social rent#2008-2012	0.066	[0.041; 0.090]
social rent#2013-2016	0.031	[0.010; 0.052]
private rent#1991-1999	0.037	[0.024; 0.050]
private rent#2000-2007	0.035	[0.023; 0.047]
private rent#2008-2012	0.048	[0.037; 0.058]
private rent#2013-2016	0.036	[0.025; 0.047]

Controlling for age group.

After controlling for all confounders (see Table 2), previous results are not confirmed for the last period, and the predicted probability of conceiving a first child is significantly higher for homeowners compared to private tenants (but only slightly). Nevertheless, over the last 25 years the predicted probabilities have decreased for homeowners, whereas the proportion of young adults who are conceiving a first child as private tenants has increased, net of all confounders.

Table 2: Predicted probabilities and confidence intervals of conceiving a first child according to housing tenure and calendar period. 1991-2016.

Housing tenure#Calendar period	Pred. Prob.	Confidence interval
owner#1991-1999	0.067	[0.058; 0.076]
owner#2000-2007	0.049	[0.042; 0.057]
owner#2008-2012	0.066	[0.057; 0.076]
owner#2013-2016	0.043	[0.034; 0.052]
social rent#1991-1999	0.032	[0.014; 0.049]
social rent#2000-2007	0.045	[0.022; 0.068]
social rent#2008-2012	0.047	[0.028; 0.065]
social rent#2013-2016	0.023	[0.007; 0.038]
private rent#1991-1999	0.023	[0.014; 0.032]
private rent#2000-2007	0.021	[0.013; 0.028]
private rent#2008-2012	0.032	[0.024; 0.039]
private rent#2013-2016	0.025	[0.017; 0.033]

Controlling for age group, partnership, education, parental social class, economic activity, equivalised income (in quintiles), overcrowding, and country of birth outside UK.

In a final step, we add a second-level covariate that takes into account the variability in house prices among the LADs, through the selling price of the 25th percentile house in a given LAD among the 20% lowest (1st quintile), up to 20% most expensive (5th quintile) across the country.

We find that, among homeowners, the probability to conceive is highest in LADs where the house prices of the least expensive houses lies in the 4th quintile of the LAD's distribution. In other words, homeowners will be more likely to have a first child if they own a house in an area where the house prices are higher than the country's average. Among private renters, the probability to have a first child is highest in LADs where the house prices of the least expensive houses lies in the 3rd quintile of the LAD's distribution. Moreover, in this area that represents the country's average for house prices, the higher probability of becoming a mother whilst as homeowner as opposed to private renter is only slightly significant. To sum up, once taken into account women's socio-economic and demographic characteristics, as well as household features and contextual information, the predicted probability of conceiving a first child remains higher for homeowners compared to private tenants, but the context plays a role in shaping the propensity to childbearing in private rented accommodation as opposed to owner occupier.

Table 3: Predicted probabilities and confidence intervals of conceiving a first child according to housing tenure and lower quartile house prices. 1991-2016**.**

Housing tenure#House prices	Pred. Prob.	Confidence interval
owner#1° quintile	0.055	[0.045; 0.065]
owner#2° quintile	0.049	[0.039; 0.058]
owner#3° quintile	0.056	[0.046; 0.066]
owner#4° quintile	0.075	[0.063; 0.087]
owner#5° quintile	0.053	[0.044; 0.063]
social rent#1° quintile	0.036	[0.019; 0.054]
social rent#2° quintile	0.035	[0.014; 0.057]
social rent#3° quintile	0.050	[0.021; 0.079]
social rent#4° quintile	0.054	[0.022; 0.086]
social rent#5° quintile	0.026	[0.009; 0.042]
private rent#1° quintile	0.026	[0.016; 0.036]
private rent#2° quintile	0.022	[0.013; 0.031]
private rent#3° quintile	0.034	[0.022; 0.045]
private rent#4° quintile	0.026	[0.017; 0.034]
private rent#5° quintile	0.024	[0.017; 0.031]

Controlling for age group, calendar period, partnership, education, parental social class, economic activity, equivalised income (in quintiles), overcrowding, and country of birth outside UK.

4 Conclusion

In the UK there has been a rapid increase in the last couple of decades in the number of young people living in private rented accommodation into their late twenties and thirties. This paper suggests that as a result of this increase the likelihood of becoming a parent whilst in private rented accommodation has increased slightly. By linking the individual survey data to administrative data on house prices by local authority district, we have shown that local housing contexts do play a role. Further research is required to better understand the possible expansion of having a first child whilst in private rented accommodation among other social strata of the population, as well as the potential consequences on young families' social security as private renters in place of homeownership in the years to come.

References

1. Barber, J.S., Murphy, S.A., Axinn, W.G., Maples, J.: Discrete-time multilevel hazard analysis. Sociol. Methodol. 30, 201--235 (2000) doi:10.1111/0081-1750.00079
2. Berrington, A., Duta, A., Wakeling, P.: Youth social citizenship and class inequalities in transitions to adulthood in the UK. CPC Working Paper 81. https://eprints.soton.ac.uk/405269/ (2017)
3. Berrington, A., Stone, J.: Young adults' transitions to residential independence in Britain: the role of social and housing policy. In: Antonucci, L., Hamilton, M., and Roberts, S. (eds.) Young People and Social Policy in Europe: Dealing with Risk, Inequality and Precarity in Times of Crisis, pp. 210–235. Palgrave Macmillan, Basingstoke (2014)
4. Courgeau, D., Lelièvre, E.: Interrelations between first home-ownership, constitution of the family, and professional occupation in France. In: Trussell, J., Hankinson, R., Tilton, J. (eds) Demographic Applications of Event History Analysis, pp. 120-140. Clarendon Press, Oxford (1992)
5. Feijten, P., Mulder, C.H.: The timing of household events and housing events in the Netherlands: A longitudinal perspective. Hous. Stud. 17, 773–792 (2002)
6. Kulu, H., Steele, F.: Interrelationships Between Childbearing and Housing Transitions in the Family Life Course. Demography 50, 1687--1714 (2013). doi:10.1007/s13524-013-0216-2
7. Mulder, C.H.: Population and housing : A two-sided relationship. Dem. Res. 15(13): 401--412 (2006)
8. Mulder, C.H.: Home-ownership and family formation. J. Hous. Built Environ. 21, 281--298 (2006) doi:10.1007/s10901-006-9050-9
9. Mulder, C.H., Wagner, M.: The connections between family formation and first-time home ownership in the context of West Germany and the Netherlands. Eur. J. Popul. 17, 137--164 (2001) doi:10.1023/A:1010706308868
10. ONS: UK Perspectives 2016: Housing and home ownership in the UK, https://visual.ons.gov.uk/uk-perspectives-2016-housing-and-home-ownership-in-the-uk/. Accessed 3 July 2017
11. ONS: Housing affordability in England and Wales: 1997 to 2016, https://www.ons.gov.uk/peoplepopulationandcommunity/housing/bulletins/housingaffordabilityinenglandandwales/1997to2016#affordability-gap-widens-over-time. Accessed 3 July 2017
12. Rugg, J.J.: Young people and housing: the need for a new policy agenda. https://www.jrf.org.uk/report/young-people-and-housing-need-new-policy-agenda (2010)

Finance & Insurance

Robust statistical methods for credit risk

Metodi statistici robusti per il rischio di credito

A. Corbellini, A. Ghiretti, G. Morelli and A. Talignani

Abstract Credit risk is a relevant problem faced by banks and financial institutions. The traditional statistical models which are generally used to quantify the credit risk present several drawbacks. First, in their standard versions they are not robust and do not take into account that the data may be corrupted by several outliers. Second, when a parametric model is employed, the variable selection procedure might be severely affected by the so called masking and swamping effects. This work extends robust statistical methods to credit risk analysis, showing how the traditional approach can be greatly improved through robust methods.

Abstract *La gestione del rischio di credito è un problema particolarmente rilevante per tutte le banche e le istituzioni finanziarie. I modelli statistici tradizionalmente utilizzati per quantificare tale rischio presentano diversi svantaggi. Quando il dataset contiene alcuni outliers il fit che si ottiene attraverso i metodi di stima standard può risultare distorto e inconsistente. Inoltre, il metodo di selezione delle variabili può essere severamente influenzato dagli effetti di masking e/o swamping. L'obiettivo di questo lavoro è quello di estendere i metodi statistici robusti all'analisi del rischio di credito, mostrando come l'analisi può essere fortemente migliorata utilizzando un approccio robusto.*

A. Corbellini
Department of Economics and Management, University of Parma, Via J. F. Kennedy, 6, Parma, e-mail: aldo.corbellini@unipr.it

A. Ghiretti
Department of Statistics, Computer Science, Applications, University of Florence, Viale Morgagni, 59, Firenze e-mail: ghiretti@disia.unifi.it

G. Morelli
Department of Economics and Management, University of Parma, Via J. F. Kennedy, 6, Parma, e-mail: gianluca.morelli@unipr.it

A. Talignani
Department of Economics and Management, University of Parma, Via J. F. Kennedy, 6, Parma, e-mail: andrea.talignani@studenti.unipr.it

Key words: Credit Risk, Forward Search, Lasso, Outliers, Robustness

1 The Credit Risk framework

Credit risk is defined as the risk of default on a debt that may arise from a borrower failing to make the required payments. In order to quantify how likely a borrower will be unable to meet his debt obligation, it is customary to use the so called probability of default (PD). The PDs have been introduced in the the Basel agreements, that took into account new developments in the measurement and management of banking risks for those institutions that agreed to use the "internal ratings-based" (IRB) approach. In this approach, financial institutions and banks are allowed to use their own internal measures as primary inputs to the capital calculation. These measures, require the estimation of a set of indexes that describe the risk exposure of the institution. These risk measures are subsequently converted into risk weights and further into regulatory capital requirements by means of risk weight formulas specified by the Basel Committee. In this work we will focus on the estimate of the PD, showing how the standard statistical methods generally proposed in literature can be greatly improved when a robust approach is adopted. In literature several statistical models have been proposed to estimate the PD. Some commonly adopted examples are: logistic regression models, discriminant analysis, classification trees and so on. The main drawback of all of these procedures is that they are not robust against slight deviations from the model assumptions. In fact, real data are generally corrupted by a random number of outlying units, i.e, units that do not share the same characteristics of the majority of data. It is well known in the literature that neglect outliers might have a severe impact on the analysis, leading to biased and inconsistent estimates and misleading inference. Furthermore, when facing real data, the analyst is generally required to perform a variable selection, as many are often available, but some might not be relevant to drive the PD. As a consequence, when a parametric model such as the logistic regression is adopted, it is common practice to employ a variable selection technique. However, standard variable selection techniques such as stage-wise algorithms or the widespread LASSO may be seriously affected by few outliers, leading to miss-selected variables.

To show the great improvements that can be achieved with a robust analysis our work will consider a real data set made available by an Italian Bank, which we kindly acknowledge, but do not report for confidentiality.

2 A Forward Search approach to the LASSO

We consider a set of observations $z_i = (y_i, x_i'), i = 1, \dots, n$, where y_i is a binary variable and x_i' is a vector of p features relative to the ith company.

We estimate the default probability of a firm by a logistic discriminant function, that is,

$$Pr(y_i = 1|x_i) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} = \frac{1}{1+e^{-x_i'\beta}} \qquad i = 1,\ldots,n.$$

$x_i' = (x_{i1},\ldots,x_{ip})$ is the vector of predictors, i.e. firm-specific characteristics and financial indexes and β is a vector of p unknown parameters.
The parameters are generally estimated by maximizing the log-likelihood function,

$$\ell(\beta) = \sum_{i=1}^{n}[y_i x_i'\beta - \log(1+e^{x_i'\beta})]. \tag{1}$$

In the data set considered one of the major concerns was to extract, from the p original features, a subset of k highly significant predictors. The LASSO (Hastie and Tibshirani, 2009) [2], performs a variable selection by means of a regularization or penalization parameter.

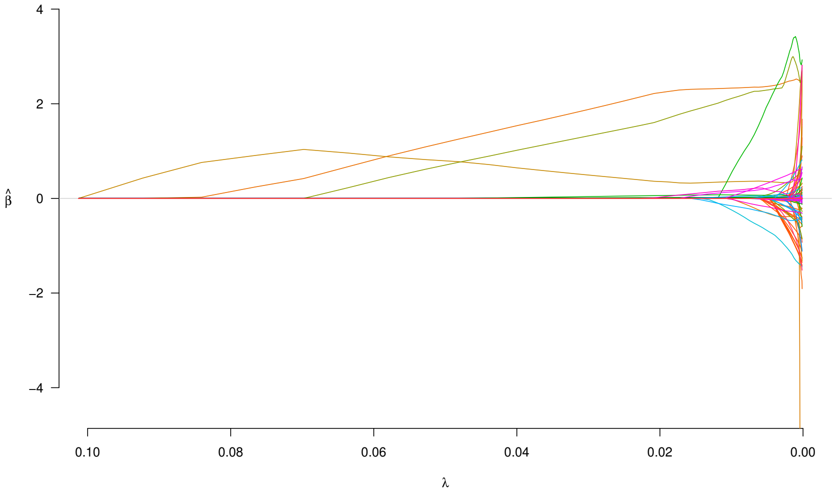


Fig. 1 Plot of the trajectories of the estimated coefficients, as a function of λ, obtained with the group LASSO with no outliers

When the LASSO is applied to the logistic model, the resulting penalized version of the log-likelihood function to be maximized becomes

$$\ell_{\lambda}^{L}(\beta) = \sum_{i=1}^{n}[y_i x_i'\beta - \log(1+e^{x_i'\beta})] + \lambda \sum_{j=1}^{p} |\beta_j|.$$

By selecting a value of λ sufficiently large, the L_1 penalty used in the LASSO, forces some of the coefficient estimates to be exactly equal to zero. The selection of λ is generally performed by cross validation and in the trivial case $\lambda = 0$ the log-likelihood (1) is maximized.
We will denote the optimal value of λ obtained by cross validation with λ_{optim}.
In our case, we have p predictors, h of which are qualitative (with $h \ll p$), and it is common to code those variables with dummies. However, it might happen that, a group of dummies represents the same variable, therefore it is compulsory to assume that the p predictors belong to G distinct groups. Notice that each quantitative variable form a distinct group. The parameter ρ_g is such that we either select or neglect the entire group. This is the standard approach suggested by Yuan e Lin (2007) [4] via the group LASSO, expanded by Meier, Van De Geer and Bühlmann (2008), [3] where groups of features can be included together into or out of a model. The group LASSO estimator is obtained by maximizing the following penalized log-likelihood function

$$\ell_{\lambda}^{GL}(\beta) = \sum_{i=1}^{n}[y_i x_i'\beta - \log(1+e^{x_i'\beta})] + \lambda \sum_{g=1}^{G} \sqrt{\rho}_g ||\beta_g||_2$$

where β_g is the vector of parameters associated with the predictors in group g.
In our work we propose a novel use of the Forward Search , see Atkinson and Riani (2000) [1], coupled with a LASSO regularization technique. This allows to detect multiple outliers and perform a selection of the most significant explanatory variables in a robust way.

In order to detect outliers and departures from the fitted regression model, the Forward Search uses the group LASSO to fit the model to subsets of m observations. The initial subset of m_0 observations is chosen robustly as follows:

1 fit the group-LASSO and select λ_{optim} by performing a cross validation

2 obtain the residuals for all the observations given the k features obtained at the previous step

3 sort all the residuals

4 repeat steps from 1-3 ten-thousand times and store the results in a data matrix

5 the subset of units that minimize the median of the residuals is selected.

After the initial subset has been obtained, the search procedure starts. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest residuals. For each m ($m_0 \leq m \leq n-1$), by a graphical monitoring of the maximum standardized residual among the units included in the set,

$s_{i,max}$, and the minimum standardized residual for the units excluded from the fit, $s_{i,min}$, it is possible to detect a potential outlier.
It is worth to be noted that, as outlined at step 2, at each iteration of the Forward Search, the residuals are calculated for all units, notwithstanding the fact that these residuals are stemming from different sets of variables. Moreover, since in the Forward Search at each step the residuals are ordered, and several units or groups of units might join and leave the subset, some problems concerning the identification of the model may arise. To overcome this drawback we impose that in every subset generated by the search at each step, there will be a percentage of defaulted firms equal to the original percentage found in the whole data. The defaulted firms included in the subset are selected among those with smallest residual.

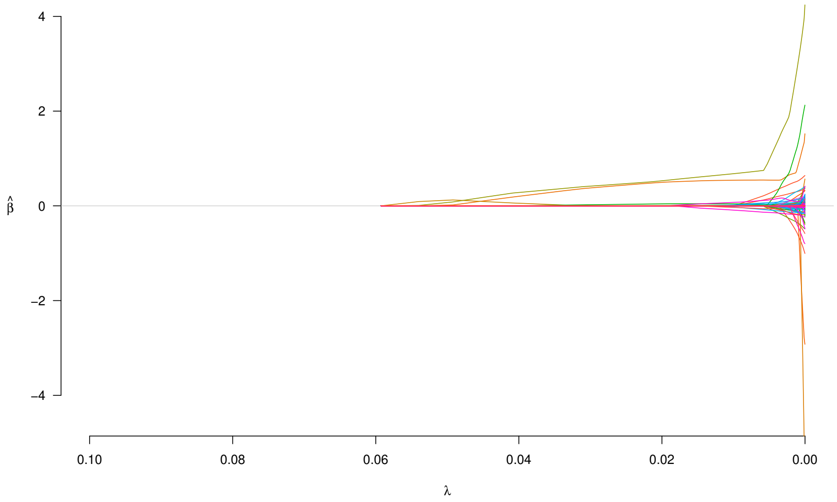


Fig. 2 Trajectory plot of the estimated coefficients, as a function of λ, obtained with the group LASSO when a 15% contamination is introduced

3 Preliminary results

As we have previously mentioned standard statistical techniques can be seriously affected by the presence of outliers in the data.
In order to show the impact that a small contamination might provoke on the group LASSO we introduce a 15% contamination into the Bank dataset and subsequently we perform the group-LASSO on the contaminated data.

The contamination is performed by adding a percentage of 15% additive outliers to the response variable Y. Figure 1 and Figure 2 show the trajectories of the estimated parameters over different values of lambda, before and after the contamination.
The different trajectories in the two plots show clearly that when some contamination is introduced the spurious units affect consistently the variable selection technique.
Figure 3 highlights the effect that the spurious units introduced have on the cross validation. First, the λ_{optim} selected in the contaminated scenario results lower than the one selected with the clean data. Second, the cross validation error in the contaminated case results sensibly larger for all the values of λ. Last, but not least, the number of groups selected with the introduction of the additive outliers drops remarkably to a number of 14.

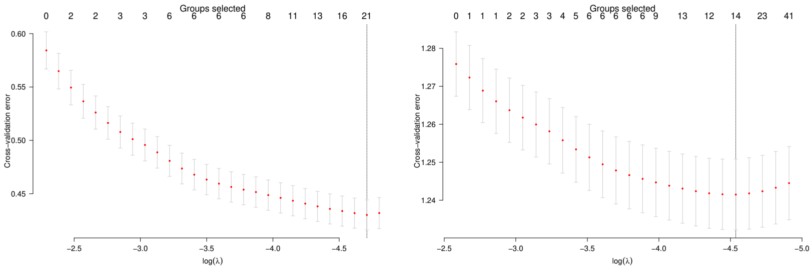


Fig. 3 Plots of λ estimates. Left panel, uncontaminated data, right panel, 15% of contamination

The significant drop in the number of selected groups as well as the increase in the cross validation error, suggest that there were swamping and masking effects induced by the introduction of several outliers.
The aim of our work is to perform a robust calibration of the value of λ whose choice is not affected by the spurious units and, at the same time, identify the most influential units by means of the forward plots.

References

1. Atkinson, A., Riani, M.: Robust Diagnostic Regression Analysis. Springer Verlag (2000).
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2009).
3. Meier, Lukas and Van De Geer, Sara and Bühlmann, Peter: The group lasso for logistic regression. Journal of the Royal Statistical Society, Series B **70**(1), 53-71 (2008)
4. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, **68**(1), 49-67 (2007)

Depth-based portfolio selection

Data depth per la selezione di portafoglio

Giuseppe Pandolfo, Carmela Iorio and Antonio D'Ambrosio

Abstract The estimation of multivariate location and scatter is the cornerstone of the classical multivariate statistical methods widely used in portfolio selection problems. However, they are not robust. We propose to use an alternative non-parametric approach based on the weighted L^P depth as robust location and scatter estimator in order to deal with extreme events in asset returns analysis. We first review weighted L^P depth along with its main properties and then discuss its application to portfolio selection through a small simulation study.

Abstract *Le depth functions La stima dei parametri di location e dispersione rappresentano la pietra miliare dei classici metodi statistici multivariati utilizzati nella selezione di portafoglio. Tuttavia questi metodi non sono robusti. Proponiamo, quindi, l'utilizzo di un metodo non parametrico basato sulla weighted L^P depth come stimatore robusto di location e dispersione nel caso di eventi estremi nell'analisi dei rendimenti. Forniamo la definizione e le proprietlla weighted L^P depth*, e mostriamo la sua applicabilitla selezione di portafoglio attraverso uno studio di simulazione.

Key words: Data depth, Robust estimation, Financial assets, Contaminated model.

Giuseppe Pandolfo
Department of Industrial Engineering, University of Naples Federico II, Napoli, e-mail: giuseppe.pandolfo@unina.it

Carmela Iorio
Department of Industrial Engineering, University of Naples Federico II, Napoli, e-mail: carmela.iorio@unina.it

Antonio D'Ambrosio
Department of Economics and Statistics, University of Naples Federico II, Napoli, e-mail: antdambr@unina.it

1 Introduction

Portfolio selection concerns how to allocate capital over a number of assets to maximize the "return" on the investment and minimizing the "risk". The first mathematical model for portfolio selection was introduced by Markowitz (1952). In this model, the "return" on a portfolio is measured by the expected value of the random portfolio return, and the associated "risk" is given by the variance of the portfolio return. This mean-variance model has had a great impact on the economic modelling of financial markets and represents the milestone of the modern portfolio theory. Despite the success of the mean-variance model, its solutions are often very sensitive to perturbations in the parameters of the problem. Indeed, when the sample distribution deviates even slightly from the assumed distribution, the efficiency of classical estimators may be drastically reduced. Robust estimators, on the other hand, are not as efficient as maximum likelihood estimators when the underlying model is correct, but their properties are not as sensitive to deviations from the assumed distribution. In order to reduce the sensitivity of the Markowitz-optimal several techniques were proposed in the literature. Among the proposals we recall Vaz-de Melo and Camara (2005) who used the M-estimators, Perret-Gentil and Victoria-Feser (2005) who adopted a translated biweight S-estimator, while Welsch and Zhou (2007) used the minimum covariance determinant estimator and the winsorization. DeMiguel and Nogales (2009) proposed a class of policies that are constructed using both on M- and S-estimators. For an overview on the robust methods used for in portfolio selection we refer to Fabozzi, Huang and Zhou (2010).

In recent years, attention on portfolio selection strategies based on non-parametric and semi-parametric techniques have been also shown to exist (see e.g., Ben Salah et al., 2018 and Iorio et al., 2018)

Despite the growing interest on non-parametric and/or robust methods for portfolio selection, to the best of authors knowledge there is no literature on the exploitation of data depth functions to this purpose. Hence, in this paper we propose to adopt a non-parametric approach based on the weighted L^p depth to perform robust location and scatter estimation in financial applications.

In the following we first recall the definition of data depth function and review the notion of L^p depth, then the results of a small simulation are offered to the reader.

2 Data depth concept

Many statistical techniques in multivariate analysis assume normality of the distribution of the data. This assumption is often disputable and thus other, non-parametric, approaches are worth to be considered. One is based on the so called *data depth*, which is a way to measure the depth or outlyingness of a given point with respect to a multivariate data cloud or its underlying distribution. This concept was originally introduced to generalize the concepts of the median and the quantiles to a multivariate framework. The principle is very simple. For a distribution F in $\mathbb{R}^d$, a depth function, denoted by $D(x;F)$, provides a certer-outward ordering of points in $x \in \mathbb{R}^d$.

Zuo and Serfling (2000) provided general notions of depth function on $\mathbb{R}^d$ and presented four reasonable properties that a depth function (bounded and non-negative) should possess, that is:

P1 **Affine invariance**. The depth of a point $x \in \mathbb{R}^d$ should not be dependent on the underlying coordinate system or, in particular, on the scales of the underlying measurements.

P2 **Maximality at center.** For a distribution having a uniquely defined "center", the depth function should attain maximum value at this center.

P3 **Monotonicity relative to deepest point.** As a point $x \in \mathbb{R}^d$ moves away from the "deepest point" (the point at which the depth function attains maximum value) along any fixed ray through the center, the depth at x should decrease monotonically.

P4 **Vanishing at infinity.** The depth of a point x should approach zero as $\|x\|$ approaches infinity.

Let $\mathscr{P}$ denote the class of distributions on Borel sets on $\mathbb{R}^d$, while F_X denote the distribution of a given random vector X belonging to the class of random vectors X

There are several notions of data depth in the literature.

Definition 1. Let the mapping $D(\cdot,\cdot) : \mathbb{R}^d \times \mathscr{P} \to \mathbb{R}_+$ satisfy P1, P2, P3 and P4. That is, assume:

(i) $D(Ax+b, F_{AX+b}) = D(x, F_X)$ holds for any random vector $X \in \mathbb{R}^d$ and any $d \times d$ nonsingular matrix A, and any d dimensional vector b.

(ii) $D\left(\theta, F\right) = \sup_{x \in \mathbb{R}^d} D(x,F)$ holds for any $F \in \mathscr{P}$ having centre θ.

(iii) For any $F \in \mathscr{P}$ having deepest point θ, $D(x,F) \leq D(\theta + \alpha(x?\theta), F)$ holds for $\alpha \in [0,1]$; and

(iv) $D(x,F) \to 0$ as $\|x\| \to \infty$, for each $F \in \mathscr{P}$.

Then $D(\cdot;F)$ is called a statistical depth function.

The sample version of a depth function $D(\cdot;F)$ is denoted by $D(\cdot;F_n)$, where F is replaced with an empirical measure F_n, computed on a sample $X_n = \{x_1, \ldots, x_n\}$.

There are several notions of data depth function available in the literature. The halfspace, simplicial, Mahalanobis and L^p depths are some of the most popular ones. The notion of data depth has been also extended to the functional space (see e.g., Lopez-Pintado and Romo, 2009) and on the spheres (see e.g., Liu and Singh, 1992 and Pandolfo et al., 2017).

In this paper, we adopt the notion of weighted L^p depth introduced by Zuo (2004) because of its ease of computation and (local and global) robustness properties.

2.1 The weighted L^p depth

Zuo and Serfling (2000) defined the L^p depth based on the L^p-norm. Different distances (norms) were used with equal weights. However, in practice, the importance (weight, cost, penalty, or incentive) may not be the same for different distances (norms). This motivates to adopt this notion of depth, that is define as follows:

$$WL^pD(x;F) = \frac{1}{1+Ew\left(\|x-X\|_p\right)},$$

where w is a weight function on $[0,1)$, $X \sim F$ and $\|\cdot\|$ denotes the L^p-norm (when $p=2$ we have the Euclidean norm), w is assumed to be non-decreasing and continuous on $[0,\infty)$. The weighted L^p depth possesses some desirable properties of depth functions. It is translation invariant (can be affine invariant for $p=2$ under some modification), maximized at the center of a (centrally) symmetric distribution for convex w, decreasing when a point moves along a ray stemming from the deepest point, and vanishing at infinity. For more related discussions see Zuo and Serfling (2000).

The weighted L^p depth-induced medians (multivariate location estimator) are globally robust with the highest breakdown point for any reasonable estimator. The weighted L^p medians are also locally robust with bounded influence functions for suitable weight functions. Unlike other existing depth functions and multivariate medians, the weighted L^p depth and medians are computationally feasible and easy to calculate in high dimensions. The price to be paid is the lack of affine invariance.

3 Simulation study

In this section, we present a small Montecarlo simulation to investigate the performance of the weighted L^p depth-based estimators of the mean and covariance matrix, for both contaminated and non-contaminated simulated data.

Following Toma and Leoni-Aubin (2015), and DeMiguel and Nogales (2009), we use simulations to generate asset returns data following a distribution that deviates slightly from the normal distribution. Specifically, we considered the multivariate normal distribution $F \sim N(\mu_F, \Sigma_F)$ with mean $\mu_F = 0$ and Σ_F a $N \times N$ covariance matrix with variances equal to 1 and covariances all equal to 0.2, with N denoting the total number of assets. We generated samples of size $T = 100$ according to the following contaminated model:

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon G, \tag{1}$$

where G is a contaminating distribution and ε is the fraction of the data that follows the contaminating distribution $G \sim N(\mu_G, \Sigma_G)$ with $mu_G = -4$ and $\Sigma_G = 4\Sigma_F$. We considered $N \in \{2,5,10,20\}$ and three different contamination levels $\varepsilon \in \{0\%, 2.5\%, 5\%\}$ to investigate how the estimates change when the asset returns deviate from normality.

The estimates of location and scatter were obtained through the weighted L^2, with all the observations having same weight. More in detail, in case of location, the estimator is defined as:

$$L(F) = \int x w_1 \left(L^2 D(x,F)\right) dF(x) \Big/ \int w_1 \left(L^2 D(x,F)\right) dF(x),$$

then a weighted L^2 depth scatter estimator is defined as

$$S(F) = \frac{\int (x - L(F))(x - L(F))^T w_2(D(x,F))\, dF(x)}{\int w_2(D(x,F))\, dF(x)}$$

where w_2 are suitable weight function that can be different from w_1.

For each setting, we generated $R = 250$ samples and for each sample we computed the depth estimates of location and scatter. The performances are evaluated through the empirical mean squared error (EMSE) given by

$$EMSE = \frac{1}{R}\sum_{i=1}^{R} \left\|\hat{\theta}_i - \theta_0\right\|^2$$

where $\theta_0 = (\mu_F, vech(\Sigma_F))'$ and $\hat{\theta}_i = \left(\hat{\mu}_i, vech\left(\hat{\Sigma}_i\right)\right)'$ is an estimate corresponding to the i-th sample, while $vech(\Sigma)$ is "the vector half", namely the $N(N+1)/2$-dimensional column vector obtained by stacking the columns of the lower triangle of Σ, including the diagonal, one below the other.

Results are presented in Table 1. The mean squared errors generally increase along with the number of assets (i.e., the sample size). However, for low dimensions ($N = 2$ and 5), the L^2 depth-based method estimates appear to be less affected by the contamination.

4 Final comments

In this paper we suggest to exploit the use of the weighted L^p depth function to perform robust estimation in financial settings. The very first results obtained through simulations are promising. Further research are needed to determine how to assign (depth-)weights to the observations, and to investigate the behaviour in real data applications.

Table 1: Empirical mean squared errors of WL^2 depth estimates.

N	ε		
	0%	2.5%	5%
2	0.10	0.12	0.48
5	0.71	0.95	2.57
10	2.89	3.76	9.46
20	9.23	12.58	32.50

References

1. Ben Salah, H., Chaouch, M., Gannoun, A., de Peretti, C., Trabelsi, A.: Mean and median-based nonparametric estimation of returns in mean-downside risk portfolio frontier. Annals of Operations Research **262**, 653–681 (2018)
2. DeMiguel, V., Nogales, J.F.: Portfolio selection with robust estimation. Operations Research **57**, 560–577 (2009)
3. Fabozzi, F.J., Huang, D., Zhou, G.: Robust portfolios: contributions from operations research and finance. Annals of Operations Research **176**, 191–220 (2010)
4. Iorio, C., Frasso, G., D'Ambrosio, A., Siciliano, R.: A P-spline based clustering approach for portfolio selection. Expert Systems With Applications **104**, 88–103 (2018)
5. Liu, R.Y., Singh, K.: Ordering directional data: Concepts of data depth on circles and spheres. The Annals of Statistics **20**, 1468–1484 (1992)
6. Lopez-Pintado, S., Romo, J.: On the concept of depth for functional data. The Annals of Statistics **104**, 718–734 (2009)
7. Markowitz, H.: Portfolio selection. The Journal of Finance **7**, 77–91 (1952)
8. Pandolfo, G., Paindaveine, D., Porzio, G.C.: Distance-based depths for directional data. Working Papers ECARES 2017 – 35 (2017) Available at `https://ideas.repec.org/p/eca/wpaper/2013-258549.html`
9. Perret-Gentil, C., Victoria-Feser, M.P.: Robust mean-variance portfolio selection. FAME research paper 140 (2005) Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=721509`
10. Toma, A., Leoni-Aubin, S.: Robust portfolio optimization using pseudodistances. PLoS ONE (2015) Available at `e0140546.https://doi.org/10.1371/journal.pone.0140546`
11. Vaz-de Melo, B., Camara R.P.: Robust multivariate modeling in finance. International Journal of Managerial Finance **4**, 12–23 (2005)
12. Welsch, R.E., Zhou, X.: Application of robust statistics to asset allocation models. REVSTAT – Statistical Journal **5**, 97–114 (2007)
13. Zuo, Y., Serfling, R.: General notions of statistical depth functions. Annals of Statistics textbf28, 461–482 (2000)
14. Zuo, Y.: Robustness of weighted L^p-depth and L^p-median. Allgemeines Statistisches Archiv textbf28, 215–234 (2004)

Estimating large-scale multivariate local level models with application to stochastic volatility

Stima di modelli local level di grandi dimensioni con applicazioni ai modelli di volatilità stocastica

Matteo Pelagatti and Giacomo Sbrana

Abstract We derive the closed-form solution to the Riccati equation for the steady-state Kalman filter of the multivariate local linear trend model. Based on this result we propose a fast EM algorithm that provides approximated maximum likelihood estimates of the model's parameters and apply it to large-scale stochastic volatility models.

Abstract *Deriviamo la soluzione dell'equazione di Riccati relativa al filtro di Kalman in steady-state per il local linear trend multivariato. Utilizziamo tale risultato per proporre un algoritmo EM, che calcola un'approssimazione della stima di massima verosimiglianza dei parameteri del modello. Applichiamo tale metodo di stima a modelli di volatilità stocastica di grandi dimensione.*

Key words: Algebraic Riccati equation, state-space models, Kalman filtering, smoothing, EM, multivariate stochastic volatility.

1 Introduction

Multivariate stochastic volatility (SV) models are useful for portfolio managers only if they can be applied to portfolios of tens or hundreds of assets. Indeed, in the GARCH literature the most successful multivariate models are those, such as the Constant Conditional Correlation (CCC) and the Dynamic Conditional Correlation (DCC) that, by splitting the estimation processes is a sequence of computationally feasible steps, make the application to large portfolios possible. However, the SV

Matteo Pelagatti
Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy
e-mail: matteo.pelagatti@unimib.it

Giacomo Sbrana
NEOMA Business School, 1 Rue du Marechal Juin, 76130 Mont-Saint-Aignan, France e-mail: giacomo.sbrana@neoma-bs.fr

literature is till lacking of a computationally feasible procedure to estimate large scale models, possibly dividing the estimation process into sub-steps.

In this work we consider the multivariate extensions of the SV models of Harvey et al (1994) and Alizadeh et al (2002), which have a state-space representation as multivariate local linear trend models, provide the steady-state Kalman filter recursions by solving the related Riccati equation in closed form and propose an EM algorithm that approximates the (quasi) maximum likelihood estimate of SV model.

2 Main results

It is well known that, for time-invariant state-space models, the Kalman filter eventually converges to the steady-state solution such that the error covariance matrix satisfies the so called algebraic Riccati equation. Explicit solutions for such matrix equation are in general not available. A notable exception is represented by the univariate local level model (see Harvey, 1989) because the Kalman filter covariance matrix reduces to a scalar. In this paper we show that an analytical solution exists also for the multivariate local level model (also known as multivariate exponential smoothing or exponentially weighted moving average or EWMA). In what follows we use this convention: $\boldsymbol{M}$ is a matrix and $\boldsymbol{M}'$ is its transpose, $\boldsymbol{m}$ is a column vector such that $\boldsymbol{m}'$ is a row vector. A lower-case letter, such as x, represents a scalar. Finally, 0 is used indiscriminately for matrices, vectors and scalars.

Consider the time-invariant state-space representation of the multivariate exponential smoothing process:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim \mathrm{WN}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon) \\ \boldsymbol{\alpha}_{t+1} &= \boldsymbol{\alpha}_t + \boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim \mathrm{WN}(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta) \end{aligned} \tag{1}$$

where WN denotes a white noise sequence. In what follows it is assumed that $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\eta$ are symmetric positive definite matrices and $\mathbb{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\eta}_t') = 0$. Note that all vectors and matrices in (1) have dimension d and $d \times d$ respectively. The Kalman filter recursions for this model can then be written as follows:

Innovation : $\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{a}_t$,
Innovation variance : $\boldsymbol{F}_t = \boldsymbol{P}_t + \boldsymbol{\Sigma}_\varepsilon$,
Kalman gain : $\boldsymbol{K}_t = \boldsymbol{P}_t \boldsymbol{F}_t^{-1}$,
Prediction : $\boldsymbol{a}_{t+1} = \boldsymbol{a}_t + \boldsymbol{K}_t \boldsymbol{v}_t$,
Prediction error : $\boldsymbol{P}_t = \boldsymbol{P}_t - \boldsymbol{P}_t \boldsymbol{F}_t^{-1} \boldsymbol{P}_t + \boldsymbol{\Sigma}_\eta$.

In the steady-state the the covariance matrix $\boldsymbol{P}_t$ converges to the so called algebraic Riccati equation, that is,

$$\boldsymbol{P} = \boldsymbol{P} - \boldsymbol{P}(\boldsymbol{P} + \boldsymbol{\Sigma}_\varepsilon)^{-1} \boldsymbol{P} + \boldsymbol{\Sigma}_\eta, \tag{2}$$

where $\boldsymbol{P}$ is a symmetric positive definite matrix. The following proposition provides the analytical (matrix) solution of (2), that is the algebraic link between $\boldsymbol{P}$ and the pair $\boldsymbol{\Sigma}_\varepsilon$, $\boldsymbol{\Sigma}_\eta$.

Theorem 1. *Consider the system as in (1) where $\boldsymbol{\Sigma}_\eta$ is positive semi-definite while $\boldsymbol{\Sigma}_\varepsilon$ is strictly positive definite. Moreover, consider the following Cholesky decomposition $\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{M}\boldsymbol{M}'$ and defining $\boldsymbol{Q} = \boldsymbol{M}^{-1}\boldsymbol{\Sigma}_\eta \boldsymbol{M}^{-1'} = \boldsymbol{\Psi}\boldsymbol{\Delta}\boldsymbol{\Psi}'$ such that for the last eigendecomposition $\boldsymbol{\Psi}$ is a matrix of eigenvectors (i.e. $\boldsymbol{\Psi}\boldsymbol{\Psi}' = \boldsymbol{I}$) and $\boldsymbol{\Delta} =$* $\mathrm{diag}(\delta_1, \delta_2, \cdots, \delta_d)$ *is a diagonal matrix of eigenvalues. Then there exists a unique positive definite solution for $\boldsymbol{P}$. The solution is*

$$\boldsymbol{P} = \frac{1}{2}\boldsymbol{M}\boldsymbol{\Psi}\left[\boldsymbol{\Delta} + (\boldsymbol{\Delta}^2 + 4\boldsymbol{\Delta})^{\frac{1}{2}}\right]\boldsymbol{\Psi}'\boldsymbol{M}'. \tag{3}$$

Going back to the Kalman filter recursion, it is immediate to see that, in steady state, the only step to compute is the *prediction step*, because all the other quantities are time-invariant. Also the implementation of the smoothing process results greatly simplified. Indeed, the smoothing algorithm proposed by de Jong (1988, 1989) (see also Ansley and Kohn, 1985; Koopman, 1997) becomes $\boldsymbol{r}_n = \mathbf{0}$, $\boldsymbol{N}_n = \mathbf{0}$,

$$\boldsymbol{r}_{t-1} = \boldsymbol{F}^{-1}\boldsymbol{v}_t + \boldsymbol{L}'\boldsymbol{r}_t \tag{4}$$

$$\boldsymbol{N}_{t-1} = \boldsymbol{F}^{-1} + \boldsymbol{L}'\boldsymbol{N}_t\boldsymbol{L} \tag{5}$$

where $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{K}$ is also time-invariant.

The maximum likelihood estimation of a model in state space form using the EM algorithm is fully discussed by Shumway and Stoffer (2017) (see also Koopman, 1993; Durbin and Koopman, 2001). In practice, the EM algorithm for the multivariate local level can be implemented using the simple updating expressions (3.5), (3.6) and (3.7) of Section 3 in Koopman (1993). More specifically, for model (1) these expressions can be restated as follows:

$$\boldsymbol{\Sigma}_\varepsilon(\iota+1) = \boldsymbol{\Sigma}_\varepsilon(\iota) + \boldsymbol{\Sigma}_\varepsilon(\iota)\boldsymbol{\Theta}_e\boldsymbol{\Sigma}_\varepsilon(\iota) \tag{6}$$

$$\boldsymbol{\Sigma}_\eta(\iota+1) = \boldsymbol{\Sigma}_\eta(\iota) + \boldsymbol{\Sigma}_\eta(\iota)\boldsymbol{\Theta}_r\boldsymbol{\Sigma}_\eta(\iota) \tag{7}$$

where $\iota = 0, 1, \ldots$. Here $\boldsymbol{\Sigma}_\varepsilon(0)$, $\boldsymbol{\Sigma}_\eta(0)$ are the starting values. In addition,

$$\boldsymbol{\Theta}_r = \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{r}_t\boldsymbol{r}_t' - \boldsymbol{N}_t) \tag{8}$$

where $\boldsymbol{r}_t$ and $\boldsymbol{N}_t$ are constructed as in (4) and (5), and

$$\boldsymbol{\Theta}_e = \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{e}_t\boldsymbol{e}_t' - \boldsymbol{D}_t) \tag{9}$$

with

$$\boldsymbol{e}_t = \boldsymbol{F}^{-1}\boldsymbol{v}_t - \boldsymbol{K}'\boldsymbol{r}_t \tag{10}$$

and

$$\boldsymbol{D}_t = \boldsymbol{F}^{-1} + \boldsymbol{K}'\boldsymbol{N}_t\boldsymbol{K}. \tag{11}$$

Using the steady-state filter and smoother, we obtain a computationally feasible procedure to approximate the maximum likelihood estimation of the multivariate local level model.

Algorithm 1 (Approximate maximum likelihood estimation) *Fix arbitrary initial covariance matrices $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\eta$ and iterate as follows.*

1. *Obtain $\boldsymbol{P}$ from equation (3) and compute the steady-state Kalman filter matrices and start the Kalman filter with $\boldsymbol{a}_1 = \boldsymbol{y}_1$ (see Harvey, 1989, p. 26).*
2. *Run $\boldsymbol{a}_t = \boldsymbol{K}\boldsymbol{y}_t + (\boldsymbol{I} - \boldsymbol{K})\boldsymbol{a}_{t-1}$ and for $t = 2, \dots, n$.*
3. *Run the steady-state smoothing formulae (4), (5), (10), (11), (8) and (9).*
4. *Run the EM step updating the parameters using (6) and (7).*
5. *Go to step 1. until the likelihood increment is negligible.*

Table 1 compares the execution time and the precision of our algorithm to exact MLE on simulated vector time series of dimension up to $d = 20$. For $d = 100$ our algorithm provides estimation in few minutes, while it was not possible to get results for MLE.

Table 1 Comparisons of Kalman filter based maximum likelihood estimates versus estimates obtained by Algorithm 1 (MAD = mean absolute difference, MAE = mean absolute error).

	Execution time (sec)			MAD		MAE MLE		MAE Alg. 1	
d	Classic	Alg. 1	Ratio	$\boldsymbol{\Sigma}_\varepsilon$	$\boldsymbol{\Sigma}_\eta$	$\boldsymbol{\Sigma}_\varepsilon$	$\boldsymbol{\Sigma}_\eta$	$\boldsymbol{\Sigma}_\varepsilon$	$\boldsymbol{\Sigma}_\eta$
3	1.3	0.1	10.6	0.002	0.030	0.006	0.006	0.007	0.031
5	5.7	0.5	11.8	0.002	0.022	0.006	0.007	0.006	0.023
10	60.3	1.6	37.0	0.001	0.012	0.006	0.007	0.006	0.015
20	1490.4	10.1	147.8	0.001	0.006	0.006	0.008	0.006	0.011

3 Application to multivariate stochastic volatility models

In this section, we demonstrate how the stochastic volatility (SV) models of Harvey et al (1994) (from now on HRS) and Alizadeh et al (2002) (from now on ABD) can be successfully estimated on large portfolios using our Algorithm 1. In particular, while Harvey et al (1994) discuss how to estimate small scale multivariate stochastic volatility model using state space methods, Alizadeh et al (2002) cover only the univariate model. Therefore, we first show how to extend the model by Alizadeh et al (2002) to the multivariate case and then apply our approximate maximum likelihood estimation to a large portfolio of stocks.

Let y_{it} be the, possibly mean-adjusted, return of stock $i \in \{1, 2, \dots, d\}$ at time $t \in \{1, 2, \dots, n\}$, then the multivariate stochastic volatility model Harvey et al (1994) consider is defined by

$$\begin{aligned} y_{it} &= \exp(h_{it}/2)\zeta_{it}, \\ h_{it+1} &= h_{it} + \eta_{it}, \end{aligned} \tag{12}$$

where $\exp(h_{it})$ plays the role of the time-varying variance, whose logarithm evolves according to a random walk. The random vectors $\boldsymbol{\zeta}_t$ and $\boldsymbol{\eta}_t$ obtained by stacking the random variables $\{\zeta_{1t}, \dots, \zeta_{dt}\}$ and $\{\eta_{1t}, \dots, \eta_{dt}\}$, respectively, into vectors are assumed to be normally distributed with zero means and covariance matrices $\boldsymbol{\Sigma}_\zeta$ and $\boldsymbol{\Sigma}_\eta$. The matrix $\boldsymbol{\Sigma}_\zeta$ is constrained to be a correlation matrix. Harvey et al (1994) model the logarithm of the squared return, so that equation (12) can be rewritten as

$$\begin{aligned} \log(y_{it}^2) &= h_{it} + \log(\zeta_{it}^2), \\ h_{it+1} &= h_{it} + \eta_{it}, \end{aligned} \tag{13}$$

This set of equations can be easily cast into the linear state space form as

$$\begin{aligned} \boldsymbol{w}_t &= \boldsymbol{h}_t + \boldsymbol{\varepsilon}, \\ \boldsymbol{h}_{t+1} &= \boldsymbol{h}_t + \boldsymbol{\eta}_{it}, \end{aligned} \tag{14}$$

where the i-th element of $\boldsymbol{w}_t$ is equal to $\log(y_{it}^2) + 1.27$ and $\boldsymbol{\varepsilon}_t$ is a non-Gaussian random i.i.d. sequence whose ij-th element of the covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ is given by equation B.9 in Harvey et al (1994).

Harvey et al (1994) propose to estimate the unknown covariance matrices $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\eta$ by Gaussian quasi maximum likelihood, approximating the distribution of $\boldsymbol{\varepsilon}_t$ with a normal with the same mean and covariance matrix. The log-variances h_{it} can be estimated using Kalman filtering and state smoothing, which in this case provide just best linear estimates.

Alizadeh et al (2002) propose a stochastic volatility model based on the logarithm of stock price ranges (daily maximum minus daily minimum) instead of log-squared returns. Let $v_{it} = \log(P_{it}^{\max} - P_{it}^{\min})$ be the sequence of daily log-ranges, with $P_{it}^{\max}$ and $P_{it}^{\min}$ representing the maximum and minimum price of stock i for the day t. If each price process is well described by a Brownian motion, than the distribution of the log-range is approximately Gaussian (cf. Alizadeh et al, 2002, Table I and Figure 1). Assuming that the approximation remains valid also in a multivariate context, then we can write a multivariate stochastic volatility model based on log-ranges exactly as in equation (14), where now the generic element of $\boldsymbol{w}_t$ is $w_{it} = v_{it} - 0.43$, and the errors ε_{it} have all the same standard deviation $\sigma_\varepsilon = 0.29$ (cf Alizadeh et al, 2002, Table I).

We estimated both SV models for a portfolio composed by 94 daily stock returns belonging to the SP100 index ranging from 2007-01-04 to 2017-04-28 ($n = 2598$). As customary in many large scale multivariate GARCH models, we split the estimation process in two steps: first, we estimated the correlation matrix of the returns, from which we obtained the covariance matrix $\boldsymbol{\Sigma}_\varepsilon$, and then we applied our Algorithm 1 by running the EM update only for the matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, as in equation (7), and using the $\boldsymbol{\Sigma}_\varepsilon$ estimated in the first step.

If we concentrate on the correlation matrix of the disturbances that drive the h_{it} processes, we notice that all correlations are positive, but those of model HRS are generally larger than those of model ABD. The same message can be derived from the cumulated eigenvalues plot in the left panel of Figure 1: for model HRS the first three principal components cover more than 90% of the variance, while the same share of total variance is reached for model ABD with 20 components. If we use the scores of the first principal component of the estimated $\boldsymbol{h}_t$ in the two models and take the suitable transforms to derive volatility indicators, we get the volatility profiles depicted in the right panel of Figure 1. The profiles are similar, but they differ in some extreme event and in the level, which is higher for the HRS model.

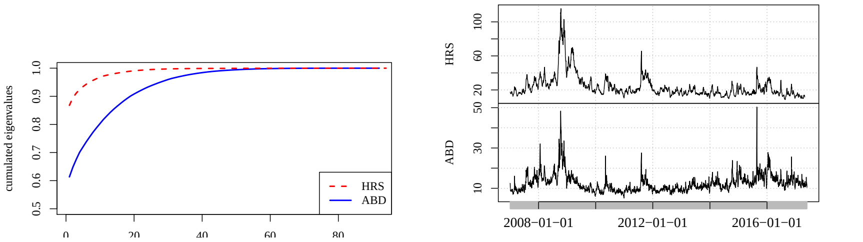


Fig. 1 Left) cumulated eigenvalues of the correlation matrix derived from $\boldsymbol{\Sigma}_\eta$ estimated in the two models. Right) Ensemble volatility indicators derived from the first principal component scores

References

Alizadeh S, Brandt MW, Diebold FX (2002) Range-based estimation of stochastic volatility models. The Journal of Finance 57(3):1047–1091

Ansley C, Kohn R (1985) Estimation, filtering and smoothing in state space models with incomplitely specified initial conditions. The Annals of Statistics 13:1286–1316

Durbin J, Koopman S (2001) Time Series Analysis by State Space Methods. Oxford University Press

Harvey A (1989) Forecasting Structural Time Series and the Kalman Filter. Cambridge University Press

Harvey A, Ruiz E, Shephard N (1994) Multivariate stochastic variance models. Review of Economic Studies 61(2):247–264

de Jong P (1988) A cross-validation filter for time series models. Biometrika 76:594–600

de Jong P (1989) Smoothing and interpolation with state-space models. Journal of the American Statistical Association 75:594–600

Koopman S (1993) Disturbance smoother for state space model. Biometrika 80(1):117–126

Koopman S (1997) Exact initial kalman filtering and smoothing for nonstationary time series models. Journal of the American Statistical Association 92(400):1630–1638

Shumway RH, Stoffer DS (2017) Time Series Analysis and Its Applications. With R Examples. Springer

Health and Clinical Data

Is retirement bad for health? A matching approach

Il pensionamento fa male alla salute? Una analisi causale

Elena Pirani, Marina Ballerini, Alessandra Mattei, Gustavo De Santis

Abstract The aim of this paper is to assess the causal impact of the transition from work to retirement on individual health in various European countries in recent years. The health effects of this transition are far from clear: the specialized literature reports both positive and negative consequences, however, most of the early studies focus on associations rather than causal relationships. We estimate causal effects of retirement on three measures of health and well-being – self-rated health, depression, quality of life – using a propensity score matching approach under the assumption of selection on observables on data coming from SHARE, the longitudinal Survey on Health, Ageing and Retirement in Europe, in the years between 2004 and 2016. Our results suggest that the transition from work to retirement negatively affect self-rated health almost everywhere in Europe; nevertheless, the quality of life seems to improve, especially in Continental and Mediterranean countries.

Abstract L'obiettivo di questo lavoro è valutare l'impatto causale del pensionamento sulla salute individuale in vari paesi europei. La letteratura riporta effetti sia positivi che negativi del pensionamento sulla salute e benessere degli individui, tuttavia, la maggior parte degli studi precedenti si concentra sulle associazioni piuttosto che sulle relazioni causali. Considerando tre misure di salute – salute percepita, depressione e qualità della vita – sulla base dei dati delle indagini SHARE svolte tra il 2004 e il 2016, stimiamo gli effetti causali del pensionamento sulla salute utilizzando l'approccio del propensity score matching sotto l'ipotesi di assenza di confondimento. I risultati suggeriscono che il pensionamento ha effetti negativi sulla salute percepita in tutti i paesi europei, mentre la qualità della vita sembra beneficiarne, soprattutto nei paesi dell'Europa continentale e mediterranea.

Key words: Retirement; Europe; Share; Causal inference.

[1] Elena Pirani, University of Florence, elena.pirani@unifi.it;
Marina Ballerini, University of Florence, marinaballerini.21@gmail.com;
Alessandra Mattei, University of Florence, mattei@disia.unifi.it;
Gustavo De Santis, University of Florence, desantis@disia.unifi.it.

1 Retirement and health: a complex connection

The effects on health of the transition from work to retirement are unclear. Several scholars argue that retirement itself is a stressful event (e.g., Carp 1967; MacBride 1976), which can lead to a break with support networks and friends, and may be accompanied by feelings of loneliness, uselessness, or obsolesce (MacBride 1976). Others claim instead that retirement is a health-preserving life event: it is a relief from work-related stress (Eibich 2015), and encourages health-improving behaviors – such as quit smoking – or increased physical activity (Eibich 2015; Insler 2014).

A strand of the literature reports a significant increase in health after retirement (e.g. Blake and Garrouste 2012; Charles 2004; Coe and Zamarro 2011; Insler 2014; Latif 2013; Neuman 2008), whereas other researchers find significant negative effects on both objective and subjective health measures (e.g. Behncke 2012; Dave et al. 2008; Sahlgren 2012), and also on cognitive functions (Bonsang et al. 2012; Mazzonna and Peracchi 2012). Bound and Waidmann (2007) showed a short-term positive relationship between retirement and health for men but not for women.

In a large part of the previous studies, the focus was on the *association* between health and retirement, and comparisons between the retired and those still working were usually not adjusted for health characteristics before retirement: this adjustment is instead crucial for drawing inference on the causal effects of retirement on health (Coe and Zamarro 2011).

In this paper, we aim to assess the *causal* impact on health of the transition from work to retirement by applying a propensity score matching approach under the assumption of selection on observable to the data of the Survey on Health, Ageing and Retirement in Europe (SHARE). We refer to several European countries, and analyse the heterogeneity of the causal effects across different geographical areas.

2 Data and method

Our empirical analyses were based on SHARE, the Survey on Health, Ageing and Retirement in Europe, which is a panel including five regular waves plus a wave on people’s life histories (wave 3, SHARELIFE, which, however, is not considered here because it collected very different information compared to the regular waves and excluded some of the variables we need). In each wave, SHARE data cover the key areas of life (health, socio-economic status social and family networks, etc.) of more than 60,000 individuals aged 50 or over. We focused on the period between 2004 and 2016 and on a subset of the SHARE countries that participated in at least three consecutive waves (Austria, Belgium, Czech Republic, Denmark, Estonia, France, Germany, Italy, the Netherland, Slovenia, Sweden, Switzerland, and Spain), and for which we had all the information we needed for our analysis.

We selected three health and well-being indicators. As a general measure of health, we used self-rated health, dichotomizing the original 5-point scale into good (excellent, very good, good) and poor (fair, poor) perceived health (preliminary analyses considering the original formulation proved consistent results). Because of its subjective nature, self-rated health may change across populations (Prinja et al., 2012); however, various studies proved its power in predicting objective health conditions (Egidi and Spizzichino, 2006), physical and emotional well-being (Bayliss et al., 2012), and even mortality (Idler and Benyamini, 1997). Moreover, we used a composite indicator of depression constructed from the 12 basic items of the EURO-D scale (Prince et al., 1999): depressed mood, pessimism, suicidality, guilt, sleep, interest, irritability, appetite, fatigue, concentration, enjoyment and tearfulness. This scale, which ranges from 0 (not depressed) to 12 (very depressed), was developed in an effort to derive a common scale of depression symptoms, especially in later life, based on different indicators in several European countries. Finally, we considered a theoretically grounded measure of quality of life, i.e., a composite indicator based on four subscales corresponding to four life domains (Hyde et al., 2003; Mehrbrodt et al., 2017): control (C), autonomy (A), self-realization (S) and pleasure (P). After reversing the original scale, this indicator, which was proved to represent a reliable indicator of quality of life in the context of research on ageing (Wiggins et al., 2008), scored between 0 (high quality of life) and 36 (low quality of life).

Because our objective was the estimation of a causal effect of entering retirement on individual well-being in the short run, we focused on the subsample of those who were in the labor market (employed or self-employed) in the waves $t=1,2,4$ of the survey and were either in the labor market or retired in the waves $t=2,4,5$. Those who got out of the labor market for other reasons (e.g., unemployment) were discarded from our subsample.

We followed the “potential outcome” approach (e.g., Imbens and Rubin, 2015). For each unit i, $i=1,\ldots, n$, we considered a vector X_i of background variables. Let D_i denote the treatment variable indicator, equal to 1 if unit i retires between two consecutive waves of the survey – waves $w1$ and $w2$; waves $w2$ and $w4$, or waves $w4$ and $w5$ – and zero otherwise. Under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980), for each unit i there are two potential outcomes at a future point in time after treatment (at waves $w4$, $w5$ or $w6$): the value of the health outcome Y if unit i retired – $Y_i(1)$ – and the value of Y at the same future point in time if the unit did not retire – $Y_i(0)$. The causal effect of the transition from work to retirement for each unit is defined as a comparison of the treatment and control potential outcomes, $Y_i(1)$ and $Y_i(0)$, typically their difference. In this paper we focus on Average Treatment Effects on the Treated (ATT effects), that is, the effects of the transition from work to retirement averaged over the subpopulation of units who actually retired: ATT = $\mathrm{E}[Y_i(1) - Y_i(0) \mid D=1]$.

Because we used observational data, we needed to introduce some assumptions on the treatment-assignment mechanism to draw inference on the causal effects of interest. We assumed unconfoundedness (or selection on observables), which implies that, conditioning on the observed covariates, an experimental-like context is reproduced. Formally, unconfoundedness requires that the treatment assignment is independent of the potential outcomes: $D_i \perp (Y_i(0), Y_i(1))|\mathrm{X}_i$. We also assumed that there was sufficient overlap in the joint distribution of the covariates between treated and control subjects: $0 < P(D_\mathrm{i}=1|X_i=x) < 1$ for each *i*. Under these assumptions, we applied a statistical matching technique, the purpose of which was to select a sub-group of control subjects (who did not retire between two subsequent waves) who were, in all respects, as similar as possible to the treated subjects, i.e., those who retired from work. We matched individuals based on the propensity score, or the probability of entering retirement conditional on the observed covariates. The propensity score is a balancing score, that is, covariates are independent of the treatment conditional on the propensity score. Moreover, if the unconfoundeness and the overlap assumptions hold conditional on covariates, they also hold conditional on the propensity score (Rosenbaum and Rubin 1983). Therefore, matching on the propensity score is sufficient to remove confounding.

The variables on which we constructed our propensity score included individual socio-demographic characteristics (e.g., age, living arrangement, relatives alive, level of education) and health-related behaviours and health conditions (e.g., smoking and drinking, mobility index, as well as self-rated health, quality of life, and depression). We also introduced the type and the sector of work, even if we acknowledge that these aspects only partially account for important aspects of working life, such as stressing factors or autonomy in decisions, which could importantly contribute to individual’s health and wellbeing. We imposed an exact matching on country of residence and gender. On the basis of the estimated propensity score, for each of the 1124 retired individuals we selected as a match the closest individual – i.e., a person of the same sex, from the same country and with very similar pre-treatment characteristics – among the 6250 potential controls (1-to-1 nearest neighbor matching; Abadie and Imbens 2011).

3 Preliminary results

The check of the covariate balance (i.e., similarity of treated and controls individuals in terms of covariates) proved that the matching procedure was successful: after matching, the differences between the two groups (retired an non-retired) in terms of socio-demographic, work-related and health covariates either disappeared (best case) or were drastically reduced (not shown here). We proceeded then to the estimation of the causal effect of retirement on health, by computing the Average Treatment Effects on the Treated (ATT) using a matching estimator.

The first part of Table 1 reports the ATT effects of retirement on the three well-being indicators considered, computed for all the SHARE countries together. Our results show a worsening of self-rated health after retirement; conversely, they also convey the impression of a slightly improvement in terms of (less) depression and (higher) quality of life. Note however that these two latter estimated effects are very small in absolute terms and not statistically significant.

In order to account for the heterogeneity of people living in the different European countries, we estimated the ATT effects separately for the Nordic countries (Denmark and Sweden), Continental countries (Austria, Belgium, France, Germany, the Netherland, and Switzerland), Mediterranean countries (Spain and Italy), and East European countries (Czech Republic, Slovenia, Estonia). The negative effect of retirement on self-rated health persisted for all European areas, even if with a loss of significance for East European countries.

Considering European countries altogether masks some territorial differences for the other two well-being indicators: both depression and quality of life levels increase for retired people in Continental and Mediterranean countries (although the effect is not statistically significant for the second group, maybe due to the small sample size). On the contrary, in Nordic countries retirement seems to determine a detrimental effect also on depression and quality of life (but again results are not statistically significant), whereas for Eastern European countries the effect on the two well-being variables diverges.

Table 1: Estimated ATT effects and their standard errors, for all countries, and by groups of countries. SHARE 2004-2016.

			Self-rated health	Depression	Quality of life
All countries	treated (n=1124),	ATT	0.09	-0.12	-0.17
	matched controls (n=1124)	*Std Err*	*0.01*	*0.07*	*0.21*
By group of countries:					
Nordic	treated (n=194),	ATT	0.09	0.13	0.29
	matched controls (n=194)	*Std Err*	*0.03*	*0.15*	*0.44*
Continental	treated (n=474),	ATT	0.08	-0.10	-0.91
	matched controls (n=474)	*Std Err*	*0.02*	*0.11*	*0.31*
Mediterranean	treated (n=116),	ATT	0.12	-0.14	-0.13
	matched controls (n=116)	*Std Err*	*0.05*	*0.22*	*0.68*
East European	treated (n=340),	ATT	0.04	-0.24	0.68
	matched controls (n=340)	*Std Err*	*0.03*	*0.15*	*0.44*

In our analysis the richness of background (i.e., pre-treatment) information allowed us to adjust treatment comparisons for a large set of pre-treatment characteristics – in terms of health, life-style behaviors, socio-demographic characteristics and factors linked to the (previous) working condition – and thus the assumption of selection on observable appears to be plausible. Under this assumption, we found that the effects of retirement on health vary not only according to the context of reference, but also depending on the specific health/well-being indicator considered.

It is thus worth to investigate more in detail the mechanisms through which retirement affects the various dimensions of health and well-being. Specifically, aspects such as family types and intergenerational relationships, social relationships, embeddedness in social network, and job characteristics will be examined in our future research.

Acknowledgements: We thank the CREW group for early feedback on this paper (CREW=Care, Retirement and Wellbeing of Older People across Different Welfare Regimes; https://crew-more-years-better-lives.org/).

References

1. Abadie A, Imbens GW. (2011). Bias-corrected matching estimators for average treatment effects. *J Bus Econ Stat.*; 29(1), 1–11.
2. Bayliss E.A., Ellis J.L., Shoup J.A., Zeng C., McQuillan D.B., Steiner, J.F. (2012). Association of patient-centered outcomes with patient-reported and icd-9-based morbidity measures. *Ann. Fam. Med.*, 10, 126-133.
3. Behncke S. (2012). Does retirement trigger ill health?, *Health Economics*, 21, 282-300
4. Blake H., and Garrouste C. (2012). Collateral effects of a pension reform in France, *Health, Econometrics and Data Group (HEDG) Working Papers* 12/16, University of York
5. Bonsang E., Adam S., and Perelman S. (2012), Does retirement affect cognitive functioning?, *Journal of Health Economics*, 31, 490– 501
6. Bound J., and Waidmann T. (2007). Estimating the Health Effects of Retirement, University of Michigan *Retirement Research Center working paper* 2007-168
7. Carp F.M. (1967). Retirement crisis, *Science* 157, 102–103
8. Charles K. (2004). Is retirement depressing? Labor force inactivity and psychological well-being in later life, *Research in Labor Economics*, 23, 269-299
9. Coe N.B., and Zamarro G. (2011), Retirement effects on health in Europe, *Journal of Health Economics*, 30, 77-86
10. Dave D., Rashad I. and Spasojevic J. (2008). The Effects of Retirement on Physical and Mental Health Outcomes, *Southern Economic Journal*, 75, 2, 497-523
11. Egidi V., Spizzichino D. (2006). Perceived health and mortality: a multidimensional analysis of ECHP Italian Data. Genus LXII (3e4), 135e154
12. Eibich P. (2015). Understanding the effect of retirement on health: Mechanisms and Heterogeneity, *J. Health Econ*., 43, 1–12
13. Hyde M., Wiggins R.D., Higgs P., & Blane D.B. (2003). A measure of quality of life in early old age: the theory, development and properties of a needs satisfaction model (CASP-19). *Aging & mental health*, 7(3), 186-194.
14. Idler L.E., Benyamini Y. (1997). Self- rated health and mortality: a review of 27 community studies. *J. Health Soc. Behav.*, 38, 21-37.
15. Imbens GW, Rubin DB (2015). *Causal inference for statistics, social, and biomedical sciences: an introduction*. 1st ed. New York: Cambridge University Press.
16. Insler M. (2014). The Health Consequences of Retirement, *The Journal of Human Resources*, 49, 1
17. Latif E. (2013). The impact of retirement on mental health in Canada, *Journal of Mental Health Policy and Economics*, 16(1), 35-46.
18. MacBride A. (1976). Retirement as a life crisis: myth or reality? *Canadian Psychiatric Association Journal* 72, 547–556
19. Mazzonna F. and Peracchi F. (2014). Unhealthy Retirement?, *EIEF Working Paper 14/09*
20. Mehrbrodt T., Gruber S. & Wagner S. (2017). Scales and Multi-Item Indicators. Share Manuals, retrieved on www.share-project.org/fileadmin/pdf_documentation/SHARE_Scales_and_Multi-Item_Indicators.pdf
21. Neuman K. (2008). Quit Your Job and Get Healthier? The Effect of Retirement on Health, *Journal of Labor Research*, 29, 177-201
22. Prince M.J., Reischies F., Beekman A.T., et al.(1999). Development of the EURO-D scale-a European, Union initiative to compare symptoms of depression in 14 European centres. *British Journal of Psychiatry*,174, 330–8.
23. Rosenbaum P.R, Rubin D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
24. Rubin DB. (1980). Discussion of “Randomization analysis of experimental data in the Fisher randomization test”, by Basu, *Journal of the American Statistical Association*, 75, 591–593.
25. Sahlgren G. (2012), Work ‘til you drop: Short- and longer-term health effects of retirement in Europe, *IFN Working Paper* n. 928
26. Wiggins R.D., Netuveli G., Hyde M., Higgs P., Blane D. (2008). The Evaluation of a Self-enumerated Scale of Quality of Life (CASP-19) in the Context of Research on Ageing: A Combination of Exploratory and Confirmatory Approaches. *Social Indicators Research*, 89, 61–77.

The emergency department utilisation among the immigrant population resident in Rome from 2005 to 2015

L'accesso al pronto soccorso da parte degli immigrati residenti a Roma tra il 2005 e il 2015

Eleonora Trappolini, Laura Cacciani, Claudia Marino, Cristina Giudici, Nera Agabiti, Marina Davoli

Abstract

Health inequalities affecting migrant population have been largely analysed in the literature, but only scarce studies have been published on migrant's utilisation of Emergency Departments (EDs). The aim of this study is to analyse trends of EDs utilisation for the population resident in Rome during the period 2005-2015, overall and for specific causes. By comparing immigrants' access with that of the host population, the study evaluates differences in the healthcare utilisation rate between these populations. The analysis is based on a dynamic cohort aged 25 to 64 years old in each year and defined using data from the Municipal Register of Rome. Results on the overall utilisation rate show a lower use of the EDs by the immigrant population with respect to the host population, although this gap has started to decrease in the years after the 2008. Even if the overall utilisation rates show similar trends between Italians and immigrants, different patterns have been detected for specific causes.

Abstract

Le disuguaglianze di salute legate alla condizione migratoria sono state oggetto di numerose analisi in letteratura. Tuttavia, nel panorama europeo, gli studi sull'accesso degli immigrati al servizio di pronto soccorso risultano ad oggi scarsi e spesso contraddittori. Questo contributo è volto ad analizzare l'accesso al pronto

[1] Eleonora Trappolini, Sapienza University of Rome; email: eleonora.trappolini@uniroma1.it
Cristina Giudici, Sapienza University of Rome; email: cristina.giudici@uniroma1.it
Dipartimento di Epidemiologia del Servizio Sanitario Regionale del Lazio – ASL Roma 1:
Laura Cacciani: l.cacciani@deplazio.it
Claudia Marino: c.marino@deplazio.it
Nerina Agabiti: n.agabiti@deplazio.it
Marina Davoli: m.davoli@deplazio.it

soccorso da parte della popolazione italiana e straniera residente a Roma tra il 2005 e il 2015.
L'analisi si basa su una coorte dinamica di età compresa, ogni anno, tra 25 e 64 anni, definita utilizzando i dati dell'Anagrafe Comunale di Roma. I risultati mostrano un minore accesso all'emergenza per gli stranieri rispetto agli Italiani, anche se il gap mostra una riduzione dopo il 2008. In generale le due popolazioni mostrano comportamenti simili, mentre emergono differenze dal confronto per specifiche cause di accesso.

Key words: Emergency Department utilisation, Immigrant population, Dynamic cohort, Time trends, Rome

1 Introduction

During the last decades, the number of immigrants in Italy has continued to increase, reaching 5,047,028 (8.3%) in 2016 (Istat, 2016). On this basis, migration is no longer a transient phenomenon as in the 1970s, but it should be considered as a structural component of our society.

The WHO Commission on Social Determinants of Health recently highlighted the rise of new health inequalities between and within countries due to differences in social class, gender and ethnicity, as far as inequalities in the access to healthcare services among migrants and natives (Malmusi et al., 2010). Access to healthcare should be seen as no less important than housing and education for the wellbeing. Actually, it can be considered a proxy of the integration process (Ingleby et al., 2008).

At present, only scarce and often contradictory data have been published on migrant's utilisation of healthcare services in Europe (Norredam et al., 2004; Rué et al., 2008), and even less information is available on migrant's utilisation of Emergency Departments (EDs). Although the literature shows different scenarios (Zinelli et al., 2014; De Luca et al., 2013), there is a general consensus about the immigrants' higher use of emergency room services than that of non-migrants (Cots et al., 2007; Ruud et al., 2015). In Italy, a retrospective analysis based on 2005 Italian Health Conditions Survey, carried out by the Italian National Statistical Office, registered a higher Emergency Department utilisation rate by immigrants, notably from Morocco, other African countries and Albania (De Luca et al., 2013). These differences may be partly explained first of all by obstacles in the access to primary care, which are related to the migrant status, but also by different factors mainly related to the perceptions of illness or to the so called *health literacy.*

In Italy, some recent studies investigated the health of immigrants (Pacelli et al., 2016; Cacciani et al., 2011); however, as far as we know, no evidence is available about the health service utilisation of Italians and immigrants during the recent Great Recession.

While immigrant flows have increased rapidly in the last decade, we hypothesize that under the conditions to which immigrants are exposed because of their status, they might not use the medical system in the same way as the Italian population.

The objective of this study is twofold. Firstly, it investigates the possible differences in the EDs utilisation between immigrant and Italian residents in Rome. Secondly, it compares the EDs utilisation between the periods before and after the 2008.

2 Materials and methods

An observational study based on a dynamic cohort, defined using data from the Municipal Register of Rome, has been performed to evaluate the healthcare utilisation rate of Italians and that of the immigrant population, during the last decade.

The study population includes all residents in Rome, aged 25 to 64 years old between 1st January 2005 and 31st December 2015 in each year (2,184,467 individuals). Emergency Department data, gathered from the Regional Health Information System on Emergency Care in Lazio, were linked to the subjects using an individual anonymised code. An average of 390,000 visits by year has been observed. The available data include information about gender, birthdate, birthplace, citizenship, as well as medical information. Diagnoses and procedures are coded according to the ICD-9-CM.

The outcome variable is the overall number of EDs contacts. The immigrant status is the exposure variable (immigrants vs Italians). The most appropriate information to identify immigrants is still being debated (Malmusi et al., 2010). This study defines as immigrants those individuals without the Italian citizenship, distinguishing them between immigrants coming from High Migratory Pressure Countries (HMPCs)[1] and coming from Highly Developed Countries (HDCs). Since we are dealing with people resident in Rome, other migrant categories have been excluded. In order to explore differences in the emergency access, we performed descriptive analyses which aimed to investigate the pattern of EDs use by immigrants living in Rome (overall and for specific causes[2]), comparing it with that of the resident Italian population. Crude utilisation rates (URs), using person years, and direct age-standardized URs, by gender and for both immigrants and Italians, have been computed using the Italian population residing in Lazio at 1st January 2014 as standard population. The following age groups have been considered: 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64.

Negative binomial regression was used to estimate rate ratios (RRs) with 95% confidence intervals in order to evaluate differences in the EDs utilisation between

[1] Central-Eastern Europe (including Poland and Romania), North Africa, Sub-Saharan Africa, Asia (except for Israel and Japan), and Central and South America

[2] Mental disorders (ICD-9-CM 291-312), injuries (ICD-9-CM 800-959), and cardiovascular diseases (ICD-9-CM 401-445).

immigrants and Italians, adjusting for gender, age, socio-economic position (SEP)[1], and stratifying by time period.

3 Results

The total number of ED contacts during the period 2005-2015 was 4,297,981.

Immigrants as a whole had a lower UR than Italians, and this trend was similar in all age groups, except for the youngest women aged 25-29 (39.9 x 100 py compared to 37.5 x100 py for Italians). Women had higher overall URs than men in all groups of geographical origin (Fig. 1.a).

Of the two immigrant groups examined, those coming from HDCs registered the lowest rates. As figures show, even if the highest rate belongs to Italians, a similar trend between Italians and immigrants coming from HMPCs has been detected. Although the overall UR, both for men and women, varied slightly over the study period (Fig. 1.a), the access for specific causes showed different patterns. For the specific causes analysed, men always registered higher URs compared to women, except for both the causes related to mental disorders, where we observed a higher but constant UR among women coming from HMPCs, and injuries where women coming from HDCs registered a higher UR (Fig. 1.b). Concerning contacts for injuries we observed decreasing patterns especially among Italian men (Fig. 1.c). While the UR for mental disorders was fairly constant and that for injuries was declining, the contacts for cardiovascular-related causes registered an increasing pattern among all groups, especially for immigrants. In addition, immigrant men registered a higher UR with respect to Italian women (Fig. 1.d).

Figure 1. Overall and cause-specific (mental disorders, injuries, and CVDs) age-standardized UR of residents in Rome, by gender and area of origin. Lazio, years 2005-2015

Figure 1a. Overall

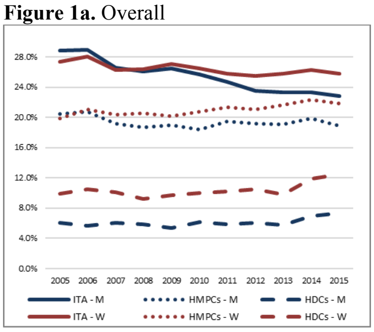


Figure 1b. Mental disorder

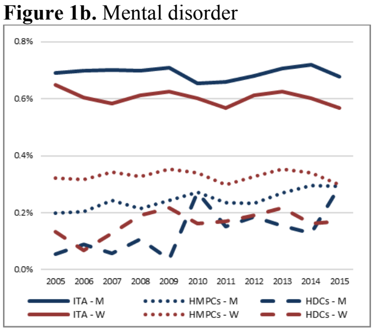
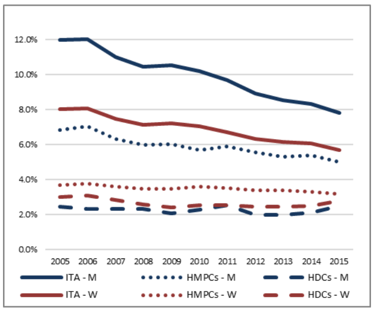


Figure 1c. Injuries

Figure 1d. Cardiovascular diseases

[1] The SEP indicator (from 1 higher to 4 lower) is based on the characteristics of the census block of residence (that is the smallest territorial unit for which population data were available).



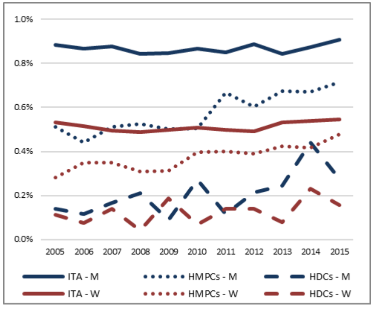


Table 1 shows the results from the Negative Binomial regression analysis. Italians were used as the reference group when comparing area of origin, men were used as a reference group when comparing gender and those with a high socio-economic position were used as a reference group when comparing the SEP. Age was considered as a continuous variable. An overall highly significant association was found between EDs utilisation rates and area of origin (p<0.0001). When controlling for age, gender and socio-economic position, immigrants registered a lower UR compared with Italian residents. A highly significant association was also found between the socio-economic position and utilisation. Table 1 shows how URs increase with decreasing SEP. Moreover, women registered higher URs with respect to men and the URs decrease with age. Comparing the period before and after the 2008, results show that patient behaviours are fairly constant in the two periods, however during the years analysed a reduction in the gap between the Italian population and the immigrant population (HMPCs) has been observed.

Table 1. Negative Binomial regression: rate ratios (with 95% confidence intervals) of ED utilisation for immigrant compared to Italian resident population in Rome, adjusted for age, gender, and socio-economic position, and stratified by time period. Lazio, years 2005-2015.

	Pre-2008		Post-2008	
	RRs	95% CI	RRs	95% CI
Area of origin:				
Italy	1		1	
HMPCs	0.743	(0.736 - 0.750)	0.800	(0.787 - 0.811)
HDCs	0.335	(0.324 - 0.345)	0.360	(0.351 - 0.369)
Gender and Age:				
Man	1		1	
Woman	1.013	(1.000 - 1.010)	1.088	(1.083 - 1.093)
25-64	0.983	(0.982 - 0.984)	0.986	(0.986 - 0.987)
Socio-ec. position:				
Higher	1		1	
Middle	1.099	(1.091 - 1.106)	1.104	(1.097 - 1.111)
Fairly	1.230	(1.221 - 1.239)	1.240	(1.232 - 1.248)
Lower	1.508	(1.496 - 1.519)	1.508	(1.498 - 1.519)

Our results show lower utilisation of the ED (overall) among immigrants during the period 2005-2015, in Rome. This result is consistent with previous reports of healthcare utilisation by the immigrant population and was probably due to the *healthy immigrant effect* (Carrasco-Garrito et al., 2007; Norredam et al., 2007) where recently arrived immigrants have better health status than natives because of a previous *selection process* in each country of origin. Consideration should be also given to the fact that we are dealing with immigrants who are resident in Rome, it means they have a residence permit and most of them have a work, thus they are also positively selected compared to other migrant categories. Moreover, other studies have shown no relationship between the perceptions of need, willingness to seek health services and ethnicity (Adamson et al., 2003), and more evidence is provided towards the hypothesis that barriers occur at the access of the services (Okie et al., 2007).

However, it would be interesting to analyse the access for specific causes, in which we detected different patterns, and to explore the hospitalization. Furthermore, even if the relationship between the SEP and the ED utilisation was expected, in this context the SEP was used for adjusting the exposure, but because the mechanism behind the interaction of this two variables is complex, further insights are necessary. Concerning comparative analyses, we will also should consider other factors which may have contributed to the modification in the EDs utilisation, as the Great Recession and the austerity policies implemented by the region.

References

1. Adamson J, Ben Shlomo Y, Chaturvedi N, Donovan J: Ethnicity, socio-economic position and gender-do they affect reported health-care seeking behaviour? pp. 895-904, Soc Sci Med (2003)
2. Cacciani L., Asole S., Polo A., Franco F., Lucchini R., De Curtis M., et al., Perinatal outcomes among immigrant mothers over two periods in a region of central Italy, p. 294. BMC Public Health (2011)
3. Carrasco-Garrido P, De Miguel AG, Barrera VH, Jimenez-Garcia R., Health profiles, lifestyles and use of health resources by the immigrant population resident in Spain, pp. 503-507, Eur J Public Health (2007)
4. Cots F., Castells X., Ollé C., Manzanera R., Varela J., Vall O., Perfil de la casuística hospitalaria de la población inmigrante de Barcelona, pp. 376-384. Gaceta Sanitaria (2007)
5. De Luca G., Ponzo M., Andres A.R., Health care utilization by immigrants in Italy, pp. 1–31. International Journal Health Care Finance Econ (2013)
6. Ingleby D., New perspectives on migration, ethnicity and schizophrenia, Willy Brandt Series of Working Papers in International Migration and Ethnic Relations 1/08, Malmö University, Malmö (2008)
7. Istat (2016), Bilancio demografico nazionale: http://www.istat.it/it/files/2017/06/bilanciodemografico201613giugno2017.pdf?title=Bilancio+demografico+nazionale++13%2Fgiu%2F2017+-+Testo+integrale.pdf
8. Malmusi D., Borrell C., Benach J., Migration-related health inequalities: showing the complex interactions between gender, social class and place of origin, pp.1610-1619. Social Science & Medicine (2010)
9. Norredam M., Krasnik A., Sorensen T.M., Keiding N. et al., Emergency room utilization in Copenhagen: a comparison of immigrant groups and Danish-born residents, pp. 53-59. Scandinavian Journal Public Health (2004)
10. Norredam M, Mygind A, Nielsen AS, Bagger J, Krasnik A: Motivation and relevance of emergency room visits among immigrants and patients of Danish origin, pp. 497-502, Eur J Public Health (2007)

11. Okie S: Immigrants and health care--at the intersection of two broken systems, pp.525-529, N Engl J Med (2007)
12. Pacelli B., Zengarini N., Broccoli S., Caranci N., Spadea T., Di Girolamo C., et al., Differences in mortality by immigrant status in Italy. Results of the Italian Network of Longitudinal Metropolitan Studies, pp. 691–701. Eur J Epidemiol (2016)
13. Rué M., Cabré X., Soler-Gonzalez J., Bosch A., Almirall M., Catalina Serna M. Emergency hospital service utilization in Lleida, pp. 1-8. BMC Health Service Research (2008)
14. Ruud S.E., Aga R., Natvig B., Hjortdahl P., Use of emergency care services by immigrants-a survey of walk-in patients who attended the Oslo Accident and Emergency Outpatient Clinic, pp. 15-25. BMC Emerg Med (2011)
15. Zinelli M., Musetti V., Comelli I., et al., Emergency department utilization rates and modalities among immigrant population. A 5- year survey in a large Italian urban emergency department, pp.22–25. Emerg Care Journal (2014)

Multi-State model with nonparametric discrete frailty

Modelli multi-stato con termine di frailty discreto nonparametrico

Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson and Linda Sharples

Abstract In this work, we propose a novel semi-Markov multi-state model with a nonparametric discrete frailty and an application to an administrative clinical database about heart failure patients from a Northern Region of Italy. In particular, we investigate a illness-death model with recovery in which the states space is composed by hospital admission, hospital discharge and death, as unique absorbing state. The available data are grouped longitudinal time-to-event data, indeed for each patient we know the times of admission and discharge of all hospitalizations (2005-2012), the time of death (if it occurs) and the healthcare provider (grouping factor). Thanks to this model, we can investigate the effect of covariates, detect the presence and a pattern of latent populations of healthcare providers across transitions.

Abstract *In questo lavoro proponiamo un modello multi-stato di tipo semi-Markov con una termine random discreto noparametrico e un'applicazione ad un database amministrativo clinico riguardante pazienti affetti da scompenso cardiaco in una regione del Nord Italia. In particolare, ci concentriamo su un modello di tipo illness-death con guarigione, in cui lo spazio degli stati è composto da ammissione in ospedale, dimissione dall'ospedale e morte, unico stato assorbente. I dati che abbi-*

Francesca Gasperoni
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: francesca.gasperoni@polimi.it

Francesca Ieva
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: francesca.ieva@polimi.it

Anna Maria Paganoni
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: anna.paganoni@polimi.it

Chris Jackson
MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge (UK), CB2 0SR, e-mail: chris.jackson@mrc-bsu.cam.ac.uk

Linda Sharples Department of Medical Statistics, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, e-mail: Linda.Sharples@lshtm.ac.uk

amo a disposizione sono dati longitudinali di tipo tempo all'evento e raggruppati, infatti per ogni paziente siamo a conoscenza dei tempi di ammissione, dimissione (nell'arco di tempo 2005-2012) e momento del decesso, se avvenuto, e la struttura ospedaliera (fattore di raggruppamento). Tramite questo modello possiamo investigare l'effetto delle covariate, individuare la presenza di possibili popolazioni latenti di strutture ospedaliere su ciascuna transizione e possibili pattern fra le transizioni.

Key words: Multi-state model, nonparametric discrete frailty, clinical administrative database

1 Multi-State models with frailty

Multi-State models are mathematical models through which we deal with both competing and progressive events. In multi-state framework, we model the history of a statistical unit (i.e., the clinical history of a patient) through a multi-state process $(X_t)\ t \geq 0$, where X_t denotes the state occupied by the statistical unit at time t. We define the set of the possible states as $S = \{0,1,2,...,J\}$ and $X_t \in S$. The mathematical quantities that define a multi-state process are the transition probability matrix, **P**(s,t) and the initial state of the process X_0. The elements of the transition probability matrix are defined as:

$$P_{lj}(s,t) := P(X_t = j | X_s = l, H_s), \quad s \leq t \quad j,l \in S; \tag{1}$$

which means that the element (l,j) of **P**(s,t) is the probability of moving from state l to state j between times s and t, given H_s, which is the history of the process up to time s. More formally, H_s is defined as the filtration generated by the process itself, $H_s = \mathscr{F}_{s^-}$. A typical assumption consists in the Markovianity of the process, which states that there is a complete independence between the future and the past:

$$P_{lj}(s,t) := P(X_t = j | X_s = l), \quad s \leq t \quad j,l \in S. \tag{2}$$

Eq (2) states that the next state to be visited and the transition time depends only on the present state. Putter et al. [13] referred to Markov multi-state models as 'clock forward' models, while mentioned 'clock reset' models as semi-Markov multi-state models. According to semi-Markov assumption, the evolution of the process depends on the last occupied state and on the sojourn time. Another important quantity in multi-state models is the transition intensity, which is the instantaneous probability of executing a transition:

$$\lambda_{lj}(t) := \lim_{\Delta t \to 0} \frac{P(X_{t+\Delta t} = j | X_t = l)}{\Delta t}, \quad j,l \in S. \tag{3}$$

If the transition intensity is time-dependent we have time-inhomogenous Markov process, if it is constant we have time-homogenous Markov process.

Several authors explored the theory behind multi-state processes [1] [6] [8] [13], and few packages have been published in R [2] [7]. However, only few authors dealt with multi-state models with frailty. Frailty terms are random terms that are usually introduced in modelling time-to-event data with a twofold aim: on one hand, for taking into account the unexplained heterogeneity of the data and, on the other hand, for taking into account a grouped structure of data. The greatest part of them proposed semi-markov Markov multi-state models [3] [10] [11]. The proposed frailties were generally transition-specific and group-specific (shared frailty) [10] [11], or subject-specific [3] [15]. Ripatti et al. [14] proposed mixed frailty terms, some of which shared across transitions, while Liquet et al. [10] proposed a joint frailty to link two transitions. However, all of them included parametric frailties, such as Gamma, Normal or Compound-Poisson.

2 Semi-Markov multi-state model with a nonparametric discrete frailty term

In this work, we propose a novel multi-state model for grouped and longitudinal data. In particular, we introduce a transition-specific nonparametric discrete frailty term. This kind of frailty has been proposed in the simpler framework of time-to-event data [5] and it allows to avoid any a priori specification of the shape of the frailty and to detect possible clusters of groups (latent populations of known groups, such as hospitals).

The hazard function for individual i in group j in transition l is:

$$\lambda^l(t; X_{ij}^l, w_k^l, z_{jk}^l) = \prod_{k=1}^{K^l} \left[\lambda_0(t)^l w_k^l \exp((X_{ij}^l)^T \beta^l) \right]^{z_{jk}^l}. \tag{4}$$

Mathematical assumptions and notations are the same as the ones in time-to-event framework [5]. In Eq.(4), we can recognize a Cox model (a nonparametric part made by an unspecified baseline function $\lambda_0(t)$, and a parametric exponential part made by a vector of patient-specific covariates X_{ij} and the associated regression parameters β) with a nonparametric frailty term $\mathbf{w}^l$. This term is modeled through a random variable with discrete distribution, with an unknown number of points in the support. In particular, we assume that each group j can belong to one latent population k, $k = 1, ..., K^l$, with probability π_k^l. In this case, $[w_1, ..., w_K]^l$ are the points in the support of $\mathbf{w}^l$, K^l is the support's cardinality and $\mathbf{P}\{w^l = w_k^l\} = \pi_k^l$. Also the number of latent population K depends on l, which means that the number of latent population detected is transition-specific. In order to build the model, we introduce an auxiliary indicator random variable z_{jk}^l which is equal to 1 if the j-th group belongs to the k-th population in the l-th transition, so $z_{jk}^l \overset{i.i.d}{\sim} Bern(\pi_k^l)$. The requirement $\sum_{k=1}^K z_{jk}^l = 1$, for each j and l, is equivalent to the assumption that each group belongs to only one population in each transition.

3 Application to a clinical administrative database

The proposed semi-Markov multi-state model with nonparametric discrete frailty has been applied to clinical administrative data related to patients with a diagnosis of heart failure and treated in the Lombardia Region, Italy. Despite of the fact that multi-state models are perfect tools for studying chronic conditions, they have been modestly used for investigating heart failure [4].

We propose a illness-death multi-state model with recovery, see Fig. 1, in which the state space S is composed by: hospital admission, hospital discharge and death (as unique absorbing state). We model the hazard function for each possible transition, $\lambda^l(t)$ ($l \in \{1,2,3,4\}$), according to Eq (4) and we set the same regression parameters for all transitions: age, gender, a binary variable which is 1 if the patient has more than three comorbidities and the number of procedures.

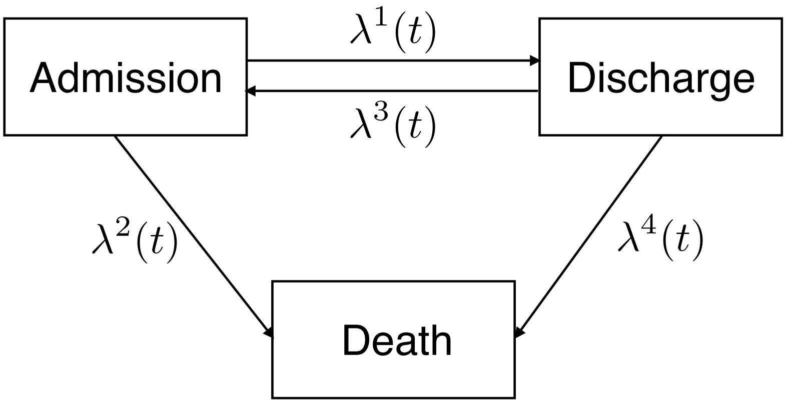


Fig. 1 Multi-state model proposed for the application to clinical administrative data.

The observed time window in the clinical administrative database that we studied is 2005 - 2012. In order to observe the complete clinical history of patients, we selected a cohort with only those patients whose first discharge was recorded between 2006 and 2007. Moreover, we selected those healthcare providers with 20-1,500 patients, in order to have more robust results. Than, the initial cohort is composed by $40,048$ patients.

For the sake of brevity, we report here only the results related to the first transition. In the first transition, we detected 7 latent populations (see Fig. 2). We observed that a patient hospitalized in a healthcare provider that belongs to latent population 7 has a higher risk (of 3.79 points) with respect to a patient with the same characteristics and hospitalized in a healthcare provider that belongs to latent population 1. The effect of covariates is coherent to what has been observed in a previous work on the same pathology [4]. Aging, having more than three comorbidities and having undergone a procedure lead to a lower risk of transition, which means a longer length of stay. Males have a higher risk of being discharged with respect to women.

The same evaluation can be done for the other transitions. For all of them, we decided to select the number of latent populations according to BIC index, which is, together with AIC, a very popular index in finite mixture model literature [12]. The other index that we show is the one proposed by Laird [9], which tends to overestimate the number of latent populations.

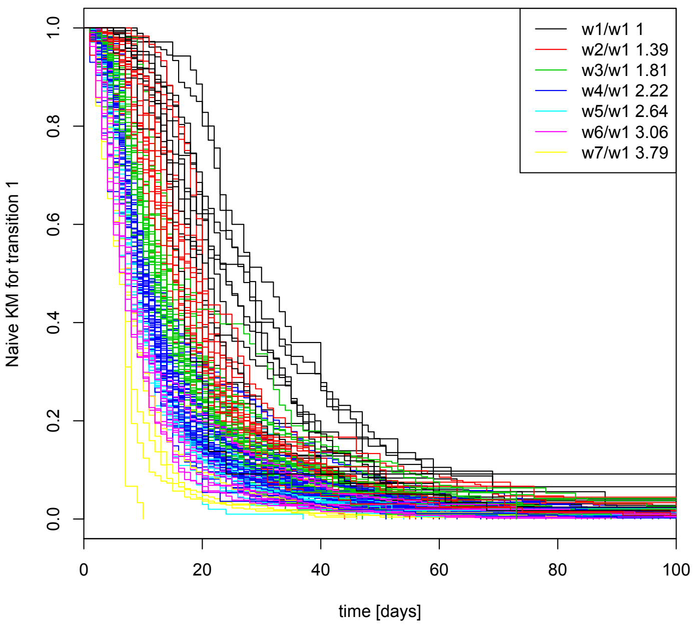


Fig. 2 Naive Kaplan-Meier for transition 1, a zoom in the first 100 days. The KM are naive because the considered censoring is not independent from the realization of the event of interest [13].

4 Conclusions

In this work, there are two sources of novelty: the inclusion of nonparametric discrete frailty in a semi-Markov model and the detection of latent populations of healthcare providers across transitions. The obtained results can be easily read and exploited by healthcare managers in decision making process, but also by clinicians in explorative analysis of clinical administrative databases.

Table 1 Estimates obtained through the proposed procedure in the first transition.

Transition	Latent populations	π	$\mathbf{w}/w_1$	β
1	K_{BIC} = 7 K_{AIC} =10 K_{Laird} = 11	π_1 = 0.09 π_2 = 0.14 π_3 = 0.25 π_4 = 0.28 π_5 = 0.12 π_6 = 0.08 π_7 = 0.04	w_1/w_1 = 1 w_2/w_1 = 1.39 w_3/w_1 = 1.81 w_4/w_1 = 2.22 w_5/w_1 = 2.63 w_6/w_1 = 3.06 w_7/w_1 = 3.79	β_{AGE} = -0.006 β_{SEX} = 0.069 β_{3COM} = -0.267 β_{NPRO} = -0.387

References

1. Andersen, P. K., Keiding, N.: Multi-state models for event history analysis. Statistical methods in medical research, **11**, 91–115. (2002)
2. de Wreede, L. C., Fiocco, M., Putter, H.: mstate: an R package for the analysis of competing risks and multi-state models. Journal of Statistical Software, 38, 1–30. (2011)
3. Foucher, Y., Saint-Pierre, P., Daures, J., Durand, J.: A semi-Markov frailty model for multistate and clustered survival data. Far East Journal of Theoretical Statistics, **19**, 185. (2006)
4. Gasperoni, F., Ieva, F., Barbati, G., Scagnetto, A., Iorio, A., Sinagra, G., Di Lenarda A.: Multistate modelling of heart failure care path: A population-based investigation from Italy. PloS one. **12**, e0179176. (2017)
5. Gasperoni, F.; Ieva, F.; Paganoni, A.M.; Jackson C.H.; Sharples L.D.: Nonparametric frailty Cox models for hierarchical time-to-event data. MOX report. **45**. (2017)
6. Hougaard, P.: Multi-state models: a review. Lifetime data analysis, **5**, 239–264. (1999)
7. Jackson, C. H.: Multi-state models for panel data: the msm package for R. Journal of Statistical Software, 38(8), 1–29. (2011)
8. Kalbfleisch, J. D., Lawless, J. F.: Likelihood analysis of multi?state models for disease incidence and mortality. Statistics in medicine, **7**, 149–160.(1988)
9. Laird, N.: Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association, **73**, 805–811. (1978)
10. Liquet, B., Timsit, J. F., Rondeau, V.: Investigating hospital heterogeneity with a multi-state frailty model: application to nosocomial pneumonia disease in intensive care units. BMC medical research methodology, **12**, 79. (2012)
11. Ma, T. Y., Joly, I., Raux, C.: A shared frailty semi-parametric markov renewal model for travel and activity time-use pattern analysis. (2010)
12. McLachlan, G., Peel, D.: Finite mixture models. John Wiley & Sons. (2004)
13. Putter, H., Fiocco, M., Geskus, R. B.: Tutorial in biostatistics: competing risks and multi?state models. Statistics in medicine. **26**, 2389–2430. (2007)
14. Ripatti, S., Gatz, M., Pedersen, N. L., Palmgren, J.: Three-state frailty model for age at onset of dementia and death in Swedish twins. Genetic epidemiology, **24**, 139–149. (2003)
15. Yen, A. M., Chen, T. H., Duffy, S. W., Chen, C. D.: Incorporating frailty in a multi-state model: application to disease natural history modelling of adenoma-carcinoma in the large bowel. Statistical methods in medical research, **19**, 529–546. (2010)

A Functional Urn Model for CARA Designs

Giacomo Aletti, Andrea Ghiglietti, and William F. Rosenberger

Abstract We present a general class of covariate-adjusted response-adaptive (CARA) designs introduced in [1], which is based on a new functional urn model. We show strong consistency concerning the allocation probability and the proportion of subjects assigned to the treatment groups, in the whole study and for each covariate profile, allowing the distribution of the responses conditioned on covariates to be estimated nonparametrically. We also establish joint central limit theorems for these quantities and the joint sufficient statistics, which allow construction of inference procedures.
Abstract *In questo lavoro presentiamo una classe generale di disegni covariate-adjusted adattivi alla risposta (CARA) introdotti in [1], che è basato su un nuovo modello d'urna funzionale. Inoltre, dimostriamo la forte consistenza della probabilità di allocazione e della proporzione di soggetti assegnati ai gruppi dei trattamenti, nell'intero studio e per ciascun valore delle covariate, permettendo alla distributione delle risposte condizionate alle covariate di essere stimata in maniera non parametrica. Infine, abbiamo stabilito alcuni teoremi centrale del limite congiunti di queste quantità e delle statistiche sufficienti, che permoettono la costruzione di procedure inferenziali.*

Key words: asymptotics, clinical trials, covariate-adjusted response-adaptive designs, randomization

1 Introduction

In CARA designs the patients in the trial are randomly assigned to $d \geq 2$ treatment groups with an allocation probability that depends on the current patient covariate profile and on the previous patients' covariates, allocations and responses (e.g. see [3]). In this framework, it is desirable that the proportion of subjects of each covariate profile assigned to the treatments converges to a desired target, defined as a function of the response distribution conditionally on the covariates.

Giacomo Aletti
ADAMSS Center and Università degli Studi di Milano e-mail: giacomo.aletti@unimi.it,

Andrea Ghiglietti
Università degli Studi di Milano e-mail: andrea.ghiglietti@unimi.it

William F. Rosenberger
George Mason University, Fairfax, VA, USA e-mail: wrosenbe@gmu.edu

Ideally, the analysis of the ethical and inferential properties of the experimental designs should be based on theoretical results concerning the asymptotic behavior of the allocation proportion and adaptive estimators, and none of the previous work on CARA designs is able to provide such results. In fact, since the allocation and the estimation process depend on both the responses and the covariates, CARA designs are very complex to be formulated in a rigorous mathematical setting. Two papers, in particular, formalize CARA in a rigorous mathematical framework. The first of these is the groundbreaking paper of [4], in which consistency and second-order asymptotic results concerning both adaptive estimators and allocation proportions have been proved for a very wide class of CARA designs. In the second [2], compound optimal design theory was used to find target allocations of interest, and these target allocations are attained using an accelerated biased coin design.

Here we present a class of CARA designs introduced in [1], in which the allocation probability may depend on nonparametric estimates of the response distribution, and the patients' covariate profiles are not identically distributed.

2 The model

For any $n \geq 0$, let $\mathbf{Y}_n = (Y_n^1, .., Y_n^d)^\top$ be a vector of functions, with $Y_n^j : \tau \mapsto (0,1)$, where τ is the covariate space. For any $t \in \tau$, $\mathbf{Y}_n(t)$ represents an urn containing $Y_n^j(t)$ balls of color $j \in \{1, .., d\}$ and $\mathbf{Z}_n(t) = \mathbf{Y}_n / \sum_{j=1}^d Y_n^j$ indicates the proportion of the colors.

When subject n enters the trial, his covariate profile T_n is observed. Then, a ball is sampled at random from the urn identified by T_n (i.e. with proportions $\mathbf{Z}_{n-1}(T_n)$), its color is observed and represented by $\bar{\mathbf{X}}_n$: $\bar{X}_n^j = 1$ when the color is $j \in \{1, .., d\}$, $\bar{X}_n^j = 0$ otherwise. Then, subject n receives the treatment associated to the sampled color and a response $\bar{\xi}_n$ is collected. The functional urn is then updated as: $\mathbf{Y}_n = \mathbf{Y}_{n-1} + D_n \mathbf{X}_n$, where $\mathbf{X}_n$ and D_n are appropriately defined. Specifically, the weighting function $\mathbf{X}_n : \tau \mapsto [0,1]^d$ should be such that, for any $t \in \tau$ and $j \in \{1, .., d\}$, $\sum_{j=1}^d X_n^j(t) = 1$ and $\mathbf{E}[X_n^j(t) | \mathscr{F}_{n-1}, T_n] = Z_n^j(t)$, where $\mathscr{F}_{n-1}$ is the σ-algebra of the information related with the first $(n-1)$ patients. This is straightforward for $t = T_n$ by setting $X_n^j(T_n) = \bar{X}_n^j$, since $\bar{X}_n^j$ is conditionally on $\mathscr{F}_{n-1}$ and T_n Bernoulli distributed with parameter $Z_n^j(T_n)$. Then, we define a family of Bernoulli random variables $\{\check{X}_n^j(t); t \in \tau\}$ with parameters $\{Z_n^j(t); t \in \tau\}$, representing the color that would be sampled in the trial if the covariate profile of subject n was equal to any $t \in \tau$. Finally, we use the quantile function that links this family to compute $\mathbf{X}_n(t)$ for all $t \in \tau$ as $\mathbf{X}_n := \mathbf{E}[\check{\mathbf{X}}_n | \mathscr{F}_{n-1}, T_n, \bar{\mathbf{X}}_n]$. Analogously, we can define the replacement functional matrix $D_n : \tau \mapsto [0,1]^{d \times d}$ as $D_n := \mathbf{E}[\check{D}_n | T_n, \bar{\mathbf{X}}_n, \bar{\xi}_n]$, where $\check{D}_n(t)$ is a function of a random variable having the same distribution of the response observed from a subject with covariate profile t, i.e. the response that would be observed in the trial if the covariate profile of subject n was equal to any $t \in \tau$. Naturally, $D_n(T_n) = \check{D}_n(T_n)$. Since the quantile functions of the response distributions are typically unknown, D_n is computed by using the corresponding (parametric or nonparametric) estimators obtained with the information in $\mathscr{F}_{n-1}$.

The key feature of the design is that quantile functions are used to update *all* urns, not just the urn for which $T_n = t$. In theory there could be an uncountably infinite number of urns, with only a finite subset of them used for patient allocation. However, in clinical practice, mathematically "continuous" covariates are really not continuous; for instance, cholesterol is represented by integer values, likely in some range, that would, for all intents and purposes, make it a finite discrete covariate. However, the procedure is well-defined for uncountably infinite urns, and first order asymptotic properties can be obtained, although some of the covariate-specific metrics do not make sense in that context. When we move to second-order asymptotics, we partition τ into K strata, which could be intervals of a continuous set.

3 Consistency Results

We now present some consistency results for (i) the probability of allocation of the subjects for each covariate profile ($\mathbf{Z}_n(t)$), (ii) the proportion of subjects associated to each covariate profile assigned to the treatments ($\mathbf{N}_{t,n}/\sum_{j=1}^{d} N_{t,n}^{j}$, where $\mathbf{N}_{t,n} := \sum_{i=1}^{n} \bar{\mathbf{X}}_i \mathbf{1}_{\{T_i=t\}}$), (iii) the proportion of subjects assigned to the treatments ($\mathbf{N}_n/n$, where $\mathbf{N}_n := \sum_{i=1}^{n} \bar{\mathbf{X}}_i$). Consider the following assumptions:

(A1) for any $t \in \tau$ and $n \geq 1$, $D_n^{\top}\mathbf{1} = \mathbf{1}$ (*constant balance*);
(A2) denoting by $H(t) := \mathbf{E}[\breve{D}_1(t)]$ the average replacement when the covariate profile is t, we assume that $H(t)$ is irreducible, diagonalizable and there exists $\alpha > 0$ such that $\mathbf{E}[|\mathbf{E}[D_n(t)|\mathscr{F}_{n-1}, T_n, \bar{\mathbf{X}}_n] - H(t)|\mathscr{F}_{n-1}] = O(n^{-\alpha})$.

Denote by $\mathbf{v}(t)$ the right eigenvector of $H(t)$ associated to $\lambda = 1$, with $\sum_{j=1}^{d} v^j(t) = 1$, and let μ_{n-1} be the probability distribution of T_n conditioned on $\mathscr{F}_{n-1}$. Then,

(a) for any probability measure ν on τ, we have $\int_{\tau} \|\mathbf{Z}_n(t) - \mathbf{v}(t)\| \nu(dt) \overset{a.s.}{\to} 0$;
(b) if $\sum_{i=1}^{n} \mu_{i-1}(\{t\}) \overset{a.s.}{\to} \infty$, we have $\|\mathbf{N}_{t,n}/\sum_{j=1}^{d} N_{t,n}^{j} - \mathbf{v}(t)\| \overset{a.s.}{\to} 0$;
(c) if $\int_{\tau} |\mu_n(dt) - \mu(dt)| \overset{a.s.}{\to} 0$, we have $\|\mathbf{N}_n/n - \int_{\tau} \mathbf{v}(t)\mu(dt)\| \overset{a.s.}{\to} 0$.

The convergence results consider a general covariate space τ. In order to show second-order properties, we now partition τ into K finite elements, which could, for instance, be K intervals of a continuous covariate space. This partitioning induces K urns used to allocate subjects with covariate profiles in the set $\{1,...,K\}$. In clinical trials practice, K must be considerably smaller than the total sample size.

4 Central Limit Theorems

We now present further assumptions that are required for establishing the second-order asymptotic properties.

- **Finite partition of the covariate space.** We assume that the covariate space τ is composed by a finite number $K \in \mathbb{N}$ of distinct elements. When τ contains infinite elements, we can take a partition of τ, i.e. $\{\tau_1,..,\tau_K\}$ such that $\cup_k \tau_k = \tau$ and $\tau_{k_1} \cap \tau_{k_2} = \emptyset$ for $k_1 \neq k_2$, and consider these sets to be the elements of τ, i.e. $\tau := \{\tau_1,..,\tau_K\}$. To facilitate the notation, without loss of generality in the sequel we redefine $\tau = \{1,..,K\}$ and $\mu_{n-1}(t) = \mu_{n-1}(\{t\}) = \mathbb{P}(T_n = t|\mathscr{F}_{n-1})$ for any $t \in \tau$.
- **Conditional response distributions.** The analog of the null hypothesis in classical inferential statistics is given here by assuming that the conditional response distributions $\pi_t^1,..,\pi_t^d$ are known for any $t \in \tau$. As a direct consequence, we have that $D_n = D_n^*$ and $H_n = H$ with probability one for any $n \geq 1$.
- **Eigenvalues of the limiting generating matrix.** Denoting $\lambda_H^*(t)$ the eigenvalue of $Sp(H(t)) \setminus \{1\}$ with largest real part, assume that $\max_{t\in\tau} \mathscr{R}e(\lambda_H^*(t)) < 1/2$.
- **Dynamics of adaptive estimators.**

- *(Covariate-stratification approach)* For some $t \in \tau$ and $j \in \{1,..,d\}$, consider that there are features of interest θ_t^j related with the distribution π_t^j of the responses to treatment j conditionally on the covariate profile t. Then, we assume that the corresponding adaptive estimator $\hat{\theta}_{t,n}^j$ is strongly consistent and its dynamics can be expressed as follows: there exists $n_0 \geq 1$ such that for any $n \geq n_0$

$$\hat{\theta}_{t,n}^j - \hat{\theta}_{t,n-1}^j = -\frac{\bar{X}_n^j \mathbb{1}_{\{T_n=t\}}}{N_{t,n}^j}(f_{t,j}(\hat{\theta}_{t,n-1}^j) - \Delta\mathbf{M}_{t,j,n} - \mathbf{R}_{t,j,n}), \tag{1}$$

where
  (i) $f_{t,j}$ is a Lipschitz continuous function such that $f_{t,j}(\theta_t^j) = 0$;
  (ii) $\Delta\mathbf{M}_{t,j,n} \in \mathscr{F}_n$ is a martingale increment such that $\mathbb{E}[\Delta\mathbf{M}_{t,j,n}|\mathscr{F}_{n-1}, T_n, \bar{X}_n^j] = 0$, and it converges stably to $\Delta\mathbf{M}_{t,j}$ with kernel K independent of $\mathscr{F}_{n-1}$:
  $\mathscr{L}(\Delta\mathbf{M}_{t,j,n}|\mathscr{F}_{n-1}, T_n = t, \bar{X}_n^j = 1) \overset{a.s.}{\to} K(t,j)$;
  (iii) $\mathbf{R}_{t,j,n} \in \mathscr{F}_n$ is such that $n\mathbb{E}[\|\mathbf{R}_{t,j,n}\|^2] \to 0$.
  Moreover, let $f_{t,j}$ be differentiable at θ_t^j, denote by $\lambda^*_{\theta_t^j}$ the eigenvalue of $Sp(\mathscr{D}f_{t,j}(\theta_t^j))$ with largest real part and assume that $\min_{t\in\tau}\mathscr{R}e(\lambda^*_{\theta_t^j}) > 1/2$. We also assume that for some $\delta > 0$,

$$\sup_{n\geq 1}\mathbb{E}\left[\|\Delta\mathbf{M}_{t,j,n}\|^{2+\delta} | \mathscr{F}_{n-1}\right] < +\infty a.s., \tag{2}$$

and

$$\mathbb{E}\left[\Delta\mathbf{M}_{t,j,n}(\Delta\mathbf{M}_{t,j,n})^\top | \mathscr{F}_{n-1}\right] \xrightarrow[n\to+\infty]{a.s.} \Gamma_{t,j}, \tag{3}$$

where $\Gamma_{t,j}$ is a symmetric positive matrix.
- *(Covariate-adjusted approach)* For some $j \in \{1,..,d\}$, consider that there are features of interest β^j related with the entire family of distributions $\{\pi_t^j; t \in \tau\}$ of the responses to treatment j conditionally on the covariates. Then, we assume that the corresponding adaptive estimator $\hat{\beta}_n^j$ is strongly consistent and its dynamics can be expressed as follows:

$$\hat{\beta}_n^j - \hat{\beta}_{n-1}^j = -\frac{\bar{X}_n^j}{N_n^j}(f_j(\hat{\beta}_{n-1}^j) - \Delta\mathbf{M}_{j,n} - \mathbf{R}_{j,n}), \tag{4}$$

where the quantities in (4) fulfill the same conditions presented above for the dynamics (1).

We first provide the convergence rate and the joint asymptotic distribution concerning the quantities of interest in the design in the framework of covariate-stratification response-adaptive designs. This result is established in the following central limit theorem. We introduce the variables independent of $\sigma(\mathscr{F}_n; n \geq 1)$: $T \in \tau$ with distribution $\mu(t)$, $\bar{\mathbf{X}} \in \{0,1\}^d \in \mathscr{S}$ such that $\mathbb{P}(\bar{X}^j = 1|T) = v^j(T)$, $D := \mathbb{E}[\check{D}|T, \bar{\mathbf{X}}, \xi]$, where the distribution of ξ conditioned on $\{T = t\}$ and $\{\bar{X}^j = 1\}$ is π_t^j.

Theorem 4.1. *Define* $\mathbf{W}_n := (\mathbf{Z}_n(t), \mathbf{N}_{t,n}/w(\mathbf{N}_{t,n}), \hat{\theta}_{t,n}, t \in \tau)^\top$, $\mathbf{W} := (\mathbf{v}(t), \mathbf{v}(t), \theta_t, t \in \tau)^\top$. *Then,*

$$\mu_n(t) \xrightarrow{a.s.} \mu(t) = f_{\mu,t}(\mathbf{v}(t), \theta_t), \qquad \mathbf{W}_n \xrightarrow{a.s.} \mathbf{W}, \tag{5}$$

$$\sqrt{n}(\mathbf{W}_n - \mathbf{W}) \xrightarrow{\mathscr{L}} \mathscr{N}(\mathbf{0}, \Sigma), \qquad \Sigma := \int_0^\infty e^{u(\frac{\mathbf{I}}{2}-A)}\Gamma e^{u(\frac{\mathbf{I}}{2}-A^\top)}du, \tag{6}$$

where

$$A := \begin{pmatrix} A_{ZZ} & 0 & 0 \\ -I & I & 0 \\ 0 & 0 & A_{\theta\theta} \end{pmatrix}, \qquad \Gamma := \begin{pmatrix} \Gamma_{ZZ} & \Gamma_{ZN} & \Gamma_{Z\theta} \\ \Gamma_{ZN}^{\top} & \Gamma_{NN} & 0 \\ \Gamma_{Z\theta}^{\top} & 0 & \Gamma_{\theta\theta} \end{pmatrix},$$

and A_{ZZ}, $A_{\theta\theta}$, Γ_{NN}, $\Gamma_{\theta\theta}$ *are block-diagonal matrices whose* t^{th} *block is*

(i) $A_{ZZ}^{tt} = (I - H(t) + \mathbf{v}(t)\mathbf{1}^{\top})$;
(ii) $A_{\theta\theta}^{tt}$ *is a block-diagonal matrices whose* j^{th} *block is* $[A_{\theta\theta}^{tt}]^{jj} := \mathscr{D} f_{t,j}(\theta_t^j)$;
(iii) $\Gamma_{NN}^{tt} := \mu^{-1}(t)(diag(\mathbf{v}(t)) - \mathbf{v}(t)\mathbf{v}^{\top}(t))$;
(iv) $\Gamma_{\theta\theta}^{tt}$ *is a block-diagonal matrices whose* j^{th} *block is*
$[\Gamma_{\theta\theta}^{tt}]^{jj} := (v^j(t)\mu(t))^{-1}\mathbb{E}[\Delta\mathbf{M}_{t,j}(\Delta\mathbf{M}_{t,j})^{\top}|T = t, \bar{X}^j = 1]$;

and Γ_{ZZ}, Γ_{ZN}, $\Gamma_{Z\theta}$ *are matrices defined as follows: for any* $t_1, t_2 \in \tau$

(v) $\Gamma_{ZZ}^{t_1t_2} := \mathbb{E}[D(t_1)\mathbf{g}(t_1, T, \bar{\mathbf{X}})\mathbf{g}^{\top}(t_2, T, \bar{\mathbf{X}})D^{\top}(t_2)] - \mathbf{v}(t_1)\mathbf{v}^{\top}(t_2)$;
(vi) $\Gamma_{ZN}^{t_1t_2} := H(t_1)G(t_1,t_2)diag(\mathbf{v}(t_2)) - \mathbf{v}(t_1)\mathbf{v}^{\top}(t_2)$;
(vii) $[\Gamma_{Z\theta}^{t_1t_2}]^j := \mathbb{E}[D(t_1)\mathbf{g}(t_1,t_2,\mathbf{e}_j)\Delta\mathbf{M}_{t_2,j}^{\top}|T = t_2, \bar{X}^j = 1]$;

where $\mathbf{g}$ *is a d-multivariate function with values in* $\mathscr{S}$ *and* $G(t_1,t_2)$ *is a matrix with columns* $\{\mathbf{g}(t_1,t_2,\mathbf{e}_j); j \in \{1,..,d\}\}$.

We now provide the convergence rate and the joint asymptotic distribution of the quantities interest in the design in the framework of covariate-adjusted response-adaptive designs. This result is established in the following central limit theorem.

Theorem 4.2. *Define* $\mathbf{W}_n := (\mathbf{Z}_n(t), t \in \tau, \mathbf{N}_n/n, \hat{\beta}_n)^{\top}$, $\mathbf{W} := (\mathbf{v}(t), t \in \tau, \mathbf{x}_0, \beta)^{\top}$. *Then,*

$$\mu_n(t) \xrightarrow{a.s.} \mu(t) = f_{\mu,t}(\mathbf{x}_0, \beta), \qquad \mathbf{W}_n \xrightarrow{a.s.} \mathbf{W}, \tag{7}$$

$$\sqrt{n}(\mathbf{W}_n - \mathbf{W}) \xrightarrow{\mathscr{L}} \mathscr{N}(\mathbf{0}, \Sigma), \qquad \Sigma \ := \int_0^{\infty} e^{u(\frac{\mathbf{I}}{2} - A)} \Gamma e^{u(\frac{\mathbf{I}}{2} - A^{\top})} du, \tag{8}$$

and

$$A := \begin{pmatrix} A_{ZZ} & 0 & 0 \\ A_{NZ} & A_{NN} & A_{N\beta} \\ 0 & 0 & A_{\beta\beta} \end{pmatrix}, \qquad \Gamma := \begin{pmatrix} \Gamma_{ZZ} & \Gamma_{ZN} & \Gamma_{Z\beta} \\ \Gamma_{ZN}^{\top} & \Gamma_{NN} & 0 \\ \Gamma_{Z\beta}^{\top} & 0 & \Gamma_{\beta\beta} \end{pmatrix},$$

where again and A_{ZZ}, $A_{\beta\beta}$, $\Gamma_{\beta\beta}$ *are block-diagonal matrices whose* t^{th} *or* j^{th} *block is*

(i) $A_{ZZ}^{tt} = (I - H(t) + \mathbf{v}(t)\mathbf{1}^{\top})$;
(ii) $A_{\beta\beta}^{jj} = \mathscr{D} f_j(\beta^j)$;
(iii) $\Gamma_{\beta\beta}^{jj} := (\mathbb{E}[v^j(T)])^{-1}\mathbb{E}[\Delta\mathbf{M}_j(\Delta\mathbf{M}_j)^{\top}|\bar{X}^j = 1]$;

and

(iv) $A_{NN} := I - \sum_{s=1}^{K} \mathbf{v}(s)\mathscr{D}_N f_{\mu,s}(\mathbf{x}_0, \beta)^{\top}$;
(v) $A_{N\beta} := -\sum_{s=1}^{K} \mathbf{v}(s)\mathscr{D}_{\beta} f_{\mu,s}(\mathbf{x}_0, \beta)^{\top}$;
(vi) $\Gamma_{NN} := diag(\mathbb{E}[\mathbf{v}(T)]) - \mathbb{E}[\mathbf{v}(T)]\mathbb{E}[\mathbf{v}^{\top}(T)]$;

and A_{NZ}, Γ_{ZZ}, Γ_{ZN}, $\Gamma_{Z\beta}$ *are matrices defined as follows: for any* $t_1, t_2 \in \tau$

(vii) $A_{NZ}^{t_2} := -\mu(t_2)I$;

(viii) $\varGamma_{ZZ}^{t_1t_2} := \mathbb{E}[D(t_1)\mathbf{g}(t_1,T,\bar{\mathbf{X}})\mathbf{g}^\top(t_2,T,\bar{\mathbf{X}})D^\top(t_2)] - \mathbf{v}(t_1)\mathbf{v}^\top(t_2)$*;*
(ix) $\varGamma_{ZN}^{t_1} := H(t_1)\mathbb{E}[G(t_1,T)diag(\mathbf{v}(T))] - \mathbf{v}(t_1)\mathbb{E}[\mathbf{v}^\top(T)]$*;*
(x) $\varGamma_{Z\beta}^{t_1j} := \mathbb{E}[D(t)\mathbf{g}(t_1,T,j)\Delta\mathbf{M}_j^\top|\bar{X}^j = 1]$.

where we recall that $\mathbf{g}$ *is a* d*-multivariate function with values in* $\mathscr{S}$ *and* $G(t_1,t_2)$ *is a matrix with columns* $\{\mathbf{g}(t_1,t_2,\mathbf{e}_j); j \in \{1,..,d\}\}$.

Remark 4.1. *We recall that Theorem 4.1 allows inferential procedures based on stratified estimators, while Theorem 4.2 allows inference on covariate-adjusted regression parameters representing the covariate-adjusted treatment effect.*

References

1. Aletti G., Ghiglietti A., Rosenberger W.F. (2018). Nonparametric covariate-adjusted response-adaptive design based on a functional urn model, *Ann. Statist.*, in press.
2. Baldi Antognini, A. and Zagoraiou, M. (2012). Multi-objective optimal designs in comparative clinical trials with covariates: the reinforced doubly adaptive biased coin design, *Ann. Statist.* **40** 1315–1345.
3. Hu F., Rosenberger W.F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*, John Wiley & Sons, New York.
4. Zhang L.-X., Hu F., Cheung S. H., Chan W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs, *Ann. Statist.* **35** 1166–1182.

Assessment of the INLA approach on gerarchic bayesian models for the spatial disease distribution: a real data application

Valutazione dell'approccio basato su INLA in modelli gerarchici bayesiani per la distribuzione spaziale di malattia: un'applicazione a dati reali

Paolo Girardi (1), Emanuela Bovo (2), Carmen Stocco (2), Susanna Baracco (2), Alberto Rosano (2), Daniele Monetti (2), Silvia Rizzato (2), Sara Zamberlan (2), Enrico Chinellato (2), Ugo Fedeli (1), Massimo Rugge (2,3)

Sommario The use of approximate methods as the INLA (Integrated Nested Laplace Approximation) approach is being widely used in Bayesian inference, especially in spatial risk model estimation where the Besag-York-Molliè (BYM) model has found a proper use. INLA appears time saving compared to Monte Carlo simulations based on Markov Chains (MCMC), but it produces some differences in estimates [1, 2]. Data from the Veneto Cancer Registry has been considered with the scope to compare cancer incidence estimates with INLA method and with two other procedures based on MCMC simulation, WinBUGS and CARBayes, under R environment. It is noteworthy that INLA returns estimates comparable to both MCMC procedures, but it appears sensitive to the a-priori distribution. INLA is fast and efficient in particular with samples of moderate-high size. However, care must to be paid to the choice of the parameter relating to the a-priori distribution.

Sommario *L'uso dei metodi basati sull'approssimazione di Laplace come INLA (Integrated Nested Laplace Approximation) è ampiamente utilizzato nell'inferenza Bayesiana, specialmente in modelli di rischio spaziale dove il modello di Besag-York-Molliè (BYM) ha trovato un uso appropriato. INLA permette un risparmio di tempo computazionale rispetto alle simulazioni Monte Carlo basate su Catene Markov (MCMC), ma produce alcune differenze nelle stime [1, 2]. Vengono considerati i dati del Registro dei Tumori del Veneto con lo scopo di di confrontare le stime ottenute con INLA rispetto a due procedure basata su MCMC, WinBUGS e CARBayes, svolte in ambiente R. E' importante notare che INLA restituisce stime comparabili ad entrambe le procedure MCMC, ma è sensibile alla distribuzione a priori. INLA è un metodo rapido ed efficiente, in particolare con campioni di elevata numerosità. Tuttavia, occorre prestare attenzione alla scelta del parametro relativo alla distribuzione a priori.*

(1) Sistema Epidemiologico Regionale, Azienda Zero, Padova.
mail: paolo.girardi@aulss6.veneto.it
(2) Registro Tumori del Veneto, Azienda Zero, Padova
(3) Dipartimento di Medicina DIMED, Università di Padova, Padova

Key words: BYM model, Cancer Registry, INLA, Laplace approximation, Bayesian methods

1 Introduction

In recent literature, the use of approximate methods in Bayesian inference has reported a great popularity. The Laplace approximation proposed by [Rue H., 2009] with the INLA acronym (Integrated Nested Laplace Approximation) has been adapted to the parameter estimations of an increasing number of statistical models; in addition, several papers have reported its use in wide range of real data applications. INLA offers the opportunity to perform Bayesian analyses through numerical integration avoiding extensive iterative computation; it usually implies a lower computational time respect to the classical Monte Carlo simulations based on Markov Chains (MCMC) with dedicated software (WinBUGS, OpenBUGS or JAGS). The major gain of INLA is the replacing of long chains used by MCMC methods to produce a-posteriori estimates of the coefficients distribution with a Laplace approximation of the a-posteriori distribution. Among hierarchical Bayesian models, the Besag-York-Molliè (BYM) model [4] has became popular for the analysis of spatial distribution of occurrences in epidemiology (disease risk, mortality, etc...), in financial services (investments, prices) and in demography and sociology (deprivation index, unemployment rate, etc..). The availability of the INLA package for R software [5] has allowed an easy and friendly implementation of INLA for BYM models. However, recent publications show that INLA produces considerable differences in estimates [1, 2] and research on this topic remains already unexplored. The aim of the study is to compare risk estimates produced by INLA with those one of the MCMC simulations using a series of real data applications instead of simulations.

2 Materials and methods

2.1 Veneto Cancer Registry

We consider all the cases of malignant cancers occurring in the year 2013 in the Veneto Region, one of the largest Region in Italy covering about five million of inhabitants. The area covered by the Veneto Cancer Registry includes the 96% of the territory (Figure 1). Every cancer case has been coded with the X version of International Classification of the Diseases (ICD-X) and has been aggregated at the municipality level (n=556 municipalities). In our comparison we consider 7 different primitive sites that have different number of cases: all the sites except skin, Hodgkin's lymphoma, myeloma and pancreas cancer among men; cancer of breast, cervix and cancer of esophagus among women.

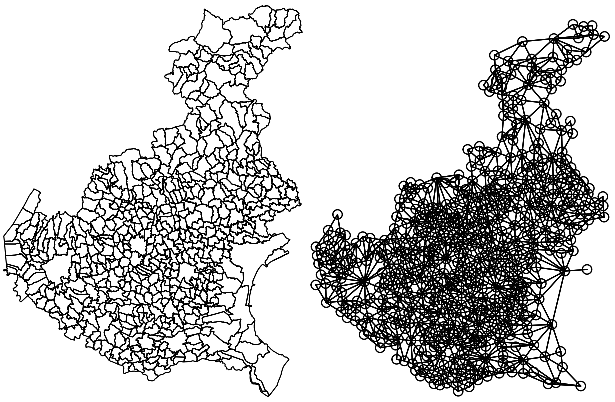


Figura 1 Boundaries and adjacency matrix of the 556 municipalities.

2.2 BYM model

The Standardized Incidence Ratios (SIR) have been estimated by means of a BYM model. The number of observed cases O_i is assumed to follow a Poisson distribution as

$$O_i \sim Poisson(\lambda_i) \tag{1}$$

where λ_i is the mean/variance parameter. Considering e_i the expected number of cases for the $i-th$ area calculated by an indirect standardization using the registry pool as reference, the estimated SIR is connected to the linear predictor η_i as follows

$$log(SIR_i) = log(\frac{\lambda_i}{e_i}) = \eta_i = (\alpha + \mu_i + \nu_i), \tag{2}$$

where α is the intercept quantifying the average incidence rate in all the 556 municipalities, while μ_i and ν_i are the correlated and uncorrelated spatial effects, following a normal distribution. While τ_ν is assumed to be distributed as a white noise ($\nu_i \sim N(0, \frac{1}{\tau_\nu})$), the μ_i distribution is modelled using an intrinsic conditional autoregressive structure (ICAR) as follow

$$\mu_i \sim N(\frac{1}{n_j} \sum_{\partial j} O_j, \frac{1}{n_j \tau_\mu}) \tag{3}$$

where O_j are the cases observed in ∂j which denotes the n_j municipalities bordering the *i-th* area, *i-th* area excluded. The precision parameters τ_μ and τ_ν follow a Gamma distribution. SIR has been estimated by INLA using R-INLA and by MCMC procedures using R2WinBUGS and CarBayes packages under R environment. For MCMC simulation, we took into account the results of 15.000 iterations discarding the first 5.000 as burn-in.

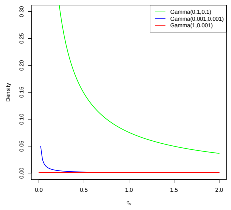


Figura 2 Distributions of the parameter τ_μ.

The study considers three different distributions for the precision of the spatial parameter τ_μ: $\Gamma(0.1,0.1)$, $\Gamma(0.001,0.001)$ and $\Gamma(1,0.001)$. The parameter of the precision related to the uncorrelated spatial effect has been fixed to be distributed as a $\Gamma(0.001,0.001)$.

3 Results

The main characteristics of the selected cancer sites are reported in Table 1. The sites are ordered decreasing the number of observed cases. A high number of male cases (15'416) is registered taking into account all the cancer sites except the skin; conversely, a low number of cervix cancers among women is reported, equal to 200, less than 1 per municipality. The average number of cases for municipality is always lower than the variance estimates indicating an over-dispersion. The p-values associated to the Moran's I test applied to empirical SIR ($\frac{o_i}{e_i}$) support the spatial independence for the distribution of each considered primitive site.

Cancer site	**Total cases**	Average	Variance	Moran's I test (p-value)
All sites (men)	15'416	27.7	5063.8	0.430
Breast (women)	4'372	7.9	535.5	0.185
Pancreas (men)	535	1.0	8.10	0.065
Cervix cancer (women)	200	0.4	1.5	0.796
Myeloma (men)	199	0.4	1.2	0.758
Hodgkin's lymphoma (men)	101	0.2	0.6	0.483
Esophagus (women)	59	0.1	0.4	0.163

Tabella 1 Characteristics of the selected cancer sites and p-value associated to Moran's I test for the empirical SIR.

SIR estimates are calculated for each selected cancer site varying the distribution of the precision parameter τ_μ computing the Pearson correlation index between INLA and MCMC-based estimates. The results are reported in Table 2. The correlation indices ranges from 0.344 in esophageal cancer with a-priori distribution $\tau_\mu \sim \Gamma(1,0.001)$, which indicates a poor agreement between INLA and CARBayes estimates, to 0.998/0.996 relatively to all male cancer sites with $\tau_\mu \sim \Gamma(0.1,0.1)$ resulting in a perfect overlapping between INLA and WinBUGS/CARBayes methods. The degree of agreement between INLA and MCMC procedures depends on: 1) the a-priori distribution of the variance of spatial component; 2) the number of incident cases. Overall, the best agreement (all r' Pearson indices >0.9) is obtained choosing a $\Gamma(0.1,0.1)$ for the τ_μ. As reported in Table 3 INLA returns estimates faster than MCMC procedure (about 15/20 times).

		Distribution of τ_μ		
Cancer site	**MCMC procedures**	$\Gamma(0.1,0.1)$	$\Gamma(0.001,0.001)$	$\Gamma(1,0.001)$
All cancers (men)	WinBUGS / CarBayes	0.998 / 0.996	0.992 / 0.987	0.990 / 0.985
Breast (women)	WinBUGS / CarBayes	0.997 / 0.995	0.994 / 0.988	0.992 / 0.983
Pancreas (men)	WinBUGS / CarBayes	0.997 / 0.995	0.987 / 0.949	0.983 / 0.976
Cervix cancer (women)	WinBUGS / CarBayes	0.986 / 0.945	0.966 / 0.947	0.961 / 0.872
Myeloma (men)	WinBUGS / CarBayes	0.910 / 0.966	0.925 / 0.963	0.948 / 0.969
Hodgkin's lymphoma (men)	WinBUGS / CarBayes	0.981 / 0.917	0.763 / 0.850	0.930 / 0.882
Esophagus (women)	WinBUGS / CarBayes	0.955 / 0.935	0.858 / 0.937	0.802 / 0.344

Tabella 2 Correlation index between INLA and MCMC-based methods on SIR estimates by distribution of τ_μ.

Although r's pearson index indicates a high agreement between INLA estimates compared to the MCMC-based methods, the graphical analysis permits to verify the presence of marked differences in the estimated risks (Figure 3), for example, relatively to the Myeloma SIR. The difference is marked considering the spatial distribution of the esophageal cancer incidence among women obtained by a comparison between INLA and CARBayes procedures (Fig. 4) that in Table 2 reports a weak agreement.

Procedures	$\Gamma(0.1,0.1)$	$\Gamma(0.001,0.001)$	$\Gamma(1,0.001)$
INLA	5.50	4.04	9.7
WinBUGS	87.25	90.99	85.93
CARBayes	68.7	64.1	68.9

Tabella 3 Computation time for INLA, WinBUGS and CARBayes esophageal SIR estimates.

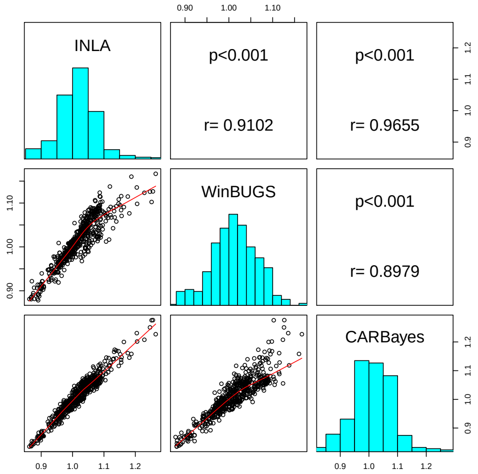
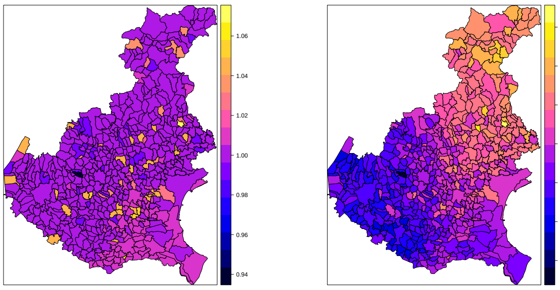


Figura 3 Esophagus SIR distribution among women with $\tau_\mu \sim \Gamma(0.1,0.1)$ estimated with INLA and with two MCMC-based procedures (WinBUGS and CARBayes).

4 Conclusions

In presence of non-informative a-priori distributions, INLA and MCMC procedures reported different estimates, even more clean-cut considering low sample size. INLA confirms to be a fast and efficient method for spatial risk estimation and, in general, for hierarchical Bayesian models [6, 7]. However, in order to avoid an over-smoothing of the risks and/or excessive imprecision of the estimates, particular attention must to be paid to the choice of the a-priori distribution for the variance of the spatial component. Further analyses are required in order to assess the comparability of INLA and MCMC estimates looking at the distribution of the uncorrelated spatial parameter and at the presence of spatial dependence.



(a) INLA SIR estimates $\tau_\mu \sim \Gamma(1, 0.001)$

(b) CARBayes MCMC SIR estimates $\tau_\mu \sim \Gamma(1, 0.001)$

Figura 4 Distribution of Esophageal SIR estimated by INLA and CARBayes.

Riferimenti bibliografici

1. De Smedt, T., Simons, K., Van Nieuwenhuyse, A., Molenberghs, G. (2015). Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. Archives of Public Health, 73(1), O2.
2. Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M., Watjou, K. (2015). Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. Spatial and spatio-temporal epidemiology, 14, 45-54.
3. Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology), 71(2), 319-392.
4. Besag, J., York, J., Mollie, A. Bayesian image restoration with two applications in spatial statistics (with discussion) Ann Inst Stat Math. 1991; 43: 1–59. doi: 10.1007. BF00116466
5. Lindgren, F., Rue, H. (2015). Bayesian spatial modelling with R-INLA. Journal of Statistical Software, 63(19).
6. Blangiardo, M., Cameletti, M., Baio, G., Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology, 4, 33-49.
7. Bilancia, M., Demarinis, G. (2014). Bayesian scanning of spatial disease rates with integrated nested Laplace approximation (INLA). Statistical Methods & Applications, 23(1), 71-94.

Medicine

Hidden Markov Models for disease progression

Hidden Markov Models per la progressione di patologia

Andrea Martino, Andrea Ghiglietti, Giuseppina Guatteri, Anna Maria Paganoni

Abstract Disease progression models are a powerful tool for understanding and predicting the development of a disease, given some longitudinal measurements obtained from a sample of patients. These models are able to give some insights about the disease progression through the analysis of patients histories and could be also used to predict the future course of the disease in an individual. In particular, Hidden Markov Models (HMMs) are a useful tool for disease modeling since they allow to model situations where the state of the disease is not observable, by giving the possibility to incorporate some priors and constraints. We applied our models to a simulated dataset by considering a generalization of HMMs with continuous time and multivariate outcome.

Abstract *I modelli di progressione di patologia sono un potente strumento per comprendere e prevedere lo sviluppo di una patologia, date delle misurazioni longitudinali ottenute da un campione di pazienti. Questi modelli sono in grado di arricchire la conoscenza sulla progressione della patologia attraverso l'analisi della storia dei pazienti e possono anche essere usati per predire il corso futuro della patologia di un individuo. In particolare, gli Hidden Markov Models (HMMs) sono un utile strumento per la modellazione di patologia in quanto permettono di costruire un modello in situazioni in cui lo stato della patologia non è osservabile, dando la possibilità di incorporare delle prior e dei vincoli. Abbiamo applicato i nostri modelli a un dataset simulato considerando una generalizzazione degli HMM a tempo continuo e risposta multivariata.*

Andrea Martino, Giuseppina Guatteri, Anna Maria Paganoni
Department of Mathematics, Politecnico di Milano, via Bonardi 9, 20133, Milan, Italy
e-mail: andrea.martino@polimi.it
e-mail: giuseppina.guatteri@polimi.it
e-mail: anna.paganoni@polimi.it

Andrea Ghiglietti
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123, Milan, Italy
e-mail: andrea.ghiglietti@unicatt.it

Key words: disease progression modeling, Hidden Markov Models, multivariate data

1 Introduction

Many chronic diseases can be naturally represented in terms of staged progression. Hidden Markov Models (HMMs) are a popular method for modeling disease progression and estimating the rates of transition between the stages of a disease. For this reason, we introduce a HMM for disease progression which takes into account the possibility of modeling multivariate observations with correlated components. Although discrete-time HMMs are often used to model disease progression, they are not very suitable in practice because the measurement data should be regularly sampled at discrete intervals and state transitions can only occur at these discrete times. Since we are interested in using our model to study the Heart Failure (HF) pathology and hospitalizations for HF patients occur irregularly in time, we consider a continuous time HMM, in which both the transitions between the hidden states and the observations can occur at arbitrary continuous times (for further details, see [1, 2]).

2 The model

A continuous time HMM is very suitable for modeling disease progression, which is a continuously evolving process. Even though the continuous time HMM adds more flexibility to the models with respect to the discrete time HMM, this comes with a higher computational cost. Indeed, in this case not only the hidden states are unobserved but the transition times are unknown too. Moreover, although HMMs often consider the case where the observations are sampled from a discrete distribution which can only take a finite number of values, in our analysis we have considered observations composed by both continuous and discrete values.

Let us denote $(\mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_K)$ the set of observations, univariate or multivariate data, observed at K irregularly-distributed continuous points in time $(t_1, t_2, \dots, t_K)$, such that we have two levels of hidden informations: the state of the Markov chain is hidden and the state transitions are also hidden, since it is not known if other transitions occur between two consecutive observations. For any observation $\mathbf{O}_k$ we denote the probability of being in a state $s(t_k)$ at time t_k, often called as *emission probability*, as $p(\mathbf{O}_k|s(t_k))$. As for continuous time Markov chains, we can define the finite and discrete state space S, the state transition rate matrix Q and the initial state probability distribution π. The elements q_{ij} in the matrix Q represent the rate of a process's transition from state i to state j for $i \neq j$, while the elements q_{ii} must be specified such that each row of the matrix sums up to zero ($q_i = \sum_{j \neq i} q_{ij}, q_{ii} = -q_i$) [1]. Moreover, if the process is time-homogeneous, the sojourn time in each state i

is exponentially distributed with parameter q_i, while q_{ij}/q_i indicates the probability of the next transition of the process from state i to state j.

As done in [3], if we consider a continuous time HMM which is fully observed, the complete joint likelihood of the data can be written as

$$CL = \prod_{k'=0}^{K'} q_{y_{k'},y_{k'+1}} e^{-q_{y_{k'}}\tau_{k'}} \prod_{k=0}^{K} p(\mathbf{O}_k|s(t_k)) = \prod_{i=1}^{|S|} \prod_{j=1,j\neq i}^{|S|} q_{ij}^{n_{ij}} e^{-q_i\tau_i} \prod_{k=0}^{K} p(\mathbf{O}_k|s(t_k))$$

where $(t'_0, t'_1, \dots, t'_{K'})$ are the K' state transition times with $Y' = \{y_0 = s(t'_0), \dots, y_{K'} = s(t'_{K'})\}$ being the corresponding states of the Markov chain, $\tau_k = t_{k+1} - t_k$ is the time interval between two observations while $\tau_{k'} = t'_{k'+1} - t'_{k'}$ is the time interval between two transitions and n_{ij} is the number of transitions from state i to state j.

3 Parameter estimation

For the computation of the rate transition matrix Q, several methods based on EM algorithms have been proposed in [3]. We now focus on the estimation of the emission probability matrix $B = \{b_j(\mathbf{O}_k)\}$ where $b_j(\mathbf{O}_k) = p(\mathbf{O}_k|s(t_k) = j)$ is the probability of being in a state j at time t_k, while observing $\mathbf{O}_k$. This estimation is usually straightforward if we are considering an univariate outcome. In general, since we consider multivariate observations, we also want to model the correlation among the variables.

As usually done, if the observations are only symbols chosen from a finite alphabet, a discrete probability density can be used to model the data and estimate the matrix B. This approach is not generally enough since we consider observations coming from both discrete and continuous distributions. In order to use a continuous observation density, we have to consider some restrictions for the estimation of the probability density function (pdf). We can represent the pdf in its most general representation as a finite mixture which can be written as

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm}(\mathbf{O}) \mathscr{D}[\mathbf{O}, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}], \quad 1 \leq j \leq N$$

where $\mathbf{O}$ is the vector being modeled, N is the number of statistical units, $c_{jm}(\mathbf{O})$ is a non negative mixture coefficient for the mth mixture in state j and $\mathscr{D}$ is a pdf with mean vector $\boldsymbol{\mu}_{jm}$ and covariance matrix $\mathbf{U}_{jm}$ for the mth mixture component in state j. We implemented a normalized version of the forward-backward algorithm and a Baum-Welch algorithm (see [5]) by using our matrix B in a continuous time framework. All the analysis have been carried out using the software R ([4]).

4 Simulation study

We applied our model to multivariate longitudinal data, which are repeated observations of multiple response variables. Since the data are correlated over time and multiple responses are measured at the same time, special treatments are required to analyze the data. In particular, the easiest approach would be to ignore the correlation, which would lead to some loss of information. Therefore, to flexibly characterize the distribution of the emission probabilities, we modelled the correlation among the observation components for each multivariate outcome.

We generated a sample $(x_1,y_1),\dots,(x_N,y_N)$ of $N = 1000$ observations for $n = 50$ statistical units, in order to have 20 observations for each statistical unit. For each one of them, we have a sequence of pairs which is the realization of a 3-state Markov process. Given the state j of the Markov process, each pair of the sample is a realization of the joint distribution (X,Y), where $X \sim Be(p)$ and $Y = XY_1 + (1-X)Y_2$, with $Y_i \sim N(\mu_{ij},\sigma_{ij}^2)$ independent of X. We used the following parameters to generate the data:

- **State 1**: $p = 0.2, \mu_{11} = 0, \mu_{21} = 3, \sigma_{11} = 0.5, \sigma_{21} = 0.8.$
- **State 2**: $p = 0.9, \mu_{12} = 1, \mu_{22} = 5, \sigma_{12} = 0.5, \sigma_{22} = 0.8.$
- **State 3**: $p = 0.7, \mu_{13} = 4, \mu_{23} = 7, \sigma_{13} = 0.5, \sigma_{23} = 0.8.$

We applied our algorithm using $m = 2,\dots,5$ states and we can see the results we obtained in Table 1. A problem which naturally arises is that of selecting an appropriate model, e.g. of choosing the appropriate number of states for the HMM, so we need some criteria for model comparison. To address this problem, for each run of the algorithm we computed the Aikake Information Criterion as $\text{AIC} = -2\log L + 2p$ and the Bayesian Information Criterion as $\text{BIC} = -2\log L + plogT$, where L is the likelihood function of the fitted model, p is the number of unknown parameters and T is the number of observations. The values we obtained are showed in Fig. 1. According to both AIC and BIC, the model with three states is the most appropriate. As we can see for $m = 3$, the estimated values are very similar to the real ones, so we can conclude that by considering the correlation among the components of the outcome variables, we were able to obtain very good results.

5 Conclusion

In this work we built a model for longitudinal data with multivariate observations and showed that, if we consider the correlation among the components of the outcome, we can obtain very good results. The next step will consist in applying our algorithm to a real case study, with the data coming from an administrative dataset about hospitalizations which is in pre-processing, in order to study the progression of the Heart Failure pathology.

$m=2$	p	$\mu_{1\cdot}$	$\mu_{2\cdot}$	$\sigma_{1\cdot}$	$\sigma_{2\cdot}$
State 1	0.4834	0.8298	3.1500	1.0470	1.2354
State 2	0.6634	3.9745	6.9547	0.4628	0.7692
$m=3$	p	$\mu_{1\cdot}$	$\mu_{2\cdot}$	$\sigma_{1\cdot}$	$\sigma_{2\cdot}$
State 1	0.1981	0.0421	2.9863	0.5422	0.8521
State 2	0.9026	1.0127	4.9409	0.5335	0.7823
State 3	0.7067	3.9732	6.9780	0.4727	0.8324
$m=4$	p	$\mu_{1\cdot}$	$\mu_{2\cdot}$	$\sigma_{1\cdot}$	$\sigma_{2\cdot}$
State 1	0.3003	0.0460	2.9873	0.4157	0.7113
State 2	0.1507	1.0128	4.8392	0.4067	0.6854
State 3	0.9120	3.7699	6.7472	0.5332	0.7222
State 4	0.6694	4.3680	7.2856	0.4727	0.7341
$m=5$	p	$\mu_{1\cdot}$	$\mu_{2\cdot}$	$\sigma_{1\cdot}$	$\sigma_{2\cdot}$
State 1	0.1431	0.1124	2.6973	0.4131	0.7034
State 2	0.2937	0.7995	2.9427	0.2844	0.5854
State 3	0.3110	1.0190	4.3863	0.7944	0.9653
State 4	0.9143	3.8483	6.9001	0.5287	0.7712
State 5	0.6684	4.3742	7.2486	0.4728	0.7338

Table 1: Results of a HMM with $m = 2, \ldots, 5$ states.

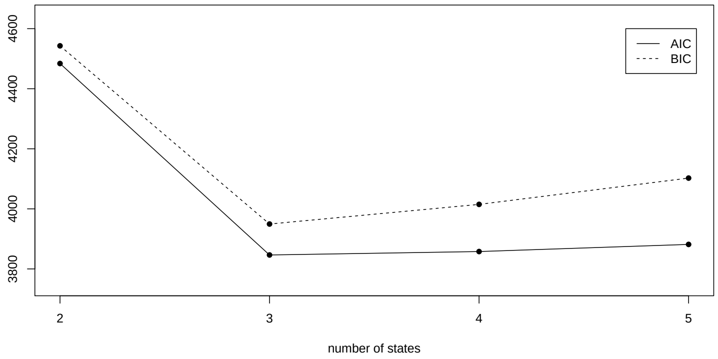


Fig. 1: Model selection criteria using AIC and BIC

References

1. Cox DR, Miller HD. The Theory of Stochastic Processes. Chapman and Hall; London: 1965.
2. Jackson CH. Multi-state models for panel data: the msm package for R. Journal of Statistical Software. 2011;38 (no. 8)
3. Y.Y. Liu, S. Li, F. Li, L. Song, J.M. Rehg, Efficient Learning of Continuous-Time Hidden Markov Models for Disease Progression, Advances in Neural Information Processing Systems, 3599-3607
4. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
5. L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE, **77**, 257–285 (1989)

A simulation study on the use of response-adaptive randomized designs

Uno studio di simulazione sull'uso di disegni casuali adattivi alla risposta

Anna Maria Paganoni, Andrea Ghiglietti, Maria Giovanna Scarale, Rosalba Miceli, Francesca Ieva, Luigi Mariani, Cecilia Gavazzi and Valeria Edefonti

Abstract Response-adaptive designs have been proposed in randomized clinical trials to achieve ethical advantages by using sequential accrual information collected during the trial to update probabilities of treatment assignments. We propose the use of a response adaptive design based on urn models in a simulation study on a randomized clinical trial on the efficacy of home enteral nutrition in cancer patients after major gastrointestinal surgery. We compare results with the adaptive design with those previously obtained with the non-adaptive approach.

Abstract *I disegni adattivi alla risposta sono stati proposti nell'ambito degli studi clinici per ottenere vantaggi etici utilizzando le informazioni sequenziali raccolte durante lo studio per aggiornare le probabilità di assegnazione ai trattamenti. In questo lavoro si propone l'uso di un disegno adattivo di risposta basato su modelli d'urna in uno studio di simulazione basato su uno studio clinico dell'efficacia della nutrizione enterale domiciliare in pazienti oncologici dopo un intervento chirurgico*

Anna Maria Paganoni, Francesca Ieva
Politecnico di Milano, Milano (Italy)
e-mail: anna.paganoni@polimi.it; e-mail: francesca.ieva@polimi.it

Andrea Ghiglietti
Università Cattolica del Sacro Cuore, Milano (Italy)
e-mail: andrea.ghiglietti@unicatt.it

Maria Giovanna Scarale
IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo (Italy)
e-mail: mgiovanna.scarale@operapadrepio.it

Rosalba Miceli, Luigi Mariani, Cecilia Gavazzi
Fondazione IRCCS Istituto Nazionale Tumori, Milano (Italy)
e-mail: Rosalba.Miceli@istitutotumori.mi.it; e-mail: luigi.mariani@istitutotumori.mi.it;
e-mail: Cecilia.Gavazzi@istitutotumori.mi.it

Valeria Edefonti
Università degli Studi di Milano, Milano (Italy)
e-mail: valeria.edefonti@unimi.it

gastrointestinale maggiore. Si confrontano i risultati del disegno adattivo con quelli precedentemente ottenuti con l'approccio non adattivo.

Key words: Randomly Reinforced Urn model, Randomized trials, Response-adaptive randomization, Simulation study

1 Introduction

In statistical literature, urn models have been widely studied as mathematical tools to implement randomization in the context of clinical trials. These designs randomly assign those subjects that sequentially enter the trial to the treatment arms according to to the proportion of balls of different color sampled from a virtual urn. Recently, interest has been increased in the use of urn models, in which the probability to sample a ball of a certain type depends on the treatment performance observed on the subjects previously randomized [2, 1]. A popular class of such designs is the Randomly Reinforced Urn (RRU) model, which has been introduced in [2] for binary treatment responses and extended in [5] to handle continuous responses. In the RRU model, an urn containing balls of two colors is sequentially sampled and the subjects in the trial receive the treatments associated to the colors of the sampled balls. In addition, the urn composition is sequentially updated by adding new balls of the same color of the sampled ones, whose number depend each time on the response observed by the correspondent patient. For the purposes of this paper, we simply remind that a RRU design assigns patients to the superior treatment with a probability that converges to one as the sample size increases. Although the theoretical result of assigning most of the patients to the superior treatment is very attractive from the ethical point of view, the RRU design have rarely been implemented in clinical trials or in simulation studies based on a real set-up. In detail, we will simulate a large number of trials that follow the RRU model starting from the real-life data collected in a previously published Home Enteral Nutrition (HEN) randomized trial [3], where a non-adaptive design was originally adopted (see [4]). Comparing the performance of the RRU with that of the original non-adaptive design, we expect that the RRU design will: 1) assign fewer patients to the inferior treatment; 2) maintain similar inferential properties. This will turn out in an advantage in terms of both statistical performance and ethical responsibility.

2 Materials and Methods

The RRU model was here implemented in a simulation study based on results from a multicenter, controlled, open-label, two-parallel groups, randomized clinical trial conducted at the Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy, and at the European Institute of Oncology, Milan, Italy, between December

2008 and June 2011 [3]. The enrolled subjects were adult ($>$ 18 years) patients with documented upper gastrointestinal cancer who were candidates for major elective surgery and showed a preoperative nutritional risk score that indicated a potential benefit from any nutritional intervention. A random permuted block design (stratified for referring center) randomly assigned patients before discharge to receive either HEN to cover the basal energy requirement (experimental group), or nutritional counselling by an expert dietitian, including oral supplements only when needed (Control Group - CG), in a 1:1 ratio. The treatment effect was defined as the difference between the mean "weight change" (weight after two months - weight at baseline) in the HEN and nutritional counselling arms (primary end-point). In total, 79 patients were initially randomized; however, as 11 patients had a missing two-months weight, the final analysis was performed on 68 patients, of which 33 patients were allocated to the HEN group and 35 to the CG. The main result of the primary end-point analysis was that the mean weight loss in the patients undertaking the HEN treatment was significantly lower than that in the CG, with a treatment effect estimated by the corresponding ANOVA model coefficient (95% confidence interval) of 3.2 (1.1-5.3) and a p-value from the corresponding two-sided t-test equal to $0.31\% < 2.94\%$. For this reason, the trial was stopped at the interim analysis and results from this analysis were published in [3]. So, the HEN was found to be the superior treatment in this trial.

To simulate the RRU design starting from the HEN trial data to derive the results for comparing the RRU design with the non-adaptive one, we performed the following main steps:

(A)using the HEN trial dataset [3]:

  (1)we estimated the parameters of the Gaussian distribution of the responses to the HEN group;
  (2)we estimated the parameters of the Gaussian distribution of the responses in the CG;
  (3)we computed the empirical distribution of the difference between arrival times of consecutive subjects;

(B)we simulated N independent trial samples based on the RRU model; for each sample, responses to both treatments and intervals between arrival times were randomly generated from distributions introduced in point (A);
(C)we computed from these N trials:

  (1)the empirical distribution of the number of subjects assigned to the inferior treatment $\mathscr{W}$;
  (2)the empirical power of the corresponding t-test.

The previous steps are detailed in the following.
To start, we considered three alternative set-ups of trial sample sizes equal to (a): $n = 58$;(b): $n = 68$; (c): $n = 78$, where the total sample size 68 of the HEN trial was used as the reference set-up and we moved $\pm 15\%$ from that to get other two reasonable sample sizes.
For each set-up, we performed $N = 10,000$ simulations of independent trials based

on the RRU design: in each run we have a virtual urn to be sampled and reinforced. Formally, we denote by (R_i^j, W_i^j) the urn composition and by $R_i^j/(R_i^j + W_i^j)$ the urn proportion in simulation $j = \{1,..,N\}$ at time $i \in \{1,..,n\}$.
All the urns start with the same (fixed) initial composition, i.e. $(R_0^j, W_0^j) = (R_0, W_0)$ for any $j = \{1,..,N\}$. Then, the urn composition (R_i^j, W_i^j) is updated in the following way

$$\begin{cases} R_i^j = R_{i-1}^j + \sum_{k \in (A_{i+1}^j \setminus A_i^j)} X_k^j \xi_{Rk}^j \\ \\ W_i^j = W_{i-1}^j + \sum_{k \in (A_{i+1}^j \setminus A_i^j)} (1 - X_k^j) \xi_{Wk}^j, \end{cases}$$

where X_k^j is a Bernoulli random variable with parameter $R_{k-1}^j/(R_{k-1}^j + W_{k-1}^j)$ indicating the treatment assignement, the set A_i^j here includes all the patients who arrived two months earlier than subject i and ξ_{Rk}^j, ξ_{Wk}^j are the subjects responses to treatment R and W respectively. Indeed, in the HEN trial, responses were available only two months after treatment administration.

In addition, as normality assumptions in the original data were not rejected, responses to both treatments were generated as independent Gaussian random variables with arm-specific means and variances computed using the HEN dataset and given by: $m_R = -0.315$ and $\sigma_R = 3.868$ for treatment $\mathscr{R}$ (HEN group), $m_W = -3.571$ and $\sigma_W = 4.789$ for treatment $\mathscr{W}$ (CG). Formally, we generated the following quantities:

(1)$\xi_{R1}^j, .., \xi_{Rn}^j \sim \mathscr{N}(m_R, \sigma_R^2)$ potential responses to treatment $\mathscr{R}$ (HEN group);
(2)$\xi_{W1}^j, .., \xi_{Wn}^j \sim \mathscr{N}(m_W, \sigma_W^2)$ potential responses to treatment $\mathscr{W}$ (CG),

where either ξ_{Ri}^j or ξ_{Wi}^j is observed, as each subject just receives one treatment.

3 Results

Table 1 shows some descriptive statistics of the empirical distribution of the number of subjects assigned to the inferior treatment, $N_W(n)$, and the empirical power of the t-test, $1 - \hat{\beta}$, for the different sample sizes n, in comparison with the corresponding results for the non-adaptive design, n_W and $1 - \beta$.

For all sample sizes under consideration [cases (a)-(b)-(c)], the mean and the median of $N_W(n)$ were smaller than n_W, the number of subjects assigned to the inferior treatment by the non-adaptive design. It follows that the RRU design provided 50% of probability (or more) to assign fewer subjects to the inferior treatment, as compared to the non-adaptive design. Although higher than n_W for all the sample sizes considered, the third quartile of $N_W(n)$ in the RRU design was very close to n_W for any n under consideration. In addition, the obtained values for the t-test empirical power under the RRU design were close, but slightly smaller than, the corresponding power values derived in the non-adaptive design.

Further information on the distribution of $N_W(n)$ is provided by the boxplots re-

n	$N_W(n)$				n_W	$1-\beta$	$1-\hat{\beta}$
	1^{st} quartile	Mean	Median	3^{rd} quartile			
(a) 58	19	25.6	25	31	29	0.88	0.83
(b) 68	**22**	**29.6**	**29**	**36**	**35**	**0.92**	**0.88**
(c) 78	25	33.6	33	41	38	0.94	0.92

Table 1 Summary statistics (1^{st} and 3^{rd} quartiles, mean, and median) of the empirical distribution of the number of subjects assigned to the inferior treatment, $N_W(n)$, and empirical power, $1-\hat{\beta}$, of the t-test for equal mean weight changes (corresponding to the treatment coefficient in the ANOVA model) for the different sample sizes n in the Randomly Reinforced Urn design, in comparison with the corresponding results for the non-adaptive design, n_W and $1-\beta$. We reported in bold typeface the results obtained with the same sample size of the original Home Enteral Nutrition trial.

ported in Figure 1. For any sample size, the median of $N_W(n)$ was below the dashed line indicating the number of subjects assigned to the inferior treatment by the non-adaptive design. Similarly, we confirmed that, although higher, the third quartile was closer than the median to the dashed line for the three cases under consideration. In addition, the probability that $N_W(n)$ was less than n_W was close to 75% for any sample size under consideration. Finally, although mostly symmetric, the empirical distributions of the number of subjects assigned to the inferior treatment showed a high level of variability. This variability increases, as the total sample size increases.

References

1. Atkinson, A.C., Biswas, A., 2014. Randomised response-adaptive designs in clinical trials. Chapman and Hall/CRC.
2. Durham, S.C., Flournoy, N. and Li, W., 1998. A sequential design for maximizing the probability of a response. The Canad. J. Stat., 26 (3), 479-495.
3. Gavazzi, C., Colatruglio, S., Valoriani, F., Mazzaferro, V., Sabbatini, A., Biffi, R., Mariani, L., Miceli, R. 2016. Impact of home enteral nutrition in malnourished patients with upper gastrointestinal cancer: a multicenter randomized clinical trial. Europ. J. Cancer, 64, 107-112.
4. Ghiglietti, A., Scarale, M.G., Miceli, R., Ieva, F., Mariani, L., Gavazzi, C., Paganoni, A.M., Edefonti, V. 2018. Urn models for response-adaptive randomized designs: a simulation study based on a non-adaptive randomized trial *J Biopharm Stat*, In press.
5. Muliere, P., Paganoni, A.M., Secchi, P., 2006. A randomly reinforced urn. J. Stat. Plann. Inference, 136, 1853-1874.
6. Wei, L.J. 1978. An application of an urn model to the design of sequential controlled clinical trials. JASA, 73, 559-563.

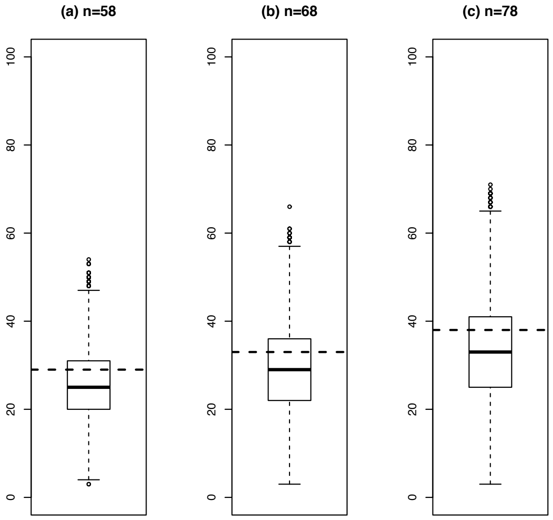


Fig. 1 Boxplots of the number of subjects assigned to the inferior treatment (Control Group) in the three cases reported above each picture: (a) $n = 58$, (b) $n = 68$, (c) $n = 78$. The dashed line indicated the number of subjects assigned to the control group in the non-adaptive trial in the three cases.

The relationship between health care expenditures and time to death: focus on myocardial infarction patients

La relazione tra spesa assistenziale e prossimità al decesso: studio dei pazienti affetti da infarto al miocardio

Luca Grassetti and Laura Rizzi

Abstract This study focuses on the relationship between health care expenditures (HCE) and time to death (TTD). This is a central theme of regional and national healthcare systems, given the current ageing of the population and the increasing expenditures. This study is aimed to investigate in-depth the causal-effect relationship between TTD and HCE. Existing regional health administrative archives have been used to build a cohort of patients with new-onset myocardial infarctions observed between 2003 and 2007. All patients are residents in the Italian region Friuli Venezia Giulia. Data on individual HCEs, socio-demographic and health status characteristics are included in the analysis, together with TTD. The econometric analysis is based on a sample selection approach necessary to account for the inflation patterns in the distributions of HCE and TTD. The main results suggest that the causal effect of TTD on HCEs ("red herring" hypothesis) is confirmed, while the reversal relationship is only partially revealed in the selection equation.

Abstract *Il presente studio analizza la relazione tra spese assistenziali e prossimità al decesso. Questo è un ambito di interesse primario dei sistemi sanitari pubblici regionali e nazionali che si trovano ad affrontare il fenomeno dell'invecchiamento della popolazione e dei trend crescenti delle spese assistenziale. L'obiettivo principale perseguito è lo studio della relazione biunivoca tra spese assistenziali e prossimità al decesso. L'analisi si basa su una coorte di pazienti affetti da infarto al miocardio, residenti nella regione italiana del Friuli Venezia Giulia, osservati nel periodo 2003-2007. La costruzione del database è derivata dai database amministrativi sanitari regionali. Il database include i dati individuali relativi alle spese sanitarie, alle caratteristiche socio-demografiche, alle proxy dello stato di salute e alla prossimità al decesso. Per l'analisi econometrica è stato adottato un approccio*

Luca Grassetti
Dept. of Economics and Statistics - University of Udine, Via Tomadini, 30/a Udine (Italy) e-mail: luca.grassetti@uniud.it

Laura Rizzi
Dept. of Economics and Statistics - University of Udine, Via Tomadini, 30/a Udine (Italy) e-mail: laura.rizzi@uniud.it

di sample selection, data la presenza di punti di massa nelle distribuzioni di probabilità delle spese sanitarie e del tempo al decesso. I risultati confermano il rilevante effetto della prossimità al decesso sulle spese assistenziali (nota come ipotesi "red herring"), mentre la relazione di causalità inversa risulta confermata solo in relazione al modello di selezione.

Key words: Administrative Data, Health Care Expenditures, Health Econometrics, Sample Selection Models, Time To Death

1 Aim and context of analysis

The aim and the context of this study are tied to central themes of modern health care policy (*Health, demographic change, and wellbeing is one of the societal challenges in the European programme Horizon 2020*) such as increasing trends of health care expenditures (HCE); the role of time to death (TTD) and the ageing of the population on HCE patterns; and the increasing burden of healthcare profiles due to medical innovations. In particular, we focus on a relevant challenge: the relationship between TTD and HCE, focusing on whether TTD affects HCE (TTD as a determinant of budget growth) or if the reciprocal relationship holds (care profiles affecting health outcomes).

The health economics literature presents different approaches to HCE: some studies consider HCE dynamics within the macro-econometric framework (see, for instance, [11]), while a wide range of works is devoted to assessing the determinants of HCE both at micro and macro-level (a review on the topic can be found in [6] and [12]). These kinds of studies focus on the analysis of the relationship between diseases and health care utilization, or on the distribution of HCE across individuals and cohorts. Moreover, HCEs are often considered in the analysis of the relationship between the ageing of the population and TTD, which has driven the *Red Herring* literature on the significant effect of TTD on the increase of HCEs ([4], [5], [8], [13] and [15] are some examples of these works). The literature devoted to the assessment of TTD determinants is mainly characterized by epidemiological (survival) analyses, while studies on the effect of HCE on TTD are rare (see [4] again). Given the differences in these existing research approaches, the following competing hypotheses are evaluated in this study: compression of morbidity (TTD causal effect hypothesis) according to which health care expenditures are determined by proximity to death; expansion-of-morbidity hypothesis based on the assumption that prolonging life means prolonging morbidity and increasing costs. We wish to explore both given hypotheses separately for different sources of HCE, by considering administrative datasets and trying to measure TTD more accurately.

The phenomena of interest of this study present the so-called inflation issue. In particular, HCEs are often zero-inflated measures, while the TTD presents a double inflated distribution (determined by sudden deaths and by survivals). In order to deal with these issues, the econometric literature has developed some "alternative" approaches (in particular, for the HCEs analysis): latent class models ([3]); sample selection models ([7]); two-part models ([1], [2] and [9]); copula probit models (as in [14]). Finally, to deal with TTD double inflated distribution it is possible to apply the solution proposed in [10].

2 Study design

Many empirical analyses on the relationship between HCE and TTD are classified as population-based studies. A specifically structured administrative database has been developed for this research. In particular, we consider a cohort type dataset of all the patients with an incident case of acute myocardial infarctions (AMI) observed in the period 2003-2007 in Friuli Venezia Giulia (Italy).

2.1 Cohort and variables

To determine the final cohort, four steps have been followed: first 11530 patients are identified selecting all the hospital admissions with a diagnosis of AMI (ICD-9 code 410*) in the Hospital Discharge Register in the period 2003-2007; within the selected cases 10700 patients are admitted with incident AMI; the size of the dataset is reduced to 9962 to consider patients who are resident in FVG at admission date. In order to deal with homogeneous patients, cases with the first intervention of AMI (no previous PTCA[1] or bypass surgery) are selected, thus identifying 9897 cases. The HCE and TTD of these patients are observed in a 5 years period (follow-up); HCEs (regarding inpatient cares, outpatient cares, and drugs) are measured by semesters while TTD is computed in days and truncated at 1825 for the survivors at the end of the 5th year.

The data are derived from the regional health data-warehouse. Data regarding HCEs are collected also for the two years before the entering event. An important role in the model specification is played by the TTD covariate definition and, given the peculiar double truncated TTD variable distribution, we considered both a factor variable identifying: sudden deaths (SD), patients died during the observational period (DDOP) and survivors (SURV), and a numerical variable (*days to death*), interacting with DDOP dummy only. The age and gender of the patients are collected, together with proxies of individual health status as dummies for pathology exemptions, dummies for PTCA or bypass surgery and a dummy for deceased. Finally, information on the area of residence (such as local health agency - LHA - and municipality) and the deprivation index (at the municipality level) are considered.

The cohort is characterized by 5766 males (58.3% of the total) with 68.2% of survivors. 45.5% of patients are more than 75 years old at the entering time and the

[1] Percutaneous transluminal coronary angioplasty

percentage of survivors is 97.8% for patients aged less than 45 decreasing to 36.6% in the oldest age class. The proportion of survivors decreases also by deprivation class (from 63.3% in the “very rich” to 56.9% in the “very deprived” class). Median values of all types of individual HCE are lower in the group of 6151 survivors and this is particularly clear for the median values of drugs expenditures (1010.9 and 1185.8 for dead patients, while 543.4 for the survivors).

3 HCE and TTD models and results

In the following sections, the HCE model is introduced and the main results on the determinants of health care expenditures are briefly discussed. The TTD model is then introduced and the main results, reported in Table 1, are discussed.

3.1 The HCE model

In the analyses, we adopted the sample selection paradigm. In particular, the Heckman-style selection model is applied separately to three kinds of health expenditures. The model for the HCEs can be defined as:

$$y_i^O = \begin{cases} 0 & \text{if } y_i^S = 0 \\ y_i^{O*} = X_i^O \beta^O + \varepsilon_i^O & \text{otherwise} \end{cases} \tag{1}$$

where y_i^O denotes the observed level of health expenditure, y_i^S is the variable defining the selection process (which can be estimated by considering a probit generalised linear model - selection equation) and y_i^{O*} represents the observed level of positive expenditures, modeled as a linear function of the explanatory variables X_i^O.

The results of this first step can be summarised as follows. The selection equations in the models for health expenditures present a significant relationship between age and outpatient and drug expenditures, while for inpatient care expenditures the “red herring” hypothesis is fully supported. Exemptions have positive effects on the probability of positive HCEs for all kinds of expenditures. The model estimation is developed separately for the different kinds of the health expenditures.

Going into greater details, males have a higher probability of HCE for inpatient care and being a male affects positively the expenditure for outpatient services. The age over TTD effect on the probability of HCE is clear for all age classes over 65 years, with an increasing effect on the amount of outpatient and drug expenditures. Survivors generally show a low probability of inpatient care and a larger probability of outpatient care. The same pattern is pointed out in expenditure model. Proxies of health conditions, represented by pathology exemptions, are, as expected, positively correlated with the probability of all types of expenditures but with greater effects on the amount of outpatient and drug demand. Moreover, strong positive relationships between types of HCEs are revealed, as expected. The HCE models point out further expected results in terms of factors affecting both the HCE likelihood and its amount. However, the TTD effect is not relevant as it is probably absorbed by the dummies included to account for sudden deaths and survivors, whose coefficients confirm the “red herring” proposition.

3.2 The TTD model

The Heckman-style selection model has also been applied in the estimation of TTD. The distribution of TTD presents two mass points generated by sudden deaths (0 days) and 5 years survivals (1825 days). For this reason, the selection model has been specified by considering an ordinal probit model (as in [10]). The bias-corrected linear model has been estimated by considering the specific inverse Mills ratios (for the ordinal model) on the patients who died during the observational period. In order to obtain better estimation results, we considered a Box-Cox transformation of the days to death.

Table 1 shows the TTD model estimation results. TTD is only partially connected to health expenditures. In particular, in the selection equation, all the coefficients of expenditure sources are significant. The HCEs affect survival status in the selection step. Notwithstanding, in the outcome equation only the logarithm of inpatient care expenditure is significant. The significance of the coefficient of the inverse Mills ratio confirms the necessity to adopt the sample selection model.

Selection Model		Estimate	p-value
Gender	*Female*	-	
	Male	-0.1770	< 0.001
Age classes	*Less than 45*	-	
	(45,55]	-0.6124	0.001
	(55,65]	-0.8933	< 0.001
	(65,75]	-1.0236	< 0.001
	More than 75	-1.6459	< 0.001
	Intervention Dummy	0.7906	< 0.001
	Bypass Dummy	-0.1367	0.010
LHA	*101*	-	
	102	0.0333	0.479
	103	0.0424	0.531
	104	0.0992	0.011
	105	-0.0030	0.952
	106	0.1096	0.008
	Emergency Dummy	-0.0729	0.084
Exemptions	*Cardiology*	0.7681	< 0.001
	Invalidity	-0.1563	< 0.001
	Cholesterol	0.2332	< 0.001
Log-HCEs	*Inpatient*	-0.0338	< 0.001
	Drugs	-0.0335	< 0.001
	Outpatient	0.0314	< 0.001
Thresholds	z_1	-2.5915	< 0.001
	z_2	-1.2976	< 0.001

Outcome Model		Estimate	p-value
	Intercept	8.4914	0.141
Gender	*Female*	-	
	Male	1.3225	0.014
Age classes	*Less than 45*	-	
	(45,55]	4.3047	0.448
	(55,65]	8.9082	0.105
	(65,75]	11.6086	0.035
	More than 75	15.2522	0.007
	Intervention Dummy	-2.3245	0.024
LHA	*101*	-	
	102	-1.4371	0.093
	103	-2.3241	0.073
	104	-2.6228	< 0.001
	105	-2.3982	0.009
	106	-2.3000	0.003
Exemption	*Invalidity*	4.0841	< 0.001
Log-HCEs	*Inpatient*	0.1516	0.015
	InvMillsRatio	9.2201	< 0.001
R^2	0.0627		

Table 1 The results of the TTD model estimation. The selection ordinal logit model results are reported in the left panel. The right panel shows the results of the outcome model.

4 Conclusions

This study has assessed the role of TTD on HCE and has confirmed, in general, the "red herring" hypothesis. Moreover, this hypothesis has not always been confirmed in our framework. The role of HCE has been studied to verify the simultaneity between HCE and TTD. This last analysis involved a non-standard econometric approach and it has been developed by using a suitable measure of TTD (presenting a double inflated distribution). The adoption of ordinal selectivity in the analysis of the role of HCE on TTD allows us to study the peculiar behaviour of the phenomenon. The results of the empirical analyses show that the role of HCE on TTD (reversal relationship) is significant, in particular for the selection equation.

References

1. Beeuwkes Buntin, M., Zaslavsky, A.M.: Too much ado about two-part models and transformation?: Comparing methods of modeling Medicare expenditures, J Health Econ, 23 (3), 525–542 (2004)
2. Deb, P., Munkin, M. K. and Trivedi, P. K.: Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. J Applied Economet, 21 (7), 1081–1099 (2006)
3. Deb, P. and Trivedi, P.K.: The structure of demand for health care: latent class versus two-part models, J Health Econ, 21 (4), 601–625 (2002)
4. Felder, S., Werblow, A. and Zweifel P.: Do red herrings swim in circles? Controlling for the endogeneity of time to death. J Health Econ 29 (2), 205–212 (2010)
5. Geue, C., Briggs, A., Lewsey, J., and Lorgelly, P.: Population ageing and healthcare expenditure projections: new evidence from a time to death approach. Eur J Health Econ, 15 (8), 885–896 (2013)
6. Koopmanschap, M., de Meijer, C., Wouterse, B., and Polder, J.: Determinants of health care expenditure in an aging society. Panel Paper, 22 (2010)
7. Madden, D.: Sample selection versus two-part models revisited: The case of female smoking and drinking, J Health Econ, 27 (2), 300–307 (2008)
8. Moorin, R., Gibson, D. and Hendrie, D.: The contribution of age and time-to-death on health care expenditure for out-of-hospital services. J Health Serv Res Po 17 (4), 197–205 (2012)
9. Mullahy, J.: Much ado about two: reconsidering retransformation and the two-part model in health econometrics, J Health Econ, 17 (3), 247–281 (1998)
10. Terza, J.: A Two Stage Estimator for Models with Ordinal Selectivity, Proceedings of the Business and Economics Section of the American Statistical Association, 484–486 (1983)
11. van Baal, P.H., and Wong, A.: Time to death and the forecasting of macro-level health care expenditures: Some further considerations, J Health Econ, 31 (6), 876–887 (2012)
12. Wang, Z.: The determinants of health expenditures: evidence from US state-level data, Appl Econ, 41 (4), 429–435 (2009)
13. Werblow, A., Felder, S., and Zweifel, P.: Population ageing and health care expenditure: a school of 'red herrings'?, Health Econ, 16 (10), 1109–1126 (2007)
14. Winkelmann R.: Copula bivariate probit model: with an application to medical expenditures, Health Econ, 21 (12), 1444–1455 (2012)
15. Wong, A., van Baal, P. H. M., Boshuizen, H. C. and Polder, J.J.: Exploring the influence of proximity to death on disease-specific hospital expenditures: a carpaccio of red herrings, Health Econ, 20 (4), 379–400 (2011)

A multivariate extension of the joint models

Un'estensione multivariata dei modelli congiunti

Marcella Mazzoleni and Mariangela Zenga

Abstract The joint models analyse the effect of longitudinal covariates onto the risk of an event. They are composed of two sub-models, the longitudinal and the survival sub-model. For the longitudinal sub-model a multivariate mixed model can be proposed. Whereas for the survival sub-model, a Cox proportional hazards model is proposed, considering jointly the influence of more than one longitudinal covariate onto the risk of the event. The purpose of the work is to extend an estimation method based on a joint likelihood formulation to the case in which the longitudinal sub-model is multivariate through the implementation of an Expectation-Maximisation (EM) algorithm.

Abstract *I modelli congiunti analizzano l'effetto delle covariate longitudinali sul rischio di un evento. Sono composti da due sotto-modelli, quello longitudinale e quello di sopravvivenza. Per il sotto-modello longitudinale si puó proporre un modello misto multivariato, mentre per quello di sopravvivenza viene proposto un modello a rischi proporzionali di Cox, dove le covariate longitudinali influenzano congiuntamente il rischio dell'evento. Lo scopo del lavoro é di estendere un metodo di stima basato sulla massimizzazione della verosimiglianza congiunta al caso in cui il sotto-modello longitudinale è multivariato attraverso l'implementazione di un algoritmo Expectation-Maximization (EM).*

Key words: Joint models, Multivariate Mixed Model, EM Algorithm, Joint Likelihood

Marcella Mazzoleni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy e-mail: marcella.mazzoleni@unimib.it

Mariangela Zenga
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy e-mail: mariangela.zenga@unimib.it

1 Introduction

The joint models analyse the effect of longitudinal covariates onto the risk of an event. They are composed of two sub-models, the longitudinal and the survival sub-model. For the longitudinal sub-model a multivariate mixed model can be proposed, considering fixed and random effects. Whereas for the survival sub-model, a Cox proportional hazards model is usually proposed, considering jointly the influence of more than one longitudinal covariate onto the risk of the event.

The joint models are often used in medical research because in clinical trails the aim is to analyse two subgroups, placebo and treated, in order to study the longitudinal covariates that could influence the survival time.

The first authors that extended the joint model to the case in which the longitudinal sub-model is multivariate are Xu and Zeger [10]. A Markov chain Monte Carlo algorithm was used to estimate parameters in the model, extending the univariate estimation method introduced by Xu and Zeger [11] and Faucett and Thomas [2]. The authors applied the model and the estimation method to the Schizophrenia trial data of risperidone. Albert and Shih [1] proposed a regression calibration approach for jointly modelling multiple longitudinal measurements and discrete time-to-event data. The authors proposed a two-stage regression calibration approach. Recently Hickey et al. [5] proposed a interesting review of all the model and estimation methods for the joint modelling of time-to-event and multivariate longitudinal outcomes. Despite developments, software to estimate the parameters of these model is still lacking. For this reason, Hickey et al. [4] implemented a new package in software R, the joineRML package. This package fits the joint model proposed by Henderson et al. [3], extended to the case of multiple continuous longitudinal measures. The association between time-to-event and longitudinal data is captured by a multivariate latent Gaussian process. The parameter are estimated using a Monte Carlo Expectation Maximization algorithm.

The purpose of the paper is to extend an estimation method based on a joint likelihood formulation used in the univariate case [6] to the case in which the longitudinal sub-model is multivariate. The parameters are estimated maximising the likelihood function, using an Expectation-Maximisation (EM) algorithm. In addition, in the M-step a one-step Newton-Raphson update is used, as for some parameters estimators, it is not possible to obtain closed-form expression. In addition, a Gauss-Hermite approximation is applied for some of the integrals involved.

2 Model and estimation method

The longitudinal and the survival sub-models compose the joint models. Concerning the survival sub-model, in this paper a proportional hazard model is used, which is defined as a function of the $m_{iq}(t)$ that denotes the true and unobserved value of the longitudinal covariate q for subject i:

$$h_i(t|M_i(t),\omega_i) = h_0(t)\exp\left[\gamma'\omega_i + \sum_q \alpha_q m_{iq}(t)\right] \tag{1}$$

where $M_i(t) = \{m_{iq}(s), 0 \leq s < t, \forall q = 1,\ldots,Q\}$ indicates the history of the true unobserved longitudinal processes up to time t, α_q quantifies the effect of the longitudinal outcome q onto the risk of an event, $h_0(t)$ indicates the baseline hazard function, and ω_i are the covariates that influence the risk of the event with coefficient γ. Concerning the longitudinal sub-model, a linear multivariate mixed model is proposed:

$$y_{iq}(t) = m_{iq}(t) + \varepsilon_{iq}(t) \tag{2}$$

where q is the longitudinal variable index, $y_{iq}(t)$ is composed by the $m_{iq}(t) = x'_{iq}(t)\beta_q + z'_{iq}(t)b_{iq}$ and by a random error term $\varepsilon_{iq}(t) \sim N(0,\sigma^2)$, and β_q are the fixed effects for $x_{iq}(t)$, while b_{iq} are the random effects for $z_{iq}(t)$. In addition, $b'_i = (b'_{1q},...,b'_{iQ}) \sim N(0,D)$ and $b_{1q},...,b_{nQ}$ and $\varepsilon_{1q},...,\varepsilon_{nQ}$ are independent.
There are two classes of estimation method, the two-stage approach and the joint likelihood formulation. The two-stage approach is biased but less computationally demanding, while the joint likelihood is more efficient but computationally slower. The two-stage approach is based on two steps. In the first one the random effects are estimated using a least-squares approach, while in the second step the estimates previously found are used to impute appropriate values of $m_{iq}(t)$ that are substituted in the classical partial likelihood of the Cox model. The joint likelihood could be based on maximum likelihood, a Bayesian estimation of joint models using MCMC, or some hypothesis concerning the normal distribution of random effects or of covariates. Rizopoulos [6] proposed a new method of estimation based on the joint likelihood formulation, maximising the log-likelihood function through the Expectation-Maximisation (EM) and the Newton-Raphson algorithm.
The aim of the paper is to extend this method of estimation [6], to the case in which the longitudinal sub-model is multivariate. Starting from the classical log-likelihood equation, for each subject i it can be defined as:

$$\begin{aligned}\log p(T_i,\delta_i,y_i;\Theta) &= \log\int p(T_i,\delta_i,y_i,b_i;\Theta)db_i\\ &= \log\int p(T_i,\delta_i|b_i;\theta_t,\beta)\left\{\prod_q p(y_{iq}|b_{iq};\theta_y)\right\}p(b_i,\theta_b)db_i\end{aligned}$$

where $\Theta = (\theta'_t,\theta'_y,\theta'_b)'$ denotes the full parameter vector, with θ_t denoting the parameters for the event time outcome, θ_y the parameters for the longitudinal outcomes, and θ_b the unique parameters of the random-effects covariance matrix. In formula, $\theta_y = [\beta',\sigma^2]$ where $\beta = [\beta_1,...,\beta_q,...,\beta_Q]$ and $\sigma^2 = [\sigma_1^2,,...,\sigma_q^2,...,\sigma_Q^2]$; $\theta_t = [\gamma',\alpha_1,...,\alpha_q,...,\alpha_Q,\theta_{h_0}]$ where θ_{h_0} is used in the case in which the baseline hazard is parametric; and $\theta_b = [vech(D)]$. It is possible to separate the log-likelihood in three parts, where each part is related only to a part of the vector of parameters involved.
For maximising the log-likelihood function the Expectation-Maximisation (EM) al-

gorithm is used where the random effects are treated as ”missing data”. Accordingly, for the E-step the expected value of the complete data log-likelihood function considering the random effects as the missing data is considered. A numerical integration procedures must be employed as an integral with respect to the random effects is employed, such as Guass-Hermite quadrature rule.
In the M-step it is possible to obtain the estimation for σ_q^2 and D in closed form solution. For the others parameters there is not a close solution, so it is necessary to use one-step Newton-Raphson update:

$$\hat{\beta}^{it+1} = \hat{\beta}^{it} - \left\{ \frac{\partial}{\partial \beta} S(\hat{\beta}^{it}) \right\}^{-1} S(\hat{\beta}^{it}) \quad ; \quad \hat{\theta}_t^{it+1} = \hat{\theta}_t^{it} - \left\{ \frac{\partial}{\partial \theta_t} S(\hat{\theta}_t^{it}) \right\}^{-1} S(\hat{\theta}_t^{it})$$

where $\hat{\beta}^{it}$ and $\hat{\theta}_t^{it}$ denote the values of β and θ_t at the current iteration. In addition, $S(\hat{\beta}^{it})$ and $S(\hat{\theta}_t^{it})$ denote the corresponding blocks of the Hessian matrix, evaluated at $\hat{\beta}^{it}$ and $\hat{\theta}_t^{it}$, respectively. For the evaluation of the blocks of the Hessian matrix, the numerical derivative routine is used.
At convergence the standard errors are evaluate with the empirical information matrix [7]:

$$I_e(\theta) = \sum_{i=1}^{n} s_i s_i' - n^{-1} \left(\sum_{i=1}^{n} s_i \right) \left(\sum_{i=1}^{n} s_i \right)' \tag{3}$$

where $s_i = \frac{\partial l_i(\theta)}{\partial \theta}$.
We implemented the algorithm in R software. The outline of the algorithm follows the points:

1. The initial values are estimated through the two-stage approach.
2. In the E-step the expected value of the complete data log-likelihood function is used considering the random effects as the missing data using, in addition, Guass-Hermite quadrature rule
3. In the M-step, for σ_q^2 and D it is possible to obtain closed form solution, while for the parameter γ, α_q and β_q a one-step Newton-Raphson update is implemented. Random effects and the baseline hazard are updated.
4. Iterate between step 2 and 3 until the algorithm converges, when the parameter estimates become stable.
5. At convergence, the standard errors for each parameter are calculated using empirical information matrix.

3 Application to Primary Biliary Cirrhosis dataset

We apply the algorithm to the primary biliary cirrhosis dataset (PBCSEQ) that is available from the package *Survival* in R [9]. The dataset established from Mayo Clinic consists of 312 clinical trial patients with primary biliary cirrhosis [8] followed up from 1974 to 1986. For each patient multiple laboratory results were collected at each visit of the follow-up. After analysing several possible models, two longitudinal covariates are considered: the level of serum bilirubin in mg/dl (*serBilir*), and the level of albumin in mg/dl (*albumin*). The observational time is expressed in days. In the survival sub-model, the exogenous covariate patient's age at registration in years (*age*) is analysed. Accordingly the longitudinal and the survival sub-models used are:

$$\begin{cases} y_{i1}(t) = \beta_{01} + \beta_{11}t + b_{i01} + b_{i11}t + \varepsilon_{i1}(t) \\ y_{i2}(t) = \beta_{02} + \beta_{12}t + b_{i02} + b_{i12}t + \varepsilon_{i2}(t) \\ h_i(t) = h_0(t)\exp[\alpha_1 m_{i1}(t) + \alpha_2 m_{i2}(t) + \gamma_1 age] \end{cases}$$

where $y_{i1}(t)$ is the *log(serBilir)* and $y_{i2}(t)$ is the *albumin*.
The results obtained using the new algorithm implemented are shown in Table 1, where every parameter results to be statically significant.

Table 1 Results of the joint model on PBCSEQ dataset

Parameter	Est.	SE	p-value
α_1 (*log(serBilir)*)	1.1700	0.1052	< 0.0001
α_2 (*albumin*)	-1.8784	0.1557	< 0.0001
γ_1 (*age*)	0.0510	0.0075	< 0.0001
β_{01} (*Intercept*)	0.6371	0.0134	< 0.0001
β_{11} (*Time*)	0.0005	$8.3434*10^{-06}$	< 0.0001
β_{02} (*Intercept*)	3.5345	0.0201	< 0.0001
β_{12} (*Time*)	-0.0003	$1.1751*10^{-05}$	< 0.0001

Log-likelihood -2957.639

In particular, the *log(serBilir)* affects positively the risk of death (a one point increase in the *log(serBilir)* is associated with a 3.2220 $(= \exp(1.1700))$ fold increase in the risk of death), while the *albumin* affects negatively the risk of death (a one point increase the *albumin* will give a 0.1528 $(= \exp(-1.8784))$ fold decrease in the risk of death). Moreover the exogenous variable, *age*, affects positively the risk of death (one point increase in *age* gives a 1.0523 $(= \exp(0.0510))$ fold increase in the risk of death). Analysing the longitudinal sub-models, the observational time affects positively $(\beta_{11} = 0.0005)$ the level of *log(serBilir)*, on the contrary it is negatively associated $(\beta_{12} = -0.0003)$ with the level of *albumin*.

4 Conclusions and ideas of further work

The aim of the paper is to extend the maximum likelihood estimation method proposed by Rizopoulos [6] to the case in which the longitudinal sub-model is multivariate. We presented the algorithm and applied it to the PBCSEQ dataset.

The results are encouraging and deal to several ideas of future work. Developing, for instance, deeper diagnostic analysis and dynamic predictions. Another idea for further work is extending the survival sub-model, studying the joint effect of more than one longitudinal covariate on more than one terminal event.

References

1. P. Albert and J. Shih. An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*, 4:1517–1532, 2010.
2. C. Faucett and D. Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15:1663–1685, 1996.
3. A. Henderson, V. De Gruttola, and M. Wulfsohn. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480, 2000.
4. G. Hickey, P. Philipson, A. Jorgensen, R. Kolamunnage-Dona, P. Williamson, and D. Rizopoulos. *joineRML: Joint Modelling of Multivariate Longitudinal Data and Time-to-Event Outcomes*, 2017. version 0.4.1.
5. G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16:117–131, 2016.
6. D. Rizopoulos. *Joint model for Longitudinal and Time-to-Event Data with applications in R*. CRC Press, Boca Raton, 2012.
7. A. Scott. Maximum likelihood estimation using the empirical fisher information matrix. *Journal of Statistical Computation and Simulation*, 72(8):599–611, 2002.
8. T. Therneau and P. Grambsch. *Modeling Survival Data: extending the Cox Model*. Springer-Verlang, New York, 2000.
9. T. Therneau and T. Lumley. *survival: Survival Analysis*, 2015. version 2.41-3.
10. J. Xu and S. Zeger. The evaluation of multiple surrogate endpoints. *Biometrics*, 57:81–87, 2001.
11. J. Xu and S. Zeger. Joint analysis of longitudinal data comprising repeated measures and times to events. *Applies Statistics*, 50:375–387, 2001.

Multipurpose optimal designs for hypothesis testing in normal response trials

Disegni ottimi multi-obiettivo per la verifica di ipotesi

Marco Novelli and Maroussa Zagoraiou

Abstract This work deals with the problem of designing multiarm clinical trials for comparing treatments in order to achieve a compromise between the power of the classical Wald test of homogeneity of the treatment effects and ethical demands. In [5] the authors derived the target allocation maximizing the non-centrality parameter of Wald test for normal responses under a suitable ethical constraint reflecting the treatment effects. Starting from these results, in this paper we provide some important properties of this constrained optimal allocation, like e.g. its D_A-admissibility and its efficiency with respect to ethical and inferential criteria, taking into account estimation precision as well. Comparisons with some allocation proportions proposed in the literature are also presented.
Abstract *Questo lavoro riguarda il problema della pianificazione ottimale di esperimenti comparativi volti ad ottenere validi compromessi tra precisione inferenziale ed esigenze etiche. Prendendo in considerazione il modello normale, in [5] è stata derivata l'allocazione ideale dei trattamenti che massimizza la potenza del test di Wald basato sui contrasti, sotto opportuni vincoli etici legati agli effetti dei singoli trattamenti. L'obiettivo di questo articolo è quello di fornire alcune importanti proprietà di tale allocazione, ossia la D_A-ammissibilità e la sua efficienza rispetto a criteri sia etici che inferenziali, riguardanti anche la precisione di stima, effettuando inoltre opportuni confronti con altre allocazioni target proposte in letteratura.*

Key words: asymptotic inference; ethics; power; multiarm clinical trials

Marco Novelli
Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, Bologna
e-mail: marco.novelli4@unibo.it

Maroussa Zagoraiou
Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, Bologna
e-mail: maroussa.zagoraiou3@unibo.it

1 Introduction

The large majority of randomized clinical trials for treatment comparisons have been designed in order to achieve balanced allocation among the treatment groups, with the aim of maximizing inferential precision about the estimation of the treatment effects. The main justification concerns the so-called "universal optimality" of the balanced design (see e.g. [8]), especially in the context of the linear homoscedastic model, since it optimizes the usual design criteria for the estimation of the treatment contrasts, (like the well-known D-optimality minimizing the volume of the confidence ellipsoid of the contrasts), and it is nearly optimal under several optimality criteria, also under heteroscedasticity [6, 7].

Taking into account the problem of testing statistical hypothesis about the equality of the treatment effects, balance is still optimal in the case of two treatments, since it maximizes the power of the test for normal homoscedastic responses and it is asymptotically optimal in the case of binary outcomes (see e.g. [2, 3]). However, in the case of several treatments the balanced allocation may not be efficient, since it is significantly different from the optimal design for hypothesis testing and could be strongly inappropriate for phase III-trials, in which the ethical demand of individual care often induces to skew the allocations to more efficacious (or less toxic) treatments. To derive a suitable compromise between these goals, Baldi Antognini et al. [5] suggested a constrained optimal target which maximizes the power of the classical Wald test of homogeneity, subject to an ethical constraint on the allocation proportions reflecting the efficacy of the treatments. The aim of the present work is to push forward the results in [5], by providing some important properties of this constrained optimal allocation like, e.g., the D_A-admissibility, and its efficiency with respect to both ethical and inferential criteria, taking into account estimation precision as well. Comparisons with some targets proposed in the literature are also presented.

2 Notation and model

Consider a clinical trial where patients come sequentially and are assigned to one of K available treatments. At each step n, let $\delta_{kn} = 1$ if the nth patient is allocated to the kth ($k = 1, \dots, K$) treatment and 0 otherwise, where $\sum_{k=1}^{K} \delta_{kn} = 1$. Let Y_n be the normally distributed response of the corresponding subject, with $E(Y_n \mid \delta_{kn} = 1) = \mu_k$ denoting the treatment effect and $V(Y_n \mid \delta_{kn} = 1) = \sigma^2$ the unknown common variance; conditionally on the allocations, the responses are assumed to be independent. Furthermore, we denote by $\boldsymbol{\pi}_n^\top = (\pi_{1n}, \dots, \pi_{Kn})$ the vector collecting the proportion of patients assigned to the treatments up to that stage, where $\pi_{kn} = n^{-1} \sum_{i=1}^{n} \delta_{ki}$ ($k = 1, \dots, K$) and $\sum_{k=1}^{K} \pi_{kn} = 1$; also let $\hat{\mu}_{kn}$ ($k = 1, \dots, K$) be the MLE of μ_k, i.e. the sample mean, so $\boldsymbol{\mu}^\top = (\mu_1, \dots, \mu_K)$ and $\hat{\boldsymbol{\mu}}_n^\top = (\hat{\mu}_{1n}, \dots, \hat{\mu}_{Kn})$ are the vectors of the treatment effects and their estimates, respectively. In what follows we assume "the larger the better" scenario and the following ordering regarding the treatment effects $\mu_1 > \mu_2 > \ldots > \mu_K$.

After n steps the Fisher information matrix (conditional on the design) associated with $\boldsymbol{\mu}$ is $\mathbf{M} = \mathbf{M}(\boldsymbol{\mu} \mid \boldsymbol{\pi}_n) = \sigma^{-2}\text{diag}\,(\pi_{kn})_{k=1,\ldots,K}$. Several authors suggested target allocations $\boldsymbol{\rho}^\top = (\rho_1, \ldots, \rho_K)$ (with $\rho_k \geq 0$ and $\sum_{k=1}^K \rho_k = 1$) in order to optimize the estimation of the treatment effects by choosing suitable criteria regarding $\mathbf{M}(\boldsymbol{\mu} \mid \boldsymbol{\rho})$.

In the context of multiarm clinical trials, the inferential attention is usually devoted to the contrasts. So, letting $\mathbf{A}^\top = [\mathbf{1}_{K-1} \mid -\mathbf{I}_{K-1}]$, where $\mathbf{1}_r$ and $\mathbf{I}_r$ represent the r-dim vector of ones and the identity matrix, respectively, then the vector of contrasts wrt the first treatment (considered as the reference) is $\boldsymbol{\mu}_c = \mathbf{A}^\top\boldsymbol{\mu} = (\mu_1 - \mu_2, \ldots, \mu_1 - \mu_K)^\top$. Under well-known regularity conditions, the corresponding MLE $\hat{\boldsymbol{\mu}}_{cn} = \mathbf{A}^\top\hat{\boldsymbol{\mu}}_n$ is strongly consistent and asymptotically normal with $\sqrt{n}(\hat{\boldsymbol{\mu}}_{cn} - \boldsymbol{\mu}_{cn}) \hookrightarrow_d \mathscr{N}\left(\mathbf{0}, \mathbf{A}^\top\mathbf{M}^{-1}\mathbf{A}\right)$. Within this framework, the balanced design $\boldsymbol{\rho}^B$, namely $\rho_k = K^{-1}$ for every $k = 1, \ldots, K$ is the so-called D_A-optimal allocation, since it minimizes $\det[\mathbf{A}^\top\mathbf{M}^{-1}\mathbf{A}]$.

Whereas, taking into account the problem of testing hypothesis on the equality of the treatments effects, i.e., $H_0 : \boldsymbol{\mu}_c = \mathbf{0}_{K-1}$, versus the alternative $H_A : \boldsymbol{\mu}_c \neq \mathbf{0}_{K-1}$, where $\mathbf{0}_{K-1}$ is the $(K-1)$-dim vector of zeros, then the optimal design maximizing the power of the classical Wald test is $\boldsymbol{\rho}^* = (1/2, 0, \ldots, 0, 1/2)^\top$ (see [5]). Clearly, this optimal allocation is unsuitable both from the ethical and the inferential point of views.

Regarding ethics, Atkinson [1] proposed a target intended to skew the assignments towards the best treatment in order to minimize the exposure of patients to toxic (or inefficacious) treatments. In particular, denoting by $\bar{\mu} = K^{-1}\sum_{k=1}^K \mu_k = \boldsymbol{\mu}^\top\boldsymbol{\rho}^B$ the overall treatment mean, the target $\boldsymbol{\rho}^A_{(\gamma)}$ proposed by Atkinson is

$$\rho^A_{(\gamma)k} = \Phi\left(\frac{\mu_k - \bar{\mu}}{\gamma}\right) \Big/ \left[\sum_{i=1}^K \Phi\left(\frac{\mu_i - \bar{\mu}}{\gamma}\right)\right], \quad k = 1, \ldots, K.$$

In the same spirit, instead of $\Phi(\cdot)$ any non-negative increasing function can be used. An example is the exponential target $\boldsymbol{\rho}^E_{(\gamma)}$ given by

$$\rho^E_{(\gamma)k} = e^{\frac{\mu_k - \bar{\mu}}{\gamma}} \Big/ \left(\sum_{i=1}^K e^{\frac{\mu_i - \bar{\mu}}{\gamma}}\right) = e^{\frac{\mu_k}{\gamma}} \Big/ \left(\sum_{i=1}^K e^{\frac{\mu_i}{\gamma}}\right), k = 1, \ldots, K.$$

Clearly, small values of γ induce a strong ethical skew, while as γ increases more emphasis is given to inferential purposes. In particular, adopting $\boldsymbol{\rho}^E_{(\gamma)}$, the allocation proportion $\rho^E_{(\gamma)1}$ to the best treatment is decreasing as γ grows, since

$$\frac{\partial \rho^E_{(\gamma)1}}{\partial\gamma} = \frac{\sum_{i=1}^K e^{\frac{\mu_i + \mu_1}{\gamma}}(\mu_i - \mu_1)}{\left(\sum_{i=1}^K e^{\frac{\mu_i}{\gamma}}\right)^2 \gamma^2} < 0.$$

Note that such monotonicity property does not hold, in general, for $\boldsymbol{\rho}^A$ as we shall show in the last section.

3 Constrained optimal allocation and its D_A-admissibility

Adopting a constrained optimization framework, Baldi Antognini et al. in [5] derived the allocation maximizing the power of Wald test under a suitable ethical constraint reflecting the efficacy of the treatments. In particular, the optimal target $\tilde{\boldsymbol{\rho}}^\top{=}(\tilde{\rho}_1,\ldots,\tilde{\rho}_K)$ maximizing the non-centrality parameter $\phi(\tilde{\boldsymbol{\rho}}) = n\sigma^{-2}\boldsymbol{\mu}_c^\top\left[\mathbf{A}^\top \mathrm{diag}(\tilde{\boldsymbol{\rho}})^{-1}\mathbf{A}\right]^{-1}\boldsymbol{\mu}_c$ of the multivariate Wald test subject to the ethical constraint $\tilde{\rho}_1 \geq \tilde{\rho}_2 \geq \ldots \geq \tilde{\rho}_K$ is $\tilde{\boldsymbol{\rho}} = (1-t[K-1], t, \ldots, t)^\top$ if $t \leq K^{-1}$, while $\tilde{\boldsymbol{\rho}} = \boldsymbol{\rho}^B$ if $t > K^{-1}$, where $t = \sum_{k=2}^K(\mu_1-\mu_k)^2 / \left\{2\left[\sum_{k=2}^K(\mu_1-\mu_k)\right]^2\right\}$. Table 1 shows how the allocation $\tilde{\boldsymbol{\rho}}$ moves away from the balanced design as the distance between μ_1 and μ_2 increases, skewing the assignments to the superior treatment.

Table 1: The behaviour of the optimal constrained target $\tilde{\boldsymbol{\rho}}$ with $K=3$ as μ_2 varies.

μ_1	μ_2	μ_3	$\tilde{\rho}_1$	$\tilde{\rho}_2$	$\tilde{\rho}_3$	t
15	14	6	0.333	0.333	0.333	0.410
15	12	6	0.375	0.312	0.312	0.312
15	10	6	0.459	0.270	0.270	0.270
15	8	6	0.492	0.254	0.254	0.254

Following the definition of admissibility proposed in [4], it is easy to show that $\tilde{\boldsymbol{\rho}}$ is D_A-admissible, i.e. it does not exist another allocation which is simultaneously superior wrt both ethics and D_A-optimality. Indeed, when $t > 1/K$, $\tilde{\boldsymbol{\rho}} = \boldsymbol{\rho}^B$ and the D_A-admissibility is trivially satisfied, while for $t \leq 1/K$, $\boldsymbol{\mu}^\top(\tilde{\boldsymbol{\rho}} - \boldsymbol{\rho}^B) \geq 0 \iff \boldsymbol{\mu}^\top\tilde{\boldsymbol{\rho}} \geq \bar{\mu} \iff \mu_1(1-Kt) \geq \bar{\mu}(1-Kt)$ which is always true since $\mu_1 > \bar{\mu}$.

4 Comparisons

We now compare the performance of $\tilde{\boldsymbol{\rho}}$, $\boldsymbol{\rho}^B$, Atkinson's target $\boldsymbol{\rho}_\gamma^A$ and the exponential one $\boldsymbol{\rho}_\gamma^E$ both with $\gamma = 1$ and $\gamma = 3$. In particular, in Table 2 we consider the following criteria: i) an ethical measure of efficiency given by the ratio between the total expected outcomes and its optimal value, i.e., $E_E(\boldsymbol{\rho}) = \sum_{k=1}^K \mu_k\rho_k/\mu_1$, ii) an efficiency measure of statistical power $E_P(\boldsymbol{\rho}) = \phi(\boldsymbol{\rho})/\phi(\boldsymbol{\rho}^*)$ and the D_A-efficiency $E_{D_A}(\boldsymbol{\rho}) = \left\{\det\left[\mathbf{A}^\top\mathbf{M}^{-1}\left(\boldsymbol{\rho}^B\right)\mathbf{A}\right]/\det\left[\mathbf{A}^\top\mathbf{M}^{-1}(\boldsymbol{\rho})\mathbf{A}\right]\right\}^{\frac{1}{K-1}}$ for estimation.

Table 2: the case $K = 5$ treatments

$\boldsymbol{\mu}^\top$	Targets	$E_E(\boldsymbol{\rho})$	$E_P(\boldsymbol{\rho})$	$E_{D_A}(\boldsymbol{\rho})$
(21,20,19,18,16)	$\boldsymbol{\rho}^A_{(1)} = (0.37, 0.332, 0.217, 0.08, 0.001)^\top$	0.952	0.147	0.282
	$\boldsymbol{\rho}^A_{(3)} = (0.305, 0.26, 0.209, 0.157, 0.07)^\top$	0.929	0.321	0.867
	$\tilde{\boldsymbol{\rho}} = (0.355, 0.161, 0.161, 0.161, 0.161)^\top$	0.916	0.503	0.930
	$\boldsymbol{\rho}^E_{(1)} = (0.641, 0.236, 0.087, 0.032, 0.004)^\top$	0.975	0.112	0.274
	$\boldsymbol{\rho}^E_{(3)} = (0.359, 0.257, 0.184, 0.132, 0.068)^\top$	0.935	0.324	0.830
	$\boldsymbol{\rho}^B$	0.895	0.474	1
(23,20,19,18,16)	$\boldsymbol{\rho}^A_{(1)} = (0.43, 0.339, 0.181, 0.05, 0)^\top$	0.913	0.264	0.186
	$\boldsymbol{\rho}^A_{(3)} = (0.364, 0.246, 0.192, 0.14, 0.058)^\top$	0.886	0.392	0.813
	$\tilde{\boldsymbol{\rho}} = (0.452, 0.137, 0.137, 0.137, 0.137)^\top$	0.887	0.554	0.840
	$\boldsymbol{\rho}^E_{(1)} = (0.93, 0.046, 0.017, 0.006, 0.001)^\top$	0.989	0.068	0.059
	$\boldsymbol{\rho}^E_{(3)} = (0.522, 0.192, 0.137, 0.099, 0.051)^\top$	0.914	0.406	0.680
	$\boldsymbol{\rho}^B$	0.835	0.438	1
(25,20,19,18,16)	$\boldsymbol{\rho}^A_{(1)} = (0.504, 0.33, 0.138, 0.028, 0)^\top$	0.893	0.367	0.112
	$\boldsymbol{\rho}^A_{(3)} = (0.41, 0.235, 0.179, 0.126, 0.049)^\top$	0.857	0.479	0.760
	$\tilde{\boldsymbol{\rho}} = (0.476, 0.131, 0.131, 0.131, 0.131)^\top$	0.859	0.618	0.814
	$\boldsymbol{\rho}^E_{(1)} = (0.99, 0.007, 0.002, 0.001, 0)^\top$	0.998	0.015	0.009
	$\boldsymbol{\rho}^E_{(3)} = (0.68, 0.128, 0.092, 0.066, 0.034)^\top$	0.922	0.427	0.486
	$\boldsymbol{\rho}^B$	0.784	0.446	1
(27,20,19,18,16)	$\boldsymbol{\rho}^A_{(1)} = (0.595, 0.297, 0.094, 0.014, 0)^\top$	0.890	0.428	0.060
	$\boldsymbol{\rho}^A_{(3)} = (0.449, 0.227, 0.168, 0.115, 0.041)^\top$	0.836	0.550	0.710
	$\tilde{\boldsymbol{\rho}} = (0.486, 0.129, 0.129, 0.129, 0.129)^\top$	0.833	0.669	0.803
	$\boldsymbol{\rho}^E_{(1)} = (0.999, 0.001, 0, 0, 0)^\top$	1	0.003	0.001
	$\boldsymbol{\rho}^E_{(3)} = (0.805, 0.078, 0.056, 0.04, 0.021)^\top$	0.941	0.352	0.309
	$\boldsymbol{\rho}^B$	0.741	0.463	1
(25,20,19,18,11)	$\boldsymbol{\rho}^A_{(1)} = (0.351, 0.323, 0.23, 0.096, 0)^\top$	0.853	0.155	$\to 0$
	$\boldsymbol{\rho}^A_{(3)} = (0.372, 0.257, 0.209, 0.159, 0.002)^\top$	0.853	0.175	0.382
	$\tilde{\boldsymbol{\rho}} = (0.402, 0.149, 0.149, 0.149, 0.149)^\top$	0.809	0.467	0.890
	$\boldsymbol{\rho}^E_{(1)} = (0.99, 0.007, 0.002, 0.001, 0)^\top$	0.998	0.006	0.002
	$\boldsymbol{\rho}^E_{(3)} = (0.699, 0.132, 0.095, 0.068, 0.007)^\top$	0.928	0.165	0.332
	$\boldsymbol{\rho}^B$	0.744	0.413	1
(25,20,19,18,13)	$\boldsymbol{\rho}^A_{(1)} = (0.4, 0.337, 0.2, 0.063, 0)^\top$	0.867	0.213	0.007
	$\boldsymbol{\rho}^A_{(3)} = (0.391, 0.252, 0.2, 0.148, 0.009)^\top$	0.856	0.252	0.537
	$\tilde{\boldsymbol{\rho}} = (0.436, 0.141, 0.141, 0.141, 0.141)^\top$	0.831	0.498	0.857
	$\boldsymbol{\rho}^E_{(1)} = (0.99, 0.007, 0.002, 0.001, 0)^\top$	0.998	0.008	0.004
	$\boldsymbol{\rho}^E_{(3)} = (0.695, 0.131, 0.094, 0.067, 0.013)^\top$	0.926	0.233	0.389
	$\boldsymbol{\rho}^B$	0.760	0.411	1
(25,20,19,18,15)	$\boldsymbol{\rho}^A_{(1)} = (0.465, 0.337, 0.16, 0.038, 0)^\top$	0.884	0.303	0.052
	$\boldsymbol{\rho}^A_{(3)} = (0.406, 0.243, 0.187, 0.134, 0.03)^\top$	0.857	0.384	0.693
	$\tilde{\boldsymbol{\rho}} = (0.464, 0.134, 0.134, 0.134, 0.134)^\top$	0.850	0.562	0.827
	$\boldsymbol{\rho}^E_{(1)} = (0.99, 0.007, 0.002, 0.001, 0)^\top$	0.998	0.012	0.007
	$\boldsymbol{\rho}^E_{(3)} = (0.686, 0.13, 0.093, 0.067, 0.024)^\top$	0.923	0.345	0.453
	$\boldsymbol{\rho}^B$	0.776	0.426	1
(25,20,19,18,17)	$\boldsymbol{\rho}^A_{(1)} = (0.547, 0.317, 0.116, 0.02, 0.001)^\top$	0.903	0.453	0.204
	$\boldsymbol{\rho}^A_{(3)} = (0.411, 0.226, 0.17, 0.118, 0.075)^\top$	0.857	0.598	0.813
	$\tilde{\boldsymbol{\rho}} = (0.485, 0.129, 0.129, 0.129, 0.129)^\top$	0.866	0.700	0.803
	$\boldsymbol{\rho}^E_{(1)} = (0.99, 0.007, 0.002, 0.001, 0)^\top$	0.998	0.020	0.011
	$\boldsymbol{\rho}^E_{(3)} = (0.671, 0.127, 0.091, 0.065, 0.047)^\top$	0.920	0.536	0.520
	$\boldsymbol{\rho}^B$	0.792	0.485	1

Considering the statistical power, $\tilde{\boldsymbol{\rho}}$ has the best performance with a gain up to 13% with respect to any second best option. The rules $\boldsymbol{\rho}^{A}_{(1)}$ and $\boldsymbol{\rho}^{E}_{(1)}$ show the lowest statistical power but, at the same time, the highest ethical efficiency. Note that, as γ grows, more emphasis is devoted to inference. However, contrary to the exponential target, the Atkinson's allocation proportion to the best treatment, $\rho^{A}_{(\gamma)1}$, is not always decreasing in γ. Moreover, $\tilde{\boldsymbol{\rho}}$ performs very well also from the ethical point of view.

Regarding the D_A-efficiency, $\tilde{\boldsymbol{\rho}}$ is substantially superior with respect to $\boldsymbol{\rho}^{A}$ and $\boldsymbol{\rho}^{E}$ guaranteing at the same time an efficiency always greater than 80.3%. Note that, adopting $\boldsymbol{\rho}^{A}_{(1)}$ and $\boldsymbol{\rho}^{E}_{(1)}$ the D_A-efficiency often tends to zero and therefore the estimation precision may vanish.

Since ethics and inference are conflicting demands, a target showing high efficiency under one criterion may perform worst under other criteria. However, $\tilde{\boldsymbol{\rho}}$ represents a valid compromise between inferential (both in terms of power and estimation precision) and ethical concerns.

Acknowledgements Research supported by the Italian Ministry of Education, University and Research under PRIN 2015 “Environmental processes and human activities: capturing their interactions via statistical methods (EphaStat)”

References

1. Atkinson, A.C.: Adaptive biased-coin designs for clinical trials with several treatments. Discussiones Mathematicae Probability and Statistics **24**, 85–108 (2004)
2. Azriel, D., Mandel, M., Rinott, Y.: Optimal allocation to maximize power of two-sample tests for binary response. Biometrika **99**, 101–113 (2012)
3. Baldi Antognini, A.: A theoretical analysis of the power of biased coin designs. J Stat Plan Inference **138**, 1792–1798 (2008)
4. Baldi Antognini, A., Giovagnoli, A.: Compound optimal allocation for individual and collective ethics in binary clinical trials. Biometrika **97**, 935–946 (2010)
5. Baldi Antognini, A., Novelli, M., Zagoraiou, M.: Optimal designs for testing hypothesis in multiarm clinical trials. Submitted.
6. Begg, C.B., Kalish, L.A.: Treatment Allocation for Nonlinear Models in Clinical Trials: The Logistic Model. Biometrics **40**, 409–420 (1984)
7. Kalish, L.A., Harrington, D.P.: Efficiency of balanced treatment allocation for survival analysis. Biometrics **44**, 409–420 (1988)
8. Silvey, S.D.: Optimal Designs. Chapman & Hall, London (1980)

Additive Bayesian networks for an epidemiological analysis of swine diseases

Reti Bayesiane additive per un'analisi epidemiologica di malattie suine

Marta Pittavino and Reinhard Furrer

Abstract Additive Bayesian networks (ABNs) are types of graphical models that extend the usual generalized linear model (GLM) to multiple dependent variables through the representation of joint probability distribution. Thanks to their flexible properties, ABNs have been widely used in epidemiological analyses. In this work we present a veterinary case study where ABNs are used to explore multivariate swine diseases data of medical relevance. We then compare the results with a classical methodology. Finally, we highlight the key difference between a multivariable standard (GLM) and a multivariate (ABN) approach: the latter attempts not only to identify statistically associated variables, but also to additionally separate these into those directly and indirectly dependent with one or more outcome variables.

Abstract *Le reti Bayesiane additive (ABNs) sono tipi di modelli grafici che estendono l'usuale modello lineare generalizzato (GLM) a variabili multiple dipendenti attraverso la rappresentazione della distribuzione di probabilità congiunta. Grazie alle loro proprietà flessibili, le reti ABNs sono state ampiamente utilizzate nelle analisi epidemiologiche. In questo lavoro presentiamo un caso di studio veterinario in cui il metodo ABN viene utilizzato per esplorare dati multivariati su malattie suine di rilevanza medica. In seguito confrontiamo i risultati con una metodologia classica. Infine, evidenziamo la differenza chiave tra un approccio standard multivariabile (GLM) e uno multivariato (ABN): quest'ultimo tenta non solo di identificare le variabili associate statisticamente, ma anche di separarle ulteriormente in quelle direttamente e indirettamente dipendenti con una o più variabili d'esito.*

Key words: abn R package; data mining; machine learning; statistical modeling; structure discovery; zoonosis.

Marta Pittavino
University of Geneva, Research Center for Statistics (RCS), Geneva School of Economics and Management (GSEM), Geneva, Switzerland e-mail: marta.pittavino@unige.ch

Reinhard Furrer
Department of Mathematics (I-MATH) and Department of Computational Science, University of Zurich (UZH), Zurich, Switzerland e-mail: reinhard.furrer@math.uzh.ch

1 Introduction

A primary objective of many epidemiological studies is to investigate hypothesized relationships between covariates of interest, and one and more outcome variables, through analyses of appropriate data. From a data analysis perspective, this is often far from being trivial. Diseases and health conditions, which are a priority for control or eradication in humans and animals, are increasingly recognized to have highly complex determinants. Typically, the unknown stochastic processes, which generated these data, are highly complex, resulting in multiple correlation/dependencies between covariates and between outcome variables. Standard epidemiological and statistical approaches cannot adequately describe such inter-dependent multifactorial relationships. ABN modelling is a data mining/machine learning methodology, which has demonstrated to be ideally suited for such analyses [1, 2].

2 Material and methods

2.1 The data

We present data on disease occurrence in pigs provided by the industry body 'British Pig Health Scheme' (BPHS). The main objective of BPHS is to improve the productivity of pig in the UK, and reducing disease occurrence is a significant part of this process. The data we consider here comprise of a randomly chosen batch of 50 pigs from each of 500 randomly chosen pig producers in the UK. In total we deal with 25'000 observations, i.e. animals entering the human food chain at an abattoir: 'finishing pigs'. Each animal is assessed for the presence of a range of different disease conditions by a specialist swine veterinarian.

Then, the resulting variables are binary due to the presence or the absence of a specific disease. We consider here the following ten disease conditions, all abbreviated to ease the notation and described in [3]: enzootic-pneumonia (EP); pleurisy (PL); milk spots (MS); hepatic scarring (HS); pericarditis (PC); peritonitis (PT); lung abscess (AB); tail damage (TD); pyaemia (PY) and papular dermatitis (PD).

The presence of any of these conditions results in an economic loss to the producer. Either directly due to the relevant infected part of the animal being removed from the food chain, or indirectly in cases such as enzootic pneumonia, which may potentially indicate poor herd health and efficiency losses on the farm. An additional loss, though not directly monetary, is the presence of tail damage which may be suggestive of welfare concerns and linked to sub-optimal production efficiency. Milk spots and hepatic scarring result from infestation with Ascaris suum, which is particularly important as this is a zoonotic helminth parasite.

2.2 Additive Bayesian networks

A Bayesian network for a set of random variables $\boldsymbol{X} = \{X_1, \dots, X_n\}$ consists of:

- A *directed acyclic graph* (DAG) structure $\mathscr{S} = (V, E)$, where V is a finite set of vertices or nodes and E is a finite set of directed edges between the vertices. A

DAG is *acyclic*; hence, the edges in E do not form directed cycles. A random variable X_j corresponds to each node $j \in V = \{1, \ldots, n\}$ in the graph. We do not distinguish between a variable X_j and the corresponding node j.

- A set of parents for a node j is denoted by $\mathbf{Pa}_j$. A node j is said to be a *parent* of a node k if the edge set E contains an edge from j to k. P_j indicates the total number of parents for a node j : $\dim(\mathbf{Pa}_j) = P_j \geq 0$. $P_j = \emptyset$ for orphan nodes.
- A set of local probability distributions for all variables in the network called $\boldsymbol{\theta}_{\mathscr{B}}$. Each node j, with parent set $\mathbf{Pa}_j$, is parametrized by a local probability distribution: $P(X_j | \mathbf{Pa}_j)$.

Edges represent both *marginal* and *conditional dependencies*. The main role of the network structure is to express the conditional independence relationships among the variables in the model through graphical separation, thus specifying the factorization of the global probability distribution: $P(\boldsymbol{X}) = \prod_{j=1}^{n} P(X_j | \mathbf{Pa}_j)$.

We denote a Bayesian network (BN) model $\mathscr{B}$ for a set of random variables $\boldsymbol{X}$ by a pair $\mathscr{B} = (\mathscr{S}, \boldsymbol{\theta}_{\mathscr{B}})$. The DAG $\mathscr{S}$ defines the *structure*, and $\boldsymbol{\theta}_{\mathscr{B}}$ the *parametrization* of the model. In order to specify a $\mathscr{B}$ for $\boldsymbol{X}$, we must therefore specify a DAG structure and a set of local probability distributions.

An additive Bayesian network $\mathscr{A}$ consists of a Bayesian network $\mathscr{B}$ that generalizes the multinomial logistic regression model $\mathscr{M}$. The multinomial logistic regression model $\mathscr{M}$ can be integrated into a BN $\mathscr{B}$ by modelling each of its conditional probability table $P(X_j = s \mid \mathbf{Pa}_j = c) = \theta_{jcs}$ with a multinomial logistic regression model, where X_j is progressively the outcome variable and the resulting regression design matrix is constructed from $\mathbf{Pa}_j$, as showed in [4] and in detail in Figure 1.

X_1 is independent: $logit(\theta_1) = \beta_{1,0}$

X_2 is independent: $logit(\theta_2) = \beta_{2,0}$

X_3 is jointly dependent upon X_1, andX_2: $logit(\theta_3) = \beta_{3,0} + \beta_{3,1}X_1 + \beta_{3,2}X_2$

X_4 is conditionally dependent upon X_3: $logit(\theta_4) = \beta_{4,0} + \beta_{4,1}X_3$

X_5 is conditionally dependent upon X_3: $logit(\theta_5) = \beta_{5,0} + \beta_{5,1}X_3$

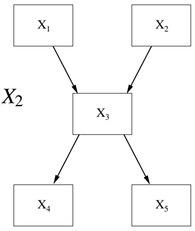


Fig. 1 A binary additive Bayesian network model $\mathscr{A}$ for five random variables.

2.3 Analysis with ABN

All analyses were conducted using the software R [5] and specifically the “abn” R package [6] which is available from CRAN “cran.r-project.org” with additional documentation and case studies at “http://www.r-bayesian-networks.org”.

Prior distributions were defined. All DAG structures were equally supported a priori with a uniform, i.e., uninformative, prior. It is possible to construct informative structural priors, i.e. penalizing models with more structural complexity, but as noted in [7] these are problematic to specify or lead to undesirable properties as in

[9]. Uninformative Gaussian priors were applied for the additive parameters at each node: specifically, independent Gaussian priors with mean zero and variance 1000.

As we are searching across DAGs - to identify optimally fitting structures - there is also the need for a prior on structures. The default being that each structure is equally supported a priori. It is possible to construct informative structural priors, for example to penalize models with more structural complexity, e.g. more arcs, but as noted in [6] these are problematic to specify in practice. In [8] an informative structural prior on the number of parents within an individual node is used, where this assumes that parent combinations with the same cardinality are equally likely. This prior gives equal weighting to a parent combination with cardinality zero and cardinality m1 which may not be entirely desirable. In the subsequent case study analyses an uninformative - flat - structural prior is used.

A two-steps procedure was used to identify a robust model.

The first step was to find an optimal ABN model $\mathscr{A}_1$. The process of identifying an optimal ABN is referred to in the literature as *structure learning* [8]. This was found with an order based exact search method [9]. The best goodness of fit to the available data was computed using the marginal likelihood (ML), equivalent to Bayes factors for models with equal structural priors and the standard Bayesian score function used in BN literature [7, 8]. The ML includes an implicit penalty for model complexity and in a binary additive Bayesian network for a node j is:

$$P(\mathscr{D}_j|\mathscr{S}) = \int_{-\infty}^{+\infty} \prod_{i=1}^{m} \left(\frac{e^{z_{ij}^T\boldsymbol{\beta}_j}}{1+e^{z_{ij}^T\boldsymbol{\beta}_j}}\right)^{x_{ij}} \left(\frac{1}{1+e^{z_{ij}^T\boldsymbol{\beta}_j}}\right)^{1-x_{ij}} \times \prod_{c=1}^{C_j} \frac{1}{\sqrt{2\pi}\sigma_c} e^{-\frac{(\beta_c-\mu_c)^2}{2\sigma_c^2}} \, d\boldsymbol{\beta}_j$$

where $\mathscr{D}_j$ are the observed data at node j, and consist of tuples of $[x_{ij}, z_{ij}^T]$. The parameter vector at node j is represented by $\boldsymbol{\beta}_j$, and has the same length as the possible parents configuration, denoted by C_j, then $\dim(\boldsymbol{\beta}_j) = C_j$. The marginal likelihood was estimated using the Laplace approximation at each node. To find the best model, the maximum number of parents allowed per node (number of covariates in each regression model at each node) was increased until the goodness of fit remained constant and thereby identified the same globally optimal ABN. The model selection procedure started from one possible parent per node and then the parent limit increased gradually until five possible parents per node [6].

In the second step, the model $\mathscr{A}_1$ was adjusted by checking it for overfitting using Markov chain Monte Carlo (MCMC) simulation implemented in JAGS (just another Gibbs sampler) [10]. A parametric bootstrapping approach was suggested in [8] which uses simulation to assess whether a chosen model comprises more complexity than could reasonably be justified given the observed data. Simulated datasets were generated with MCMC as iterations of an identical size as the original one, from the optimal model found in step one. An identical exact search for an optimal model structure was then performed exactly as in the first step, but applied to the bootstrapped data rather than original data. It was repeated 10240 times [6], a large enough number to get robust results, using the same parent limit per node as the one found in the initial search. Arcs present in less than 50% of the globally optimal ABNs - estimated from the bootstrapped data - were considered not to be robust and removed from the DAG generated in the first step. A most robust ABN model $\mathscr{A}_2$ fully adjusted for over-fitting was identified at the end of this second step, equivalent to a multivariate GLM.

2.4 Analysis with GLM

Data were analysed using the software R [5] and the “glm” function, available in the “stats” R package . As many different generalized linear models (GLMs), in particular multivariable logistic regression following the data structure, as the different number of variables were performed. Only two models based on the most significant variables and with the highest AIC score have been selected and shown.

3 Results with ABN and GLM

The resulting best fitting ABN comprised 12 arcs and a maximum number of three parents (Fig. 8 in [6]), for the variable PD (papular dermatitis) and PL (pleurisy). After the bootstrap analysis, four of the arcs in the globally optimal ABN were only weakly supported. Therefore the number of arcs was reduced from 12 to 8 (Fig. 7 in [6]). The final globally optimal additive Bayesian network model after adjustment for over-fitting is shown in Fig. 2 on the left. Three different epidemiological pathways can be identified resulting in variables indirectly linked together. The goodness of fit for this model is -44245.73. For example, forcing an additional arc connecting PC and AB gives a poorer log marginal likelihood of -44249.58.

The two GLM models with the most significant variables and the highest AIC score have PC (pericarditis) and AB (lung abscess) as response variables. Figure 2, in the middle and on the right, shows the two corresponding models. A GLM is simply a DAG where arcs are only allowed directly between the covariates and response variable. In each case we find that an arc is identified between PC and AB.

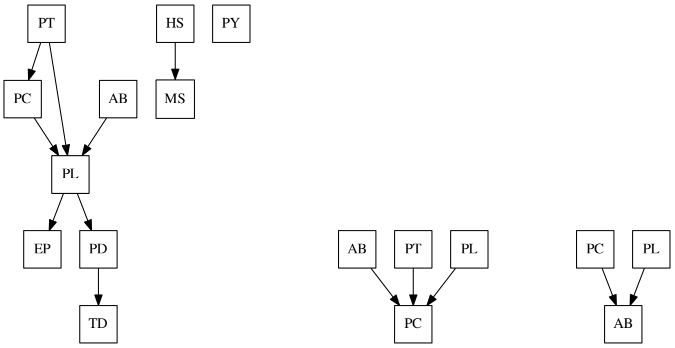


Fig. 2 Final ABN model of swine diseases data with 8 arcs after bootstrapping adjustment (left). Two globally optimal GLMs - one with PC as the dependent variable (middle), and a second with Abscess as the dependent variable (right).

4 Conclusion

In summary, we find that while the GLM analyses identifies a strongly supported statistical association between presence of pericarditis (PC) and lung abscess (AB); the ABN model does not support a direct statistical dependency between PC and AB. In the ABN model there is no arc connecting these variables, this relationship is via the intermediate variable pleurisy (PL).

This highlights the key difference between a multivariable GLM and a multivariate GLM (ABN). The former identifies variables which may be associated with the response (dependent) variable within a very restrictive model space: arcs are only allowed from covariates direct to the response variable. When considering the same data within a larger model space, which incorporates other relationships within the underlying epidemiological system which generated the observed data, then such variables may then only be supported as indirectly, rather than directly, related to the response variable. In [2, 11] there are further similar GLM and ABN examples.

These results provides a conceptual justification of the Yule-Simpson paradox, which states that an apparent relationship between variables may disappear or even be reversed when others are taken into account.

In conclusion, data analyses using ABNs have the potential to offer new insights into complex epidemiological systems.

References

1. Pittavino, M.: Additive Bayesian Networks for Multivariate Data: Parameter Learning, Model Fitting and Applications in Veterinary Epidemiology. PhD thesis. Universität Zürich (2016)
2. Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Torgerson, P., Furrer, R.: Comparison between Generalized Linear Modelling and Additive Bayesian Network. Identification of Factors associated with the Incidence of Antibodies against Leptospira interrogans sv Pomona in Meat Workers in New Zealand. Acta Tropica **173**, 191-199, ELSEVIER (2017)
3. Sanchez-Vazquez, M.J., Nielen, M., Edwards, S.A., Gunn, G.J. and Lewis F.I.: Identifying associations between pig pathologies using a multidimensional machine learning methodology. BMC Vet. Res. **8**:151, 1-11 (2012)
4. Rijmen, F.: Bayesian networks with a logistic regression model for the conditional probabilities, Int. Jour. of Appr. Reas. **48**:2, 659-666 (2008)
5. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (2017)
6. Kratzer, G., Pittavino, M., Lewis, F., Furrer, R.: abn: an R package for modelling multivariate data using additive Bayesian networks. The Comprehensive R Archive Network, 1-37 (2017)
7. Heckerman, D., Geiger, D. and Chickering, D. M.: Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning. **20**:3, 197-243 (1995)
8. Friedman, N., Goldszmidt, M., Wyner, A.: Data analysis with Bayesian networks: A Bootstrap approach, Proc. 15th Conf. on Uncert. in Artif. Intell. (UAI'99), San Francisco: Morgan Kaufmann, 196-205 (1999)
9. Koivisto, M., Sood, K. Exact Bayesian structure discovery in Bayesian networks. Jour. of Mach. Lear. Res. **5**, 549-573 (2004)
10. Plummer, M.: JAGS: a program for analysis of Bayesian graphical models using Gibbs 701 sampling. Proc. 3rd Int. Work. Dist. Stat. Comp. (DSC 2003), Vienna, Austria, 1-10 (2003)
11. Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Torgerson, P., Furrer, R.: Data on Leptospira interrogans sv Pomona in Meat Workers in New Zealand. Data in Brief **13**, 587-596, ELSEVIER (2017)

Population Dynamics

Employment Uncertainty and Fertility: a Meta-Analysis of European Research Findings

Incertezza lavorativa e fecondità: una meta-analisi degli studi europei

Giammarco Alderotti, Daniele Vignoli and Michela Baccini

Abstract The impact of employment uncertainty on fertility represents a prominent topic in demographic research since ever. With the advent of the *Great Recession*, papers addressing this relation are now booming. Although uncertainty is usually deemed to be a negative condition, however, different fertility reactions are both advocated by sociological theories and supported by micro-level evidence, which is still fragmented and contradictory. In this article, we perform a meta-analysis (i.e., a quantitative literature review) of previous research findings for Europe in order to synthesize the existing literature and offer general conclusions about the (changing) size and direction of the impact of employment uncertainty on fertility across time and between countries.

Abstract *Il tema della relazione tra incertezza lavorativa e fecondità ha sempre avuto un ruolo centrale nella ricerca in ambito demografico, e la Grande Recessione lo ha reso ancora più nevralgico negli ultimi anni. Sebbene l'incertezza sia spesso considerata con accezione negativa, può infatti generare diversi tipi di reazione a livello di fecondità. Tali reazioni trovano supporto sia nelle teorie sociologiche, sia nell'evidenza empirica a livello micro, incompleta e talvolta contraddittoria. In questo lavoro, ci proponiamo di realizzare una meta-analisi (cioè una rassegna quantitativa della letteratura) degli studi europei, con lo scopo di sintetizzare i risultati sull'argomento e offrire conclusioni generali sul cambiamento della direzione e dell'impatto dell'incertezza lavorativa sulla fecondità nel tempo e tra paesi.*

Key words: economic uncertainty, fertility, family, employment, meta-analysis

[1] Giammarco Alderotti, University of Rome "La Sapienza"; giammarco.alderotti@uniroma1.it
Daniele Vignoli, Università degli Studi di Firenze; vignoli@disia.unifi.it
Michela Baccini, Università degli Studi di Firenze; baccini@disia.unifi.it

1 Introduction and objective

The role of economic uncertainty as a key driver of fertility decision-making has long been part of demographers and sociologists' research agenda. The notion of economic uncertainty refers to clarity, or the lack thereof, about future economic activities (Bloom, 2014; Moore, 2016). For individuals, economic uncertainty reflects the likelihood of experiencing adverse labour market conditions over the life course. The crisis of the fordist model during the 1980s led to a structural incapability of jobs creation in Europe and a dramatic increase of (especially youth) unemployment rates. On the other side of the Ocean, the United States showed successful occupational outcomes. This success was imputed by social and economic observers to the 'flexibility' of the North American labour market, as opposed to European ones that were described as too 'rigid', i.e., protecting excessively permanent jobs (Cutuli and Guetto, 2013). Consequently, in the last two decades of the past Century, European labour markets experienced a strong process of deregulation, leading to a significant change in the international division of labour. The "re-regulation" of the labour market and other aspects of the globalization wave (such as privatizations and liberalizations) generated an unprecedented level of structural uncertainty in contemporary societies (Mills and Blossfeld, 2005; Standing, 1997, 2014). Such condition affected many workers and has been shown to have an impact on family formation as well (Blossfeld et al., 2006; Esping-Andersen, 1999). Since the advent of the *Great Recession*, then, papers addressing the effects of economic uncertainty on fertility intentions and behaviours are booming.

The magnitude and direction of the relationship between economic uncertainty and fertility are debated, because both negative and positive effects are supported by sociological theories and empirical evidence. An important contribution comes from the work of Prija Ranjan (1999), who concluded that increasing uncertainty about the future income leads young adults to postpone irreversible, long-term decisions, such as childbearing. On the other hand, the work by Debra Friedman and colleagues (1994) advocated that uncertainty may have a positive effect on fertility: when a woman has limited chances in the labour market, she might choose the "alternative career" of mother. Empirical evidence on this topic is mixed too and, at times, contradictory: some studies pointed out that economic uncertainty negatively affects fertility (e.g. Pailhé, 2012, Sutela, 2012, Hofman, 2013), others found a positive effect (e.g. Gutierrez, 2008, Perelli-Harris, 2006, Kreyenfeld, 2009), while some others a non-significant effect (e.g. de Lange et al., 2014).

All things considered, there is a sensible need to get a closer understanding of previous research to draw general conclusions about the size and the direction of the impact of economic uncertainty on fertility in Europe. To this end, we performed a quantitative literature review, namely a meta-analysis (e.g., Matysiak and Vignoli, 2008), to synthesise, merge and interpret all the evidence coming from the available literature on the topic. This technique allows for a systematic analysis of results coming from heterogeneous studies, standardised for the country of interest, the method applied, the sample selected, the control variables, the period covered by the study, and so forth.

2 Data and method

Economic uncertainty is a very broad and multifaceted concept, and it cannot be directly observed and measured. Nevertheless, employment uncertainty, intended as experiencing adverse labour market conditions, can be directly observed. It translates into a feeling of economic uncertainty for individuals because flexible working arrangements are often related to economic penalties (Scherer, 2009). Furthermore, a growing number of studies pointed to a connection between economic uncertainty, in the form of unemployment, temporary work contracts, and unstable labour market situations with fertility (Mills et al., 2011). Hence, in this work, we focus on employment uncertainty, measured through individuals' employment condition and type of contract. In detail, we consider: (i) fixed-time contracts, (ii) part-time employment, and (iii) unemployment as "treatments" opposed to a permanent employment position.

Four inclusion criteria were used to select the articles that could enter our meta-sample. First of all, we included articles and book chapters, disregarding conference papers and working papers to ensure a high-quality meta-analysis. Next, qualitative works are excluded as they did not test the effect under interest from the statistical point of view. Third, we included only micro-level studies to avoid ecological correlation fallacy. Finally, we restricted the search to studies in European countries because they display an interesting variation in fertility and labour market patterns, while also sharing certain economic, social, and cultural characteristics that minimize heterogeneity.

We collected articles using the electronic database Scopus (www.scopus.com), which is the largest abstract and citation database of peer-reviewed literature, summing up to more than 60 million records. After the search through electronic database, we performed a backward research to find articles that were possibly not included in the database. Finally, we sent the list of articles retrieved to a group of experts of the field, asking them to check if any important contribution was missing. At the time of writing, 76 articles have been coded and included in the meta-sample.

We perform separate meta-analyses for every treatment considered (i.e., fixed-time contracts; part-time employment; and unemployment). For each treatment, we conduct specific analyses considering the key moderators: gender and welfare regimes. In fact, the impact of economic uncertainty on fertility has been shown to present different connotations between genders (Blossfeld et al., 2005). In addition, it is a well-established fact that the pressures towards increasing labour market liberalizations in Europe have been moderated by country-specific institutions (Esping-Andersen, 1999).

3 Preliminary results and next steps

Preliminary results about the relationship between fixed-time contracts and fertility show that, on average, employment uncertainty has a significant negative effect on fertility. The OR of having a child for people with a fixed-time job with respect to people with an unlimited-time contract is 0.91. Such negative effect is even stronger for women (OR 0.88, highly significant), while it is weaker, on average, for men (OR 0.98, still significant). Interestingly, there are also important differences depending on the welfare state. For example, the OR of having a child for fixed-time workers compared to unlimited-time workers goes from 0.92 in countries such as the Netherlands and the UK, to 0.86 in Southern Europe, while it is not even significant in Germany. As a second step, we studied the role of other covariates by performing meta-regressions. Results show that fixed-time employment is more detrimental to childless individuals' fertility with respect to those who already have a child, and that some characteristics of the study (e.g., controlling for perceived economic uncertainty) play significant roles.

The next step is to move toward a Bayesian approach. The Bayesian meta-analysis (e.g., Baccini et al., 2008) makes heterogeneity assessments more robust. As a last step, we aim to perform a network meta-analysis (NMA). The NMA has been proposed as an extension of pairwise meta-analysis to facilitate indirect comparisons of multiple competing treatments – here identified by different markers of economic uncertainty: fixed-time contracts, part-time employment and unemployment – that have not yet been studied in head-to-head studies. Compared with pairwise meta-analyses, network meta-analyses allow the visualisation of a larger amount of evidence, estimation of the relative effectiveness among all treatments, and rank ordering of the treatments (Tonin et al., 2017).

References

1. Baccini, M., Biggeri, A., Accetta, G., Kosatsky, T., Katsouyanni, K., Analitis, A., ... & Forsberg, B.: Heat effects on mortality in 15 European cities. Epidemiology, 19(5), pp. 711-719 (2008)
2. Bloom, N.: Fluctuations in Uncertainty. *Journal of Economic Perspectives*, *28*(2), pp. 153–176 (2014)
3. Blossfeld, H. P., Klijzing, E., Mills, M., & Kurz, K.: *Globalization, uncertainty and youth in society: The losers in a globalizing world*. Routledge (2006)
4. Cutuli G., Guetto R.: Fixed-Term Contracts, Economic Conjuncture, and Training Opportunities: A Comparative Analysis Across European Labour Markets (2013) European Sociological Review 29(3): pp. 616-629 (2013)
5. de Lange, M., Wolbers, M. H., Gesthuizen, M., & Ultee, W. C.: The impact of macro-and micro-economic uncertainty on family formation in The Netherlands. *European Journal of Population*, *30*(2), pp.161-185 (2014)
6. Esping-Andersen, G.: *Social foundations of postindustrial economies*. Oxford University Press (1999)

7. Friedman, D., Hechter, M., & Kanazawa, S.: A theory of the value of children. *Demography*, *31*(3), pp. 375-401 (1994)
8. Gutiérrez-Domènech, M.: The impact of the labour market on the timing of marriage and births in Spain. *Journal of Population Economics*, *21*(1), pp. 83-110 (2008)
9. Hofmann, B., & Hohmeyer, K.: Perceived economic uncertainty and fertility: Evidence from a labor market reform. *Journal of Marriage and Family*, *75*(2), pp. 503-521 (2013)
10. Kreyenfeld, M.: Uncertainties in female employment careers and the postponement of parenthood in Germany. *European Sociological Review*, *26*(3), pp. 351-366 (2009)
11. Matysiak, A., & Vignoli, D.: Fertility and women's employment: A meta-analysis. European Journal of Population/Revue européenne de Démographie, 24(4), pp. 363-384 (2008)
12. Mills, M., Rindfuss, R. R., McDonald, P., & Te Velde, E.: Why do people postpone parenthood? Reasons and social policy incentives. *Human reproduction update*, *17*(6), pp. 848-860 (2011)
13. Moore, A.: Measuring economic uncertainty and its effects. *Economic Record* (2016)
14. Pailhé, A., & Solaz, A.: The influence of employment uncertainty on childbearing in France: A tempo or quantum effect?. *Demographic research*, *26* (2012)
15. Perelli-Harris, B.: The Influence of Informal Work and Subjective Well-Being on Childbearing in Post-Soviet Russia. *Population and Development Review*, *32*(4), pp. 729-753 (2006)
16. Ranjan, P.: Fertility behaviour under income uncertainty. *European Journal of Population/Revue Européenne de Démographie*, *15*(1), pp. 25-43 (1999)
17. Scherer, S.: The social consequences of insecure jobs. Social Indicators Research, 93(3), pp. 527-547 (2009)
18. Standing, G.: Globalization, labour flexibility and insecurity: the era of market regulation. *European Journal of Industrial Relations*, *3*(1), pp. 7-37 (1997)
19. Standing, G.: Understanding the precariat through labour and work. *Development and change*, *45*(5), pp. 963-980 (2014)
20. Sutela, H.: Temporary Jobs and first child fertility in Finland. *Community, Work & Family*, *15*(4), pp. 425-450 (2012)
21. Tonin, F. S., Rotta, I., Mendes, A. M., & Pontarolo, R.: Network meta-analysis: a technique to gather evidence from direct and indirect comparisons. *Pharmacy practice*, *15*(1) (2017)

What Shapes Population Age Structures in the Long Run

Verso dove tende la struttura per età corrente

Gustavo De Santis and Giambattista Salinari

Abstract We present and test a hypothesis that, to the best of our knowledge, is new in the demographic field: the age structure of any population in any period tends towards a specific shape, which can be identified in advance. This "attractor" is the age structure of the current stationary population, which we label RAS, or "reference age structure". There is no mathematical demonstration for this tendency: however, we show that it exists in practice, measure the speed of the convergence (of the current on the reference age structure), and discuss the theoretical and practical utility of the notion.

Abstract *In questo paper dimostriamo che la struttura per età corrente evolve nel tempo subendo la costante attrazione della struttura per età della popolazione stazionaria corrente, che qui chiamiamo RAS ("reference age structure", o struttura per età di riferimento). La tendenza alla convergenza non può essere dimostrata matematicamente, ma avviene empiricamente: qui lo si dimostra, si misura la velocità di questo processo e si discutono alcune delle implicazioni teoriche e pratiche della nostra scoperta, che, a quanto ci consta, rappresenta una novità in demografia*

Key words: Age structure, Stationary population, ECM (Error Correction Model).

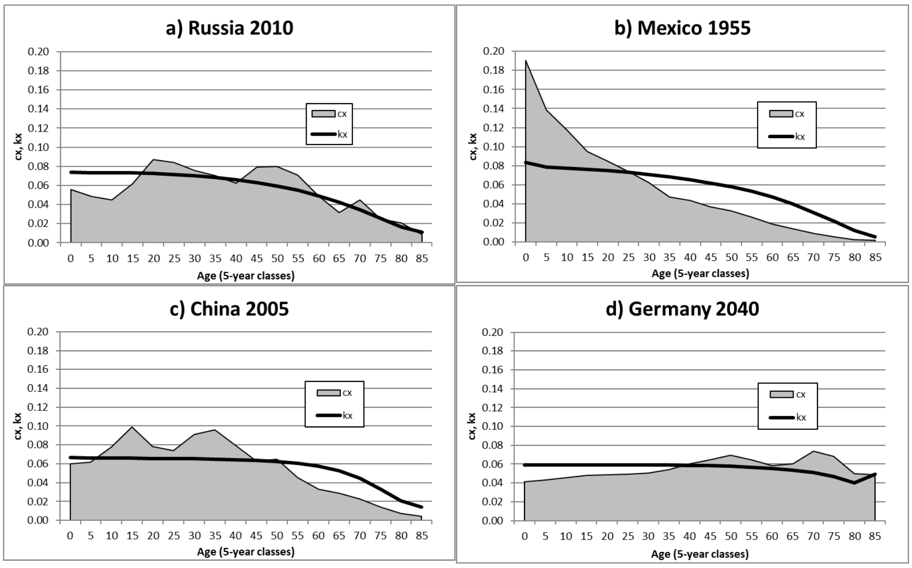
Two Age Structures: Actual and Stationary Populations

Let $c_{x,t}$ be the age structure of the population at time t, or current age structure. It is defined as the relative share of population aged x, P_x, to the total population P: $c_x=P_x/P$. Similarly, let $k_{x,t}$ be the age structure of the stationary population,

[1] Gustavo De Santis, University of Florence; email: gustavo.desantis@unifi.it;
Giambattista Salinari, University of Sassari; email: gsalinari@uniss.it.

calculated on the cross sectional life table in year t. If $L_{x,t}$ are the person-years lived at age x in year t and $T_{0,t}$ is their sum, or total number of person-years lived at all ages, the ratios $k_{x,t}$=$L_{x,t}$/T_0,t form the age structure of the stationary population in year t, which we also call "reference" age structure (or RAS) in this paper. Obviously, Σc_x=Σk_x=1. Selected examples of both are presented in Figure 1.

Figure 1: Actual (c_x) and reference (k_x) age structures of selected populations



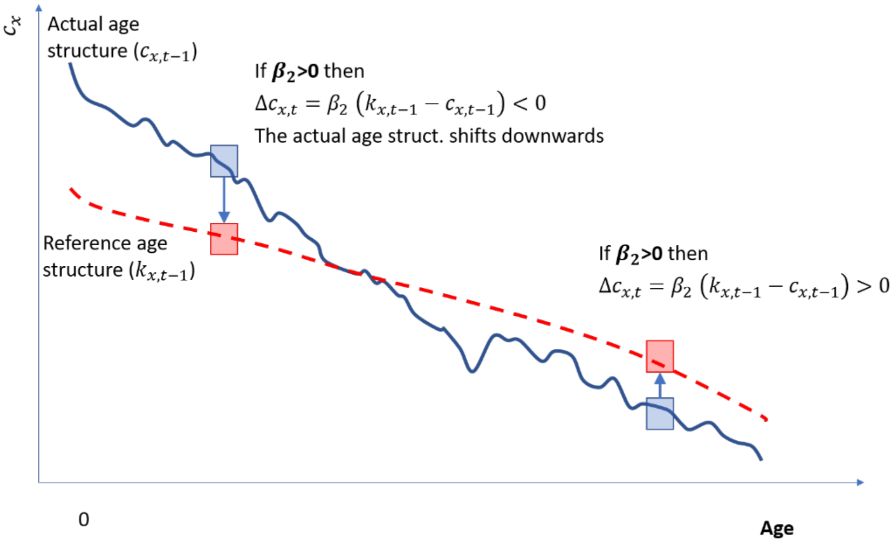
Source: UN (2017).

Our hypothesis is that the shares $c_{x,t}$ tend to move towards their reference counterpart $k_{x,t}$. To prove this, we run the following Error Correction Model (ECM) on UN (2017) data

$$\Delta c_{x,t} = \beta_0 + \beta_1 \, \Delta k_{x,t} + \beta_2 \, (k_{x,t-1}-c_{x,t-1}) + \varepsilon_{x,t} \qquad (1)$$

where $\Delta c_{x,t}$=$c_{x,t}$-$c_{x,t-1}$, $\Delta k_{x,t}$=$k_{x,t}$-$k_{x,t-1}$, and $\varepsilon_{x,t}$ is the error term. Our main object of interest is the coefficient β_2, which we expect to be significantly greater than zero. If this is true, the $c_{x,t}$ series tends to converge on the $k_{x,t}$ series. Figure 2, where we drew the profiles of the actual and of the reference age structure at time t-1 of a hypothetical population, gives a visual representation of what we expect will happen.

Figure 2: Hypothesised dynamic of the population structure


c_x
Actual age structure ($c_{x,t-1}$)
If β_2>0 then
$\Delta c_{x,t} = \beta_2 \left(k_{x,t-1} - c_{x,t-1}\right) < 0$
The actual age struct. shifts downwards
Reference age structure ($k_{x,t-1}$)
If β_2>0 then
$\Delta c_{x,t} = \beta_2 \left(k_{x,t-1} - c_{x,t-1}\right) > 0$
0
Age

Source: Illustrative data.

Data and Empirical Results

The main results of our model, based on the most recent UN (2017) data and estimates, are presented in Table 1. Note, first, that each country provides 204 observations: 17 five-year age classes (0-4 to 80-84; while the last, open-ended one, 85 and over, is not used in the estimates) for 12 five-year periods (1950 to 2015; this makes 13, but we "lose" one, because we work on differences).

The estimates of β_2 are positive and significant in all of the world and in almost all of its (sub)regions. The only exception is Middle Africa, and the reason why this happens is, we submit, the fact that this part of the world is still going through its demographic transition, which is strong enough to obscure the underlying tendency (convergence of c_x on k_x) that we are focusing on here. The speed of convergence ranges between 2% and 40%, depending on the region, and is about 6% at world level. In other words, on average, 6% of the difference between $k_{x,t}$ (reference age structure) and $c_{x,t}$ (actual age structure) observed in year t disappears 5 years later, because the current age structure tends to move in the "right" direction (see again Figures 1 and 2), which is precisely what we wanted to prove.

Table 1: Estimates of equation (1) by UN sub-regions

SubRegion	#Countries	Obs.	Intercept	β_1	β_2	R^2
World	201	41004	-2.36E-05	-0.261*	0.062*	0.06
			Africa			
Eastern Africa	20	4080	-8.50E-06	-0.183*	0.024*	0.02
Middle Africa	9	1836	-1.20E-05	-0.336*	0.002	0.01
Northern Africa	7	1428	-3.49E-05	-0.943*	0.078*	0.09
Southern Africa	5	1020	1.58E-06	-0.004	0.024*	0.03
Western Africa	16	3264	-7.30E-06	-0.380*	0.015*	0.02
			America			
Caribbean	17	3468	-6.62E-05	-0.814*	0.113*	0.12
Central America	8	1632	-7.48E-05	-0.783*	0.063*	0.14
Northern America	2	408	-7.84E-05	-0.920*	0.165*	0.14
South America	13	2652	-8.56E-05	-0.985*	0.078*	0.15
			Asia			
Eastern Asia	8	1632	-4.56E-05	-0.674*	0.182*	0.13
South Central Asia	14	2856	-1.09E-05	-0.634*	0.074*	0.06
South Eastern Asia	11	2244	8.02E-06	-0.130*	0.074*	0.09
Western Asia	18	3672	-5.86E-05	-1.257*	0.077*	0.06
			Europe			
Eastern Europe	10	2040	3.78E-05	-0.107	0.406*	0.23
Northern Europe	11	2244	4.25E-06	0.061	0.202*	0.13
Southern Europe	12	2448	-2.39E-05	-0.254*	0.188*	0.16
Western Europe	7	1428	-4.17E-05	-0.434*	0.317*	0.17
			Oceania			
Australia New Zealand	2	408	-5.42E-05	-0.536*	0.138*	0.13
Melanesia	5	1020	-4.09E-05	-1.038*	0.075*	0.14
Micronesia	3	612	-6.30E-05	-1.727*	0.089*	0.09
Polynesia	3	612	-4.26E-05	-0.676*	0.082*	0.11

Source: Own calculations on UN data. The parameters in red are non-significant (α=5%).

To the best of our knowledge, this tendency had thus far gone unnoticed in the literature. We venture that there are two main reasons why this happened. The first is that this convergence cannot be proved mathematically. Indeed, the reverse is true: there are theoretical cases (not discussed here) in which it can be proved that convergence does *not* take place. The second reason is that convergence is a constant, but weak force, as our estimates of Table 1 indicate, which is easily obscured by others, such as strong migration flows, or the demographic transition. In the long run, however, or with a sufficient number of observations on data of reasonable quality, this tendency emerges and can be detected.

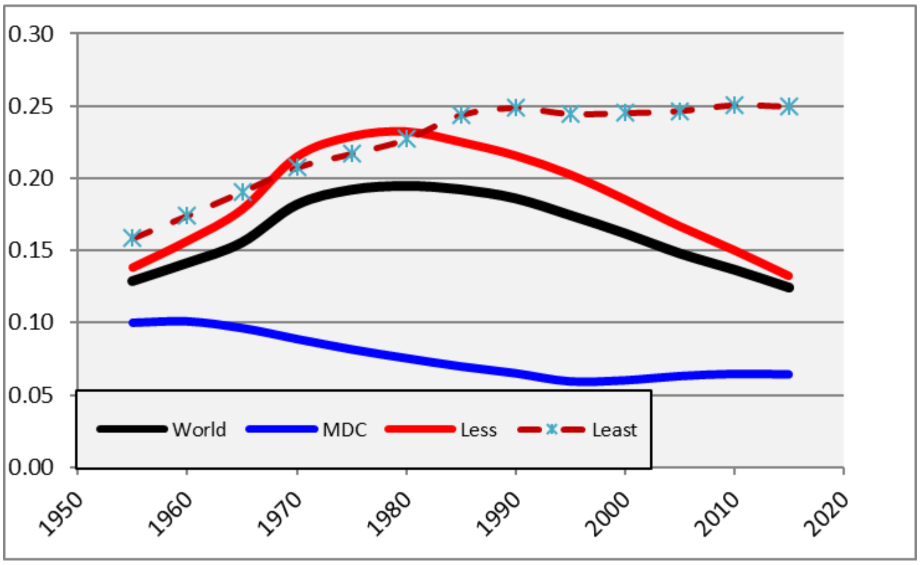
On the Distance Between the Actual and the Reference Age Structure

The distance between the actual ($c_{x,t}$) and the reference age structure ($k_{x,t}$) can be measured with the index of dissimilarity D_t, which indicates what share of the population should be in a different age class to result in a perfect coincidence between the two (c_x and k_x), in year t

$$D_t = \frac{1}{2}\sum_c \left|c_{x,t} - k_{x,t}\right| = \frac{1}{2}\sum_c \left|\frac{P_{x,t}}{P_t} - \frac{L_{x,t}}{T_{0,t}}\right| \qquad (2)$$

By construction, D_t ranges between 0 and 1, but in practice, it is virtually impossible in this kind of application (population shares by age classes) to find D_t outside the range 5-50% (discussion skipped here). Figure 3 shows that D_t (dissimilarity, or distance) has not decreased monotonically in the past 65 years, as our model (1) predicts: we argue that this is due to the fact that, in the "short" run, other forces prevail, first of all the demographic transition.

Figure 3: Dissimilarity index D_t between the actual (c_t) and the reference (k_t) age structure in selected world areas, 1950-2015



Note: MDC=More Developed Countries; Less/Least=Less/Least Developed Countries
Source: UN (2017) and own calculations.

Indeed, in more developed countries (whose demographic transition took place in the late 19th-early 20th century, i.e., left of the time scale displayed in Figure 3), D_t does decrease over time and is today about as low as it can reasonably be. In less developed countries, D_t *increased* until the 1980s (precisely because of the biases produced by the ongoing demographic transition following WW2) and decreased later. In the least developed countries (basically, Sub-Saharan Africa) convergence is not yet taking place, and the two age structures ($c_{x,t}$ and $k_{x,t}$) remain far from each other. (Convergence will occur in the next decades, according to the UN Projections, but this is not shown or discussed here.)

So what?

Our finding has several important implications, which we cannot fully develop here, for reasons of space. For instance, it affects the ongoing debate on the main causes of population ageing: is it (mainly) due to low fertility or low mortality? Our answer is simple: as the current age structure $c_{x,t}$ tends to "move" towards its reference counterpart $k_{x,t}$, which depends exclusively on survival, lower mortality is the underlying force that (slowly) drives the process of population ageing. Deviations from this path can be important, and they can last for some years, or even decades, but eventually the "attraction" of $k_{x,t}$ (i.e., the dominant role of the mortality regime) prevails.

A second application (among others) is in the field of pension systems. Our finding suggests that pension systems should be designed giving emphasis to the reference age structure $k_{x,t}$ and to its evolution over time, as the current age structure $c_{x,t}$ will eventually converge towards it. This takes time, admittedly, but pension systems are, or at least should be, intended to last for very many years: to base them on the reference age structure $k_{x,t}$ would make them much more long-lasting and robust than they usually are.

Acknowledgements

We thank the CREW group for early feedback on this paper (CREW=Care, Retirement and Wellbeing of Older People across Different Welfare Regimes; https://crew-more-years-better-lives.org/).

References

1. UN. 2017. World population prospects, Population Division, Department of Economic and Social Affairs, New York

The impact of economic development on fertility: a complexity approach in a cross-country analysis

L'impatto dello sviluppo economico sulla fecondità: l' approccio della complessità in un'analisi cross-country

Niccolò Innocenti, Daniele Vignoli and Luciana Lazzeretti

Abstract This work focuses on the role of economic indicators on fertility from a cross-country comparative perspective. It relies on a new indicator of the economy of a country and on its perspective of future development – i.e. the economic complexity. We apply a Structured Equation Model to data on 70 countries from 1964 to 2015 to assess the impact of economic complexity on economic growth (Gross Value Added, employment, innovation) and, then, on total fertility. Preliminary results show that the relation between economic complexity and economic growth is positive for all countries, while a U-shaped relation between the economic complexity and fertility seems to emerge during the analysed time-span.

Abstract *Questo lavoro si concentra sul ruolo degli indicatori economici nello spiegare le variazioni di fecondità dei paesi basandosi su un nuovo indicatore in grado di indicare anche le future prospettive di sviluppo del paese, cioè la complessità economica. Abbiamo stimato un modello ad equazioni strutturali sui dati di 70 paesi per il periodo 1964-2015 per valutare gli effetti della complessità economica sulla crescita economica (valore aggiunto, occupazione, innovazione) e successivamente sul tasso di fecondità totale. I risultati preliminari mostrano una relazione positiva tra complessità e economica e crescita economica, mentre sembrano suggerire un modello a U per la relazione tra complessità economica e fecondità durante il periodo analizzato.*

Niccolò Innocenti, University of Florence, Department of Economics and Management; email: niccolo.innocenti@unifi.it

Daniele Vignoli, University of Florence Department of Statistics, Computer Science, Applications "G. Parenti"; email: vignoli@disia.unifi.it

Luciana Lazzeretti, University of Florence Department of Economics and Management; email: luciana.lazzeretti@unifi.it

Key words: Economic Development, Fertility, Cross-country analysis.

1 Introduction

While the first decade of the new millennium witnessed moderately increasing fertility rates in some countries across Europe, since the advent of the Great Recession policy makers in many countries see their current fertility levels as worryingly low. The social, economic, and demographic implications of persistently low fertility rates are largely known: the population age structure is impaired, which puts the whole welfare system under pressure and impacts the rules governing transfers between generations (Bryant, 2007; Dribe et al., 2017). A great share of studies focused on the relation between income or wellbeing and total fertility. Results, at the end, are in two different directions. In fact, there is evidence both of a positive impact of income on fertility and of a negative impact depending on different factors, such as the level of economic development of the country and of the positive or negative trend of such development (Myrskyla et al., 2009; Wang and Sun, 2016).

Our work relies on a new indicator of the economy of a country and on its perspective of future development – i.e. the economic complexity (EC). This recent strand of research, firstly developed by Hidalgo and Hausmann (2009), sustains a view of economic growth and development that gives a central role to the complexity of a country's economy. The idea is that a country with higher economic complexity will perform better in term of future economic growth (Gross Value Added – GVA, employment, innovations) (Hartmann et al., 2017). An important factor that leads our interest on EC as a determinant of fertility is that it includes many dimension often considered as important factors able to influence the fertility of a country, more than the already mentioned economic dimension, such as the accumulation of human capital or women's emancipation, features often associated to a lower fertility (Lehr, 2003). Put directly, we argue that EC could be used to predict the fertility development of a country. The complexity indicator is constructed following the approaches and methods proposed by Hidalgo and Hausmann (2009), using export data. We rely on the idea that the economic complexity of a country depends on the goods that are produced and thus exported, taking into account the relative composition of the export baskets of all other countries. More complex economies tend to produce more rare products in term of ubiquity and that relatively few countries are able to produce, due to the need of producing many other related/diverse products. The countries with a more complex economy are then able to use their comparative advantage (or their oligopoly) to perform better in term of rents and also of opportunities to diversify in more and more rare and complex products. On the contrary, countries that produce less complex products deal with a higher competition and thus tend to have less opportunities both in term of rents and in term of comparative advantage to diversify in more complex activities.

Importantly, different levels of economic complexity imply different job options, educational qualifications, and also different behaviour in private life that are likely to affect also family formation, as well as fertility timing and quantum.

2 Methodology

The paper presents a longitudinal cross-country analysis of 70 countries from 1964 to 2015. The country's total fertility – i.e. the Total Fertility Rate (TFR) – is drawn from the World Development Indicators. It is customarily defined as the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.

The data used to construct the complexity index are about the export of all the analyzed countries during the same period (1964-2015). To construct our index of EC we followed the methods proposed by Hidalgo and Hausmann (2009) and refined by Balland and Rigby (2017). We used, international trade data with products disaggregated according to the Standard International Trade Classification (SITC) at the 4-digit level and interpret them as a bipartite network.

We used export data as an adjacency matrix M*cp*, where M*cp* is equal to 1 if the country *c* is an important exporter of product *p* and equal to 0 if this is not the case. The country *c* is an important exporter of the considered product *p* and, thus, has a Revealed Comparative Advantage (RCA) if the share of product *p* in the export basket of the country *c,* divided by the share of product *p* in world trade, is higher than a threshold value *x*.

Mathematically:

$$\mathrm{RCA}cp = (\mathrm{E}cp/\mathrm{E}c)/(\mathrm{E}p/\mathrm{E}) > x$$

We then followed the method of reflections (for a detailed description of the properties see Hidalgo and Hausmann (2009)), using the RCA to compute the two components to construct the ECI, namely the diversity (in term of products) of each countries and the ubiquity (of the products produced by the country). After a series of *n* iterations of these measures at both products and country level we ended with a stable measure of ECI (we stop the iteration when the correlation with the previous is near 1, meaning that there are no significant information to capture by an additional use of the method of reflections).

The ECI is computed for each year of the period under investigation and for each country of the sample. Than an econometric analysis is performed to assess the relation among the ECI and the TFR. The method proposed is the use of structural equations to assess the impact of ECI on economic growth (GVA, Employment etc.) and then on TFR. Finally, we performed a robustness check to assess the direct relation among ECI and TFR.

The analysis will be performed using the whole sample as well as coherent sub-group, divided by economic development of countries from high, to low-income countries.

3 Results and preliminary conclusions

Preliminary results suggest a positive correlation between ECI and economic growth during the period under analysis. Interestingly, while the relation between ECI and economic variables is positive for all samples, when we analyze the impact on TFR the results show a U-shaped relation between ECI and TFR during the analyzed period. These results are in line with previous studies on the role of economic indicators on fertility. There is a clear association among EC of countries and fertility: in low-income and lower middle-income countries (associated also with an extremely low rank in terms of EC) there is a negative relation between EC growth and fertility. When the focus is on high-income countries, the results clearly show a positive correlation between EC growth and fertility rates.

Our findings are in line with those of Myrskyla et al. (2009), claiming that there is a reversal in fertility decline associated with human development, and this change i could be driven by the economic complexity, conceived as a driver of economic and human development of countries. This implies that highly developed countries benefit, in terms of fertility, by a further complexification of their economy.

The U-shaped pattern is related to the economic transition of countries in different stages of development. In low-income countries a higher level of complexity facilitates women's emancipation, a pursuit of higher education of individuals and women's entry into the labor market, leading to fertility reduction. In high-income countries a more complex economy is associated with economic opportunity, well-being and lower levels of economic uncertainty, leading to an increase of total fertility.

We continue the analysis by estimating a Structural Equation Model on the reciprocal relations between EC, economic growth and fertility.

4 Citations and References

1. Balland, P.A., Rigby, D.: The geography of complex knowledge. Econ. Geogr. 93 (1), 1–23 (2017)
2. Bryant, J.: Theories of fertility decline and the evidence from development indicators. Popul. and Dev Rev, 33 (1), 101-127 (2007)
3. Dribe, M., Breschi, M., Gagnon, A., Gauvreau, D., Hanson, H., Maloney, N., ... & Vézina, H.: Socio-economic status and fertility decline: Insights from historical transitions in Europe and North America. Pop. Stud., 71 (1), 3-21 (2017)
4. Hidalgo, C., Hausmann, R.: The building blocks of economic complexity. Proc Natl Acad Sci, 106 (26), 10570-10575 (2009)
5. Hartmann, D., Guevara, M., Jara-Figueroa, C., Aristarán, M., Hidalgo, C.: Linking economic complexity, institutions, and income inequality. World Dev, 93, 75-93 (2017)
6. Lehr, C.: Fertility and education premiums. J Popul Econ, 16 (3), 555-578 (2003)
7. Myrskyla, M., Kohler, H., Billari, F.: Advances in development reverse fertility declines. Nature, 460 (7256), 741-743 (2009)
8. Wang, Q., Sun, X.: The Role of Socio-political and Economic Factors in Fertility Decline: A Cross-country Analysis. World Dev, 87, 360-370 (2016)

Acknowledgements

We acknowledges the support of the ERC Grant EU-FER “Economic Uncertainty and Fertility in Europe”, grant agreement n°72596 (PI: Daniele Vignoli)

A Probabilistic Cohort-Component Model for Population Forecasting – The Case of Germany

Un Modello a Componenti di Coorte per Previsioni della Popolazione – Il Caso della Germania

Patrizio Vanella[1] and Philipp Deschermeier[2]

Abstract The future development of population size and structure is of undeniable importance since planning in many areas of politics and business is conducted based on expectations about the future makeup of the population. Countries with both decreasing mortality and low fertility rates, as is the case for most countries in Europe, are in urgent need of adequate population forecasts to identify future problems for social security systems and overall macroeconomic development. The labor market will especially be affected by the retirement of the *baby boomer generation* and the resulting shortage of young skilled workers. This contribution proposes a stochastic cohort-component model that uses simulation techniques based on stochastic models for fertility, migration and mortality to forecast the population by age and sex. The results provide detailed insight into the future population structure, disaggregated into both sexes and 116 age groups. Moreover, the uncertainty in the forecast is quantified as prediction intervals for each subgroup. The underlying models for forecasting the demographic components have been developed in earlier studies and rely on principal component time series models.

Abstract *Il futuro sviluppo della grandezza e della struttura della popolazione è di importanza innegabile, poiché la pianificazione in molte aree di politica ed economia è condotta sulla base di aspettative sulla composizione futura della popolazione. Paesi con tassi di mortalità e fertilità in calo, fenomeno comune nella maggioranza dei paesi d'Europa, hanno un urgente bisogno di previsioni adeguate della popolazione per identificare i problemi futuri per il sistema d'assicurazione sociale e per lo sviluppo macroeconomico generale. Il mercato del lavoro sarà influenzato in particolare dai pensionamenti della generazione baby boom e dalla risultante scarsità di lavoratori giovani e qualificati. Questo studio propone un modello a componenti di coorte che usa tecniche di simulazione a base di modelli stocastici per fertilità, migrazione e mortalità, per predire la popolazione per età e sesso. I risultati offrono informazioni dettagliate sulla struttura della popolazione nel futuro, disaggregate per sesso e 116 gruppi d'età. Inoltre, l'incertezza della previsione è quantificata attraverso intervalli di previsione per ogni sottogruppo. I sottostanti modelli per predire le componenti demografiche sono stati sviluppati in precedenti studi e usano modelli di serie storiche (time series) con componenti principali.*

Key words: Population Forecasting, Stochastic Simulation

1 Future Development of the Demographic Components

We propose a population forecast based on a probabilistic cohort-component model developed through the combination of Vanella's (2017a: 543-549) model for forecasting age- and sex-specific survival rates *(ASSSRs)*, the forecast model for age-, sex- and nationality-specific net migration *(ASNSNM)* by Vanella and Deschermeier (2017: 6-23) and Vanella's (2017b: 11-24) model for forecasting age-specific fertility rates by sex of newborns *(ASSFR)*. Each partial model generates 10,000 scenarios stochastically by simulation of Wiener processes[3], which represent 10,000 scenarios for the future development of the population in Germany by age (*0-115*) and sex (*binary*).

Vanella (2017b: 21-22) incorporated the effect of international migration on fertility in his fertility forecast model. The inclusion of this fertility model into the population forecast model proposed here provides an opportunity to quantify the effects of international migration on the development of the population not only directly through

[1] Patrizio Vanella, Gottfried Wilhelm Leibniz Universität Hannover – Center for Risk and Insurance; email: pv@ivbl.de
[2] Philipp Deschermeier, Institut Wohnen und Umwelt Darmstadt; email: p.deschermeier@iwu.de
[3] See Vanella 2017c: 13 for an explanation of Wiener processes.

the migration process itself but also indirectly by altering the number of births due to the different reproductive behavior of migrants relative to that of the native population.

We now describe the procedure for population forecasting with our model. Let $P_{x,y,g,t}$ denote the population aged *x* years at the end of year *y* for sex *g* in trajectory *t*. The population update is performed through the following step-wise process.

Step I:
The forecast begins with an adjustment of the base population with regards to international migration flows in the first forecast year *y+1***.** The addition of international net migration aged *x+1* years of sex *g* in year *y+1* ($M_{x+1,y+1,g,t}$) to $P_{x,y,g,t}$ leads to the hypothetical subpopulation $\tilde{P}_{x+1,y+1,g,t}$ at the end of year *t+1* without any deaths:

$$\tilde{P}_{x+1,y+1,g,t} = P_{x,y,g,t} + M_{x+1,y+1,g,t}.$$

Step II:
The number of survivors from $\tilde{P}_{x+1,y+1,m,t}$ at the end of *y+1* are calculated through multiplication with the age- and sex-specific survival rate (*ASSSR*) $s_{x+1,y+1,g,t}$ for persons aged *x+1* years of sex *g* in year *y+1* and in trajectory *t*:

$$P_{x+1,y+1,g,t} = \tilde{P}_{x+1,y+1,g,t} * s_{x+1,y+1,g,t}.$$

Step III:
The mean female population in *y+1* in the reproductive age group is approximated:

$$F_{x,y+1,w,t} = \sum_{x=14}^{52} \frac{P_{x-1,y,w,t} + P_{x,y+1,w,t}}{2}.$$

This approximation presumes constant migration and mortality during a year because we rely on annual data. This approach is mathematically practical; moreover, we believe the possible error arising from this assumption is extremely small and therefore produces no fundamental error in the future population estimation.

Step IV:
The sex-specific live births $B_{y+1,g,t}$ are estimated:

$$B_{y+1,g,t} = \sum_{x=14}^{52} F_{x,y+1,w,t} * f_{x,y+1,g,t},$$

where $f_{x,y+1,g,t}$ denotes the ASSFR for females aged *x* years in year *y+1* conceiving babies of sex *g* in trajectory *t*.

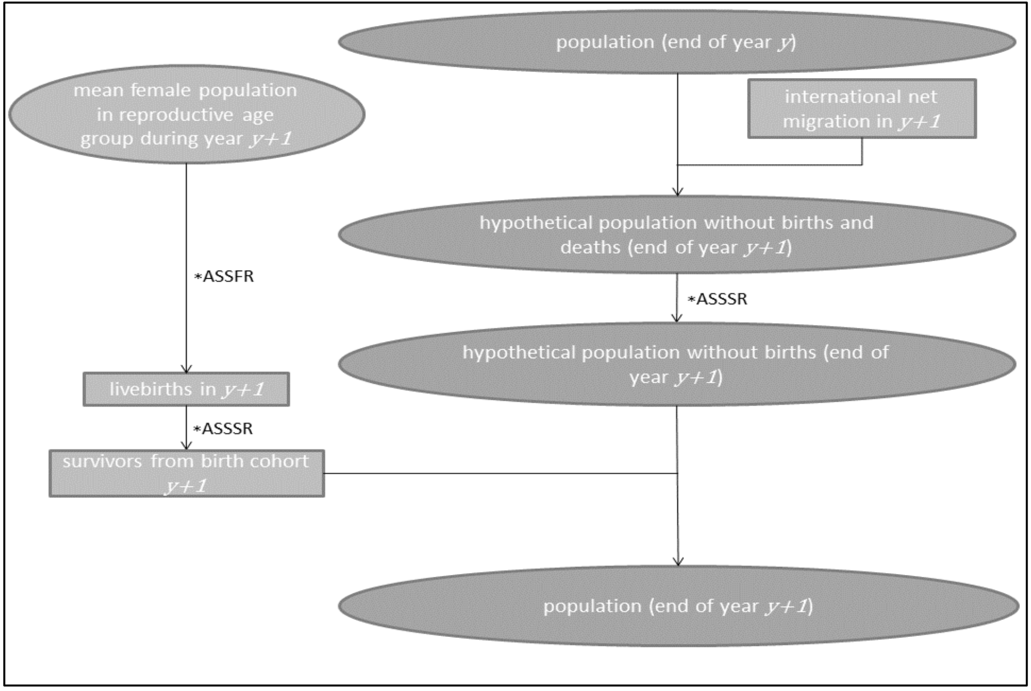
Step V:
The number of survivors among the children born in *y+1* is calculated:

$P_{0,y+1,g,t} = B_{y+1,g,t} * s_{0,y+1,g,t}$.

In this way, the population by sex and age in year *y+1* in trajectory *t* is obtained.

The algorithm is illustrated in Figure 1.

Figure 1: Process of Annual Population Update


population (end of year *y*)
mean female population in reproductive age group during year *y+1*
international net migration in *y+1*
hypothetical population without births and deaths (end of year *y+1*)
*ASSFR
*ASSSR
hypothetical population without births (end of year *y+1*)
livebirths in *y+1*
*ASSSR
survivors from birth cohort *y+1*
population (end of year *y+1*)

Source: Own design

This process is then used to stochastically forecast the population by sex and age until the year 2040. The future development of the population is forecast by estimating the future course of the demographic components of fertility (through ASSFRs), migration (ASNSNM), and mortality (ASSSRs). The underlying variables are simulated indirectly on the basis of the stochastic trajectories of the PCs, which were derived in earlier studies. A quasi-two PC time series model, as proposed by Vanella (2017a: 543-549), is used to forecast the logistically transformed ASSSRs. The first PC is a general mortality index, similar to the classic Lee-Carter index. The second PC is a behavioral index representing differences in mortality between the two sexes associated with nutritional and smoking behavior. Migration flows are accounted for via the quasi-two PC forecast model for ASNSNM proposed by Vanella and Deschermeier (2017: 15-22). In this case, the first PC is a labor market index and the second PC addresses the migration resulting primarily from humanitarian or economic crises. The fertility model builds upon the results of the mentioned migration model. This component is forecast by the quasi-three PC model proposed by Vanella (2017b: 11-24). In this case, the first PC is associated with the tempo effect in fertility, i.e., the postponement of births from younger ages to older ages. The second PC is a quantum index used to address general trends in the quantity of reproductive behavior in a society. The third PC considers the impact of the migration level on fertility and is derived via econometric specification of the influence through the second PC of the migration model proposed by Vanella and Deschermeier (2017: 12-17).

2 Population Development in Germany until 2040

The combination of the resulting trajectories for the demographic components results in a probabilistic cohort-component model for forecasting the age- and sex-specific population for the ages 0-115 years. The initial population for the forecast is the age- and sex-specific population reported by Destatis for December 31, 2015. Population numbers for ages 100 years and over are not available in detail but are rather aggregated into an upper age group. Thus, we approximated the population in this age group through an own population update based on the death counts by age and sex.

In general, we observe greater uncertainty for males. Whereas the retirement-age population can be predicted relatively well, the uncertainty in the future working-age population is especially large for males due to the high uncertainty in the migration forecast for males (Vanella, Deschermeier 2017: 17-22). The uncertainty in the population of persons under 25 years of age arises from the fact that this portion of the population has not been born yet. Specifically, the uncertainty of persons under 25 years of age arises from migration, from the mortality of persons "below 1 year" and from uncertainty due to fertility.

The aging of the population is of high social and economic importance, as was stated earlier. Therefore, in addition to the overall age structure, the median age of the male and female populations is considered as a summary indicator for the future age schedule of the population. The median age of the population can be obtained from the simulation results because it is the exact age that cuts the population in half. This computation for all 10,000 trajectories can be used to extract PIs for the median age.

We observe a rejuvenation effect for the upcoming years due to high net migration during this period. The high net migration rates around the year 2015 combined with the high forecast values for the upcoming years leave a mark in the age structure of Germany. This can be seen in the age structure for the male and female populations in the year 2040. By that time, the majority of the population that immigrated during the high influx phase will be approximately 50 years old, while the baby boomer generation will be in their seventh decade of life. Over the forecast horizon, the median age traces this development by a rejuvenation effect for men and women. The probable decrease in the number of births after the middle of the 2020s and decreasing net migration and mortality (Vanella 2017a: 550) lead to an aging of the population structure, as represented by the increasing median ages after that point. Since a larger portion of migrants is male (Vanella, Deschermeier 2017: 20-22), the rejuvenation is stronger for males than for females.The driving factor behind the presented results on the population development is net migration. Following the net migration in 2015, the first years of the forecast result in high and above average values of net migration. In the long run, net migration is expected to decrease to 212,419 in 2040. Counterbalancing net migration, we take a look at the probabilistic forecast of Germany's natural growth until 2040, which is simply derived from the birth and death forecasts. Following the trend of the recent past, the number of births will continue rising until the mid-2020s. However, as the peek remains below 900,000 births per year for the median scenario, this development cannot be labeled a second baby boom. In comparison, the number of livebirths in the 1950s and 1960s exceeded 1.1 million annually (GENESIS-Online Datenbank 2018). Due to the shift in the age structure, the number of potential mothers is expected to shrink at some point, resulting in decreasing birth numbers in the long run. The aging of the population leads to a steady increase in the number and rate of older people, who have a higher probability of dying in a given year. This results in a steady increase in the number of deaths until 2040. As the median of the simulation is greater than 900 thousand persons in each year of the forecast, the natural population development until 2040 almost certainly will be negative during the forecast horizon. Therefore, a lack of high net migration would naturally lead to a decrease in population in Germany. The connection between natural population growth and net migration is therefore obvious since an old population with low fertility requires positive net migration to avoid shrinking and overaging. In the median trajectory, the natural decrease in the population will grow annually, stressing the importance of positive net migration, especially in the younger ages, to fill the shortages occurring in the labor market. As we will show by some important measures, our model provides a wide range of detailed analyses targeting specific topics of interest. The forecast results offer the possibility for a wide range of future studies, e.g., analyzing the effects of population changes on social security, the labor market or housing demand.

References

GENESIS-Online Datenbank 2018: Lebendgeborene: Deutschland, Jahre, Geschlecht. In: www-genesis.destatis.de. URL: https://www-genesis.destatis.de/genesis/online/data;jsessionid=C4FF958DD4F6A3F4DB5C783F4B2D2890.tomcat_GO_2_1?operation=abruftabelleAbrufen&selectionname=12612-0001&levelindex=1&levelid=1515759981915&index=1, 10.01.2018.

Vanella, Patrizio 2017a: A Principal Component Model for Forecasting Age- and Sex-Specific Survival Probabilities in Western Europe. In: German Journal of Risk and Insurance 106,5: 539–554 [10.1007/s12297-017-0393-y].

Vanella, Patrizio 2017b: Age- and Sex-Specific Fertility in Germany until the Year 2040 – The Impact of International Migration. Hannover Economic Papers 606. Hannover: Gottfried Wilhelm Leibniz Universität Hannover, School of Economics and Management.

Vanella, Patrizio 2017c: Stochastische Prognose demografischer Komponenten auf Basis der Hauptkomponentenanalyse. Hannover Economic Papers 597. Hannover: Gottfried Wilhelm Leibniz Universität Hannover, School of Economics and Management.

Vanella, Patrizio; Deschermeier, Philipp 2017: Ein stochastisches Prognosemodell internationaler Migration in Deutschland. Hannover Economic Papers 605. Hannover: Gottfried Wilhelm Leibniz Universität Hannover, School of Economics and Management.

Mortality trends in Sardinia 1992-2015: an ecological study

Evoluzione della mortalità in Sardegna 1992-2015: uno studio ecologico

Vanessa Santos Sanchez, Gabriele Ruiu Marco Breschi, Lucia Pozzi

Abstract Sardinia has been for long times under the lens of biologists, demographers and social scientists for the high longevity of the population of some of its territories. However, and in some sense paradoxically, the island represents an ideal context to study the effect on human mortality produced by the environmental deterioration caused by the settlement of heavy industry. By the means of a spatial analysis we found that the largest increase in mortality rates in the period 1992-2015 are indeed in industrial areas.

Abstract *La Sardegna è stata per lungo tempo analizzata in ambito demografico, biologico e sociologico a causa della longevità registrata nelle popolazioni di alcuni suoi territori. Paradossalmente, questa regione rappresenta però un contesto ideale per studiare l'effetto prodotto sulla salute umana dalla presenza di industria pesante. Usando tecniche di analisi spaziale, si mostra infatti che il più alto incremento nei tassi di mortalità nel periodo 1992-2015 ha caratterizzato proprio le zone industrializzate.*

Key words: Mortality, Spatial epidemiology, Sardinia.

[1] Vanessa Santos Sanchez, Department of Economics and Business; University of Sassari; email: vanesasantossanchez@gmail.com; Corresponding author.

Gabriele Ruiu, Department of Economics and Business; University of Sassari; email: gruiu@uniss.it

Marco Breschi, Department of Economics and Business; University of Sassari; email: breschi@uniss.it

Lucia Pozzi, Department of Economics and Business; University of Sassari; email: lpozzi@uniss.it

1 Introduction

[1] coined the expression "blue zone of longevity" (from hereon BZL) for indicating a rather limited and homogenous geographical area where the population shares the same lifestyle and environment and its longevity has been proved to be exceptionally high. Ogliastra (South-Eastern Sardinia, Italy) together with a small number of regions around the World has been acknowledged as a BZL. The peculiar characteristic of the blue zone individuated in Sardinia, is that among centenarians the female/male ratio is about 1.34 which is a value considerably lower than in the rest of Italy ([2]). This has obviously attracted many research efforts for trying to explain the extraordinary longevity of males ([3]) whilst, at least in recent time, less attention has been devoted to analysing the overall evolution of mortality in Sardinia. Generally, the Italian public considers Sardinia a place where the environmental conditions and the healthy diet have favoured male longevity. However, this is far from being true at least in some zones of the island. In particular, Sardinia represents an ideal context for studying the effect that highly polluting industrial sites, mining sites and military bases play on human health. To our knowledge, only [4] analysed the Sardinian age-standardized mortality rates, at the municipal level, for a quite short period (1997-2001) finding that among males it was higher than in Italy (84.4 vs 80.8) while the reverse occurred in females (50.9 vs 52.0). They also find that the industrial sites of Porto Torres (chemical plants were established in this Northern Sardinian city during the sixties) was characterized by an excess mortality for respiratory diseases, diseases of the digestive tract, liver cancer and lymphohaemopoietic cancer. An anomalously high incidence of lymphohaemopoietic cancer was found also among the population living near military sites (La Maddalena, Salto di Quirra). An excess of deaths for respiratory related problems was also found in other industrial zones (Portoscuso and Sarroch). We enlarge their analysis to capture the evolution of mortality from 1992 to 2015 and applying a spatial methodology originally proposed by [5]. Extending the time period allows establishing when the pollution related to industrial activity which initially hit men, because of the fact that the workforce in these sectors is almost entirely composed by males, has begun to affect also female population, thus denoting a deterioration of the environmental conditions.

1.1 Data and Methods

An ecological study of small areas was carried out for the 377 Sardinian municipalities existing in 2015, in the periods 1992-1997, 1998-2003, 2004-2009 and 2010-2015. Individual death entries for the period 1992-2015, broken down by municipality and sex, were used as case source.

Municipal populations, broken down by age group (20 five-years groups) and sex are obtained for each year. The person years for each period were calculated by adding the population of each year.

Mortality and population data have been derived from ISTAT database.

We used the 2001 deprivation index created by [6] as indicator of socioeconomic level. This index classifies municipalities into 5 levels, according to several variables: (i) low level of education, (ii) unemployment, (iii) non-home ownership, (iv) one parent family and (v) overcrowding. This classification is based on the quintiles of the distribution of factor scores, where level 1 municipalities are the richest ones and level 5 municipalities the least rich ones.

To calculate the number of expected cases, overall Sardinian specific age group, sex and period mortality rates were multiplied by each municipal person-years for the same age, sex and period pattern. Standardized mortality ratios (SMRs) were calculated as the ratio of observed to expected deaths.

Smoothed municipal relative risks (RRs) with their corresponding 95% credibility intervals, were calculated using the conditional autoregressive model proposed by [5]. This model fits a Poisson spatial model with two types of random effects, a non structured effect that takes into account the municipal heterogeneity, and a structured effect, the spatial term, that considers municipal contiguity. To define area contiguity we used the adjacent municipal boundaries.

The model takes the following form:

$$O_i \sim Po(E_i \lambda_i)$$

$$\log(\lambda_i) = \alpha + h_i + b_i$$

Where λ_i is the RR in area i, O_i is the number of observed cases, E_i is the number of expected cases, α is the intercept, h_i is the municipal heterogeneity and b_i is the spatial term.

To analyze the effect of deprivation on mortality, the index proposed by [6] was included in the models as a covariate.

The Bayesian estimation of the models was obtained using Markov Chain Monte Carlo (MCMC) simulation methods, through the Gibbs Sampling algorithm via free distribution software WinBUGS. Convergence of the estimators was achieved before 100.000 iterations for three Markov chains, with a burn-in of 10.000 iterations. The convergence was ensured by the algorithm proposed by [7] and the effective sample size of chains.

The free software R was used to create municipal maps of SMRs, smoothed RR estimates and posterior probabilities that smoothed RR was greater than one (PRPs). To calculate PRPs we used the criterion proposed by [8], considering PRPs greater than 0.8 as statistically significant.

2 Results

From 1992 to 2015 a total of 343,584 deaths was registered in Sardinia, 182,993 in men (53.3%) and 160,591 in women (46.7%).

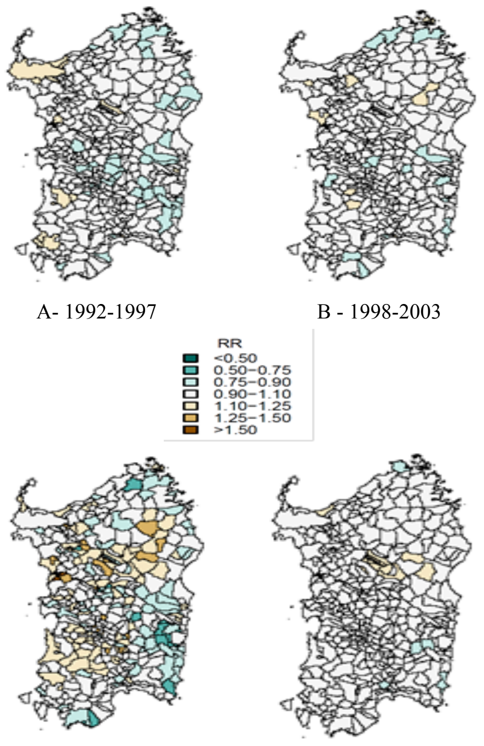
We did not observe any prevailing geographical mortality pattern in RRs for mortality throughout the 24 years of the study in either women or men.

For men, in the third period, municipalities in the northeastern provinces of Sassari, Olbia and Nuoro show an excess of mortality, with some RRs greater than 1.50 (municipalities of Sennariolo and Sagama). Figure 1 (panel a,b,c, d) shows the maps depicting the municipal distribution of mortality for each period for men.

In a different way, for women, in the last two periods, the municipalities of La Maddalena, Osilo and Ploaghe remain as the areas with the highest RRs, between 1.10 and 1.25.

Figure 2 (panel a, b, c, d) depicts the spatial pattern of the posterior probability of RR being greater than one for women. Municipalities with a statistically significant excess risk are shown in orange and areas with a low risk are shown in blue.

A high-risk pattern is observed in the province of Sassari throughout all periods. In the third period, some other municipalities present high risk in the province of Olbia, while in the last period a marked geographical pattern arises, with numerous high-risk municipalities in the northeastern areas of the island (Sassari and Olbia) together with a low-risk area in the central region.



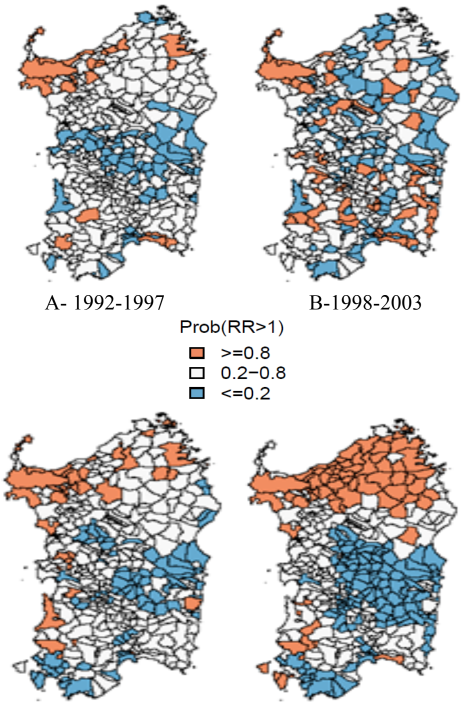
A- 1992-1997 B - 1998-2003

C- 2004-2009 D - 2010-2015

Figure 1: RR MEN 1992-2015

PRPs for men show (not reported here) in all periods an excess risk in northwestern and southwestern municipalities, in the provinces of Sassari, Medio-

Campidano and Carbonia-Iglesias. In the last two periods a high-risk pattern in the province of Nuoro emerges, the map showing a more pronounced cluster in the last period. Municipalities in the south, northeast and southeast show low risk of mortality and this pattern varies slightly throughout the study period. Table 1 shows for the first and the last period under analysis, the relationship between mortality and the deprivation index after including it in the estimated spatial model in the first and the last period of the study for men and for women.



A- 1992-1997 B-1998-2003

C- 2004-2009 D -.2010-2015

Figure 2: PRP Women

A slight increase in risk between the highest and lowest income categories can be observed in all periods, especially for men, but most credibility intervals are not statistically significant. For women, a threshold effect in the risk of death by level of deprivation emerges in the last two periods of study, while for men we cannot identify a similar temporal trend. For men, the first and second periods stand out, with almost all credibility intervals statistically significant.

Our results do not show a clear association between the deprivation index used and the risk of death, thus suggesting the necessity of studying other possible explanatory variables affecting mortality rates.

Table 1: *Deprivation index-adjusted model.*

		1 - MEN		*2-WOMEN*	
Period	***DI LEVEL***	***RR***	***IC95%***	***RR***	***IC95%***
1992-1997	1 (REF)	1	-	1	-
	2	1.11	0.98-2.23	0.99	0.93-1.04
	3	1.23	**1.02-3.22**	0.98	0.92-1.03
	4	1.76	**1.02-2.87**	1.01	0.95-1.07
	5	1.23	**1.05-2.96**	0.99	0.93-1.05
2010-2015	1 (REF)	1	-	1	-
	2	0.97	0.31-1.69	0.97	0.20-2.09
	3	1.01	0.20-2.04	0.95	0.20-2.15
	4	0.73	0.01-1.10	0.65	0.00-3.15
	5	1.18	0.34-2.64	1.19	0.22-3.97

3 Conclusions

The results of spatial analysis confirm those obtained by [3]. The largest increase in mortality rates in the period 1992-2015 was indeed recorded in the industrial areas of Northern Sardinia, and this worsening seems to be not ascribable neither to the age structure of the population nor to the level of deprivation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676060

References

1. Poulain, M., G. M. Pes, C. Grasland, C. Carru, L. Ferrucci, G. Baggio, C. Franceschi, and Deiana, L.: Identification of a Geographic Area Characterized by Extreme Longevity in the Sardinia Island: The AKEA Study. Experimental Gerontology 2004; 39 (9): 1423–1429
2. Pes, G. M., Tolu, F., Poulain, M., Errigo, A., Masala, S., Pietrobelli, A., Battistini, N. C., & Maioli, M.: Lifestyle and nutrition related to male longevity in Sardinia: An ecological study", Nutrition, Metabolism and Cardiovascular Diseases 2013; 23 (3), 212-219.
3. Caselli, G., Lipsi, R .M. :Survival differences among the oldest old in Sardinia: who, what, where and why?. Demographic Research 2006, 14 (13), 267-294.
4. Biggeri, A., Lagazio, C., Catelan, D., Pirastru, R., Casson, F., Terracini, B.: Ambiente e salute nelle aree a rischio della Sardegna. Epidemologia e Prevenzione, gen-feb 2006, supplemento 1. http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999
5. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 1991; 43: 1-20.
6. Caranci N, Biggeri A, Grisotto L, Pacelli B, Spadea T, Costa G. L'indice di deprivazione italiano a livello di sezione di censimento: definizione, descrizione e associazione con la mortalità. Epidemiol Prev 2010; 34 (4): 167-176
7. Brooks SP, Gelman A. Alternative methods for monitoring convergence of iterative simulations. J Comput Graph Stat. 1998; 7:434-455. .
8. Richardson S, Thomson A, Best N, Elliott P. Interpreting posterior relative risk estimates in disease-mapping studies. Environ Health Perspect 2004, 112: 1016-1025.

Recent Developments in Bayesian Inference

Posterior distributions with non explicit objective priors

Distribuzioni a posteriori basate su distribuzioni a priori oggettive non esplicite

Erlis Ruli, Nicola Sartori and Laura Ventura

Abstract We introduce two methods useful to derive a posterior distribution for a parameter of interest, when only the first derivative of a log-prior is available. This is typically the situation when dealing with multidimensional parameters and objective priors. An example is illustrated using a predictive matching prior.

Abstract *In questo contributo vengono introdotti due metodi utili per derivare una distribuzione a posteriori per un parametro di interesse, quando è disponibile solo la derivata prima del logaritmo della distribuzioni a priori. Tale situazione si presenta tipicamente in presenza di parametri multidimensionali e distribuzioni a priori oggettive. Il metodo viene illustrato in un modello logistico con una matching prior predittiva.*

Key words: Firth's adjustment, Logistic regression, Matching prior, MCMC, Rao score test statistic, Score function, Taylor expansion.

1 Introduction

Let $y = (y_1, \dots, y_n)$ be the available data, considered for simplicity as a random sample of size n, i.e. as a realization of a random variable $Y = (Y_1, \dots, Y_n)$ having independent and identically distributed components. Moreover, let $p(y;\theta) = \prod_{i=1}^{n} p(y_i;\theta)$ denote the probability density function of Y, with $\theta \in \Theta \subseteq \mathrm{I\!R}^k$, $k \geq 1$. We are interested in objective Bayesian inference on the unknown parameter θ, us-

Erlis Ruli
Department of Statistical Sciences, Univeristy of Padova, e-mail: ruli@stat.unipd.it

Nicola Sartori
Department of Statistical Sciences, Univeristy of Padova, e-mail: sartori@stat.unipd.it

Laura Ventura
Department of Statistical Sciences, Univeristy of Padova, e-mail: ventura@stat.unipd.it

ing the posterior distribution

$$\pi(\theta|y) \propto \pi(\theta)L(\theta), \tag{1}$$

where $\pi(\theta)$ is a prior for θ and $L(\theta) \propto p(y;\theta)$ is the likelihood function.

We consider the situation in which the prior distribution $\pi(\theta)$ is known only through its first derivative $\partial \log \pi(\theta)/\partial\theta$. This is typically the situation with defalut priors, such as matching priors (see, e.g., Datta and Mukerjee, 2004). In these cases, the posterior distribution (1) is not directly available, and it is only possibile to evaluate the first derivative of the log-posterior $t(\theta) = t(\theta;y) = \log\pi(\theta|y)$ given by

$$t_\theta(\theta) = t_\theta(\theta;y) = \frac{\partial}{\partial\theta}\log\pi(\theta|y) = \ell_\theta(\theta;y) + m(\theta), \tag{2}$$

where $\ell_\theta(\theta;y) = \partial \log L(\theta;y)/\partial\theta$ is the score function and $m(\theta) = \partial\log\pi(\theta)/\partial\theta$ is the derivative of the logarithm of the prior.

In this contribution we are interested in deriving the posterior density $\pi(\theta|y)$ such that $\partial\log\pi(\theta|y)/\partial\theta = t_\theta(\theta)$. In particular, we explore two methods for approximating $\pi(\theta|y)$ using MCMC and only $t_\theta(\theta)$ and its first derivative.

In the classical MCMC setting, the usual Metropolis-Hastings (MH) probability of acceptance of a candidate value $\theta^{(t+1)}$, given a chain at stage $\theta^{(t)}$, $\theta^{(t+1)} \sim q(\theta^{(t+1)}|\theta^{(t)})$, is

$$\min\left\{1, \frac{q(\theta^{(t)}|\theta^{(t+1)})}{q(\theta^{(t+1)}|\theta^{(t)})}\frac{\pi(\theta^{(t+1)}|y)}{\pi(\theta^{(t)}|y)}\right\}. \tag{3}$$

To evaluate (3), we must be able to evaluate

$$\frac{\pi(\theta^{(t+1)}|y)}{\pi(\theta^{(t)}|y)} = \exp\{t(\theta^{(t+1)}) - t(\theta^{(t)})\}, \tag{4}$$

in which normalizing constants, as is well known, are not needed. Here we propose two strategies for MCMC sampling even if $t(\theta) = \log\pi(\theta|y)$ is unknown, but its first and second derivatives are available in closed form. The first method (Section 2) considers an approximation based on a Rao score-type statistic based on (2). The second method (Section 3) is based on a local approximation through a Taylor expansion. We present an application to logistic regression with predictive matching priors (Section 4).

2 Method I: Approximation based on the Rao score statistic

A simple analytical way of using (2) is Bayesian statistic is to resort to a posterior distribution derived from a quadratic form of t_θ. This enables us to accomodate two important advantages of the Bayesian approach: the expressiveness of the posterior distribution and the convenient computational method of MCMC.

In particular, let $j(\theta) = -\ell_{\theta\theta}(\theta) = -\partial^2\ell(\theta)/\partial\theta^2$ be the observed Fisher information. Then the approximate posterior density takes the form

$$\pi(\theta|y) \mathrel{\dot{\propto}} \exp\left(-\frac{1}{2}t_\theta(\theta)^2 j(\theta)^{-1}\right) = \exp\left(-\frac{1}{2}\tilde{s}(\theta)\right), \tag{5}$$

where $\tilde{s}(\theta) = t_\theta(\theta)^2 j(\theta)^{-1}$ is a Rao score-type statistic based on (2) and the symbol "$\dot{\propto}$" means asymptotic proportionality to first order. In (4), (5) can be used for straightforward MCMC updating for the corresponding Bayesian posterior without any iterative optimization steps (Chernozbukov and Hong, 2003).

Here the idea is to recast (4) in terms of log-likelihood ratio type statistics and then replace the formers by Rao score tests. In particular,

$$\begin{aligned}\frac{\pi(\theta^{(t+1)}|y)}{\pi(\theta^{(t)}|y)} &= \exp\{(t(\theta^{(t+1)}) - t(\tilde{\theta})) - (t(\theta^{(t)} - t(\tilde{\theta}))\} \\ &\doteq \exp\{\tilde{s}(\theta^{(t)})/2 - \tilde{s}(\theta^{(t+1)})/2\}\end{aligned}$$

where $\tilde{\theta}$ is the posterior mode, such that $t_\theta(\tilde{\theta}) = 0$, and "$\doteq$" means asymptotic equality to first order.

3 Method II:I Local approximation through Taylor expansion

Assume, for notational simplicity, θ scalar. A Taylor expansion of $t(\theta)$ at θ_0 gives the approximation

$$t(\theta) \simeq t(\theta_0) + (\theta-\theta_0)t_\theta(\theta_0) + \frac{(\theta-\theta_0)^2}{2}t_{\theta\theta}(\theta_0), \tag{6}$$

where $t_{\theta\theta}(\theta) = (\partial t_\theta(\theta))/(\partial\theta)$. Using (6) we can approximate (4) using

$$\begin{aligned}t(\theta^{(t+1)}) - t(\theta^{(t)}) \approx\ & (\theta^{(t+1)} - \theta^{(t)})t_\theta(\theta_0) \\ &+ \frac{[(\theta^{(t+1)}-\theta_0)^2 - (\theta^{(t)}-\theta_0)^2]}{2}t_{\theta\theta}(\theta_0).\end{aligned} \tag{7}$$

Possible choices for θ_0 are $\theta^{(t)}$ or $\bar{\theta} = (\theta^{(t+1)} + \theta^{(t)})/2$. Note that

$$t_{\theta\theta}(\theta) = \ell_{\theta\theta}(\theta) + m_\theta(\theta) = \frac{\partial^2}{\partial\theta^2}\ell(\theta) + \frac{\partial}{\partial\theta}m(\theta) = \frac{\partial^2}{\partial\theta^2}\ell(\theta)\left\{1 + O(n^{-1})\right\}.$$

Hence, in (7) the quantity $t_{\theta\theta}(\theta)$ can be substituted by $\partial^2\ell(\theta)/\partial\theta^2$ with approximately the same level of accuracy.

4 Example: Predictive matching prior for logistic regression

In this section we discuss an example based on the logistic regression model and a predictive matching prior, i.e. a prior ensuring asymptotic equivalence of higher-order frequentist and Bayesian predictive densities (see, e.g., Datta and Mukerjee, 2004).

To give the expression of the proposed predictive matching prior, index notation and Einstein summation convention are convenient. Generic components of θ will be denoted by $\theta_r, \theta_s, \ldots$, with $r, s, \ldots = 1, \ldots, k$. First and second likelihood derivatives are ℓ_r and ℓ_{rs}. By equating the second-order asymptotic expansion of Corcuera and Giummolè (1999) of Bayesian predictive distributions and the frequentist modified estimative density of Komaki (1996), we obtain the proposed predictive matching prior, which is such that

$$t_{\theta_r}(\theta) = \frac{\partial_r \log \pi(\theta)}{\partial \theta_r} = -\frac{1}{2} i^{su}(\theta) \left(E_\theta\{\ell_{rsu}\} - E_\theta\{\ell_{rs}\ell_u\}\right), \tag{8}$$

where $E_\theta\{\cdot\}$ denotes expectation with respect to Y under θ and i^{rs} is the generic element of the inverse of $i(\theta) = E_\theta\{j(\theta)\}$. Note that the term on the right hand side of (8) corresponds to the Firth's adjustment (Firth, 1993) to the score function. In other words, if the prior density is choosen according to (8), then the right hand side of (2) is exactly the modified likelihood equation discussed by Firth (1993). In view of this, for general regular models, Firth's estimate coincides with the mode of the posterior distribution obtained using the default prior defined by (8). This prior thus validates the use of the method introduced by Firth (1993) for point estimation in the Bayesian framework.

Consider the binary logistic regression, where Y_i is Bernoulli with probability $\pi_i = P(Y_i = 1|x_i)$, where x_i is a known k-variate vector of regressors, $i = 1, \ldots, n$. The model can be expressed as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T \beta,$$

where β is an unknown vector of regression coefficients. The log-likelihood function for β is

$$\ell(\beta) = \sum_{j=1}^{k} \beta_j \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} \log\left(1 + \exp^{\sum_{j=1}^{k} \beta_j x_{ij}}\right).$$

In this example the posterior distribution is analytically available since the predictive matching prior coincides with Jeffreys' prior, up to the normalisation constant, and hence classical MCMC can be performed. Therefore, we use it as a benchmark in order to assess the accuracy of the proposed approximation methods. We stress, however, that in many other practical cases, e.g. with non canonical links, (8) does not lead to Jeffreys' prior and and with Firth's adjustment, the posterior is not available, and thus classical MCMC is not possible. The Fisher information is $i(\beta) = j(\beta) = X^T WX$, where X is the design matrix, and W is a diagonal matrix with elements $(w_1, \ldots, w_n)$, with $w_i = \pi_i(1-\pi_i)$ $(1 \leq i \leq n)$. The modified likelihood equation (2) (Firth, 1993) and its first derivative is

$$t^*_{\beta_r}(\beta) = \sum_{i=1}^{n}(y_i - \pi_i)x_{ir} + t_{\theta_r}(\theta), \quad 1 \leq r \leq k.$$

For this model based on the canonical link, the posterior for β is available and is given by

$$\pi_J(\beta|y,X) \propto L(\beta;y)|i(\beta)|^{1/2}. \tag{9}$$

We compare the posterior (9) with its approximate versions obtained with Methods I and II, using the *endometrial* dataset. The latter has been first analysed by Heinze and Schemper (2002) and reports histology grade (HG, the binary response variable) and three risk factors (NV a binary indicator for the presence of neovasculation, PI the pulsality index of arteria uterina and EH the endometrium height) for 79 cases of endometrial cancer.

Consider the model with all the covariates included, i.e.,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \texttt{NV}_i + \beta_2 \texttt{PI}_i + \beta_3 \texttt{EH}_i, \quad 1 \leq i \leq n. \tag{10}$$

Figure 1 compares the marginal posteriors of β_j $(0 \leq j \leq 3)$ obtained from (9) by classical MCMC with the corresponding approximations obtained by Method I (Taylor) and Method II (Rao). We generate 10^6 samples from (9) and from the two approximate posteriors, using the MH algorithm with a multivariate normal proposal distribution. The latter has scale matrix given by the negative of the inverse of the second derivative of $t(\theta)$. For (9) and the approximate posterior obtained by Method II the proposal was tuned to give approximately 0.4 acceptance rate; see Section 5 for the computational details. From Figure 1 we can conclude that the approximation obtained with Method I is very similar to the target (9), whereas the approximation obtained with Method II is less accurate.

5 Concluding remarks

In general, the higher is the acceptance rate the lower is the approximation error of Method I. We recognise that high acceptance rates in MCMC are generally not recommended because the posterior exploration of the MCMC algorithm for a finite time period may be too local. The issue of choosing an optimal acceptance rate is under investigation. However, a practical guidance to circumvent this issue would be as follows. Set a sequence of shrunken proposal matrices obtained by shrinking the main diagonal elements of the starting proposal scaling matrix, and with each of them generate an MCMC sample. Then, for this sequence of MCMC samples, which has increasing acceptance rates, monitor the shape of the resulting marginal

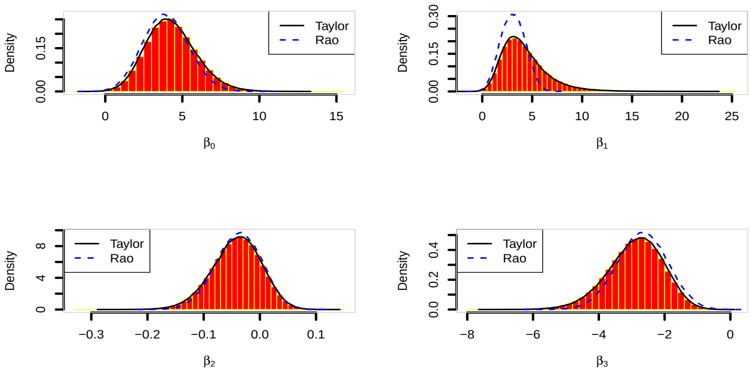


Fig. 1 Marginal posterior distributions for the logistic regression model with the *endometrial* data. The marginals of (10) are illustrated by histograms.

posteriors. If the shape of the latters is reasonably stable across two or three consequent posterior samples, then the MCMC sample with the lowest acceptance rate may by used for posterior inference. This is the strategy adopted in Section 4 which lead to an acceptance rate of 0.70.

Acknowledgement. This research work was partially supported by University of Padova (Progetti di Ricerca di Ateneo 2015, CPDA153257) and by PRIN 2015 (grant 2015EASZFS_003).

References

1. Chernozbukov, V., Hong, H.: An MCMC approach to classical estimation, J. Econometrics **115**, 1234–1241 (2003)
2. Corcuera, J.M., Giummolè, F.: A generalized Bayes rule for prediction. Scand. J. Statist. **26**, 265–279 (1999)
3. Datta, G.S., Mukerjee, R.: Probability Matching Priors: Higher-Order Asymptotics. Lecture Notes in Statistics, Springer (2004)
4. Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika **80**, 27–38 (1993)
5. Heinze, G., Schemper, M: A Solution to the problem of separation in logistic regression. Statist. Med. **21**, 2409–2419 (2002)
6. Komaki, F.: On asymptotic properties of predictive distributions. Biometrika **83**, 299–313 (1996)

A predictive measure of the additional loss of a non-optimal action under multiple priors

Una misura predittiva della perdita dovuta all'uso di un'azione non ottima in presenza di diverse a priori

Fulvio De Santis and Stefania Gubbiotti

Abstract In Bayesian decision theory, the performance of an action is measured by its posterior expected loss. In some cases it may be convenient/necessary to use a non-optimal decision instead of the optimal one. In these cases it is important to quantify the additional loss we incur and evaluate whether to use the non-optimal decision or not. In this article we study the predictive probability distribution of a relative measure of the additional loss and its use to define sample size determination criteria in one-sided testing.
Abstract L'analisi delle decisioni bayesiane prevede che la qualità di un'azione si misuri in termini della sua perdita attesa a posteriori. In alcuni casi può essere conveniente/necessario adottare una decisione non ottima al posto di quella ottima. Per valutare l'opportunità di questa scelta, è importante quantificare la perdita aggiuntiva che essa comporta. Oggetto di questo lavoro è lo studio della distribuzione predittiva di una misura relativa di tale perdita addizionale e il suo impiego per la scelta della numerosità campionaria nei problemi di test di ipotesi unilaterali.

Key words: Bayesian inference, Experimental design, One-sided testing, Predictive analysis, Sample size determination, Statistical decision theory.

1 Introduction

In a decision problem involving the unknown parameter of a statistical model, consider two decision makers who have different prior information and/or opinions on

Fulvio De Santis
Dipartimento di Scienze Statistiche, Sapienza Università di Roma,
e-mail: fulvio.desantis@uniroma1.it

Stefania Gubbiotti
Dipartimento di Scienze Statistiche, Sapienza Università di Roma,
e-mail: stefania.gubbiotti@uniroma1.it

the parameter. Let π_e and π_o denote their priors and let a_e and a_o be the actions that minimize the two posterior expected losses. Furthermore, let us suppose that the first decision makers is forced to take the action a_o, although it is not optimal from her/his point of view: under π_e, the posterior expected loss of a_o is in fact larger than the posterior expected loss of a_e. Finally, assume that the sample size of the experiment is selected by a third actor using the predictive distribution of the data based on the prior π_d, that, in general, is different from both π_o and π_e. The goal of the experiment planner is to determine the minimal sample size such that a relative predictive measure of the additional loss due to the use of a_o rather than a_e is sufficiently small.

Statistical decision problems under several actors have been previously considered, for instance, in [5], [6] and [4]. In this paper we extend to the testing problem the results of [3] (related to point-estimation) and we focus in particular on the one-sided testing set-up.

The outline of the article is as follows. In Section 2 we formalize the proposed methodology for a generic statistical decision problem: we introduce a relative measure of additional loss due to a non optimal action and the related predictive criterion for the selection of the sample size. In Section 3 the methodology is developed for a one-sided testing problem for a real-valued parameter. Results are then specialized to one-sided testing of a normal mean (Section 4) and some numerical examples are provided in Section 4.1. Finally, Section 5 contains some concluding remarks.

2 Methodology

Let $X_1, X_2, \ldots, X_n$ be a random sample from $f_n(\cdot|\theta)$, where θ is an unknown parameter, $\theta \in \Theta$. Let $a \in \mathscr{A}$ denote a generic action for a decision problem regarding θ and $L(a,\theta)$ the loss of a when the true parameter value is θ. We assume that two competing priors, π_o and π_e, are available for θ. Given an observed sample $\boldsymbol{x_n} = (x_1, x_2, \ldots, x_n)$, let $\pi_j(\theta|\boldsymbol{x_n})$ be the posterior distribution of θ from prior π_j, and

$$\rho_j(\boldsymbol{x_n}, a) = \mathbb{E}_{\pi_j}\left[L(a,\theta)|\boldsymbol{x_n}\right] = \int_{\Theta} L(a,\theta)\pi_j(\theta|\boldsymbol{x_n})d\theta$$

be the posterior expected loss of an action a, for $j = o, e$. Let a_j denote the optimal action with respect to $\pi_j(\theta|\boldsymbol{x_n})$. The performance of the action a_o when the expected loss is evaluated with respect to $\pi_e(\theta|\boldsymbol{x_n})$ is then $\rho_e(\boldsymbol{x_n}, a_o) = \mathbb{E}_{\pi_e}\left[L(a_o,\theta)|\boldsymbol{x_n}\right]$. If a_o is used instead of a_e, the *relative additional expected loss* is

$$\bar{A}_{o,e}(\boldsymbol{x_n}) = \frac{\rho_e(\boldsymbol{x_n}, a_o) - \rho_e(\boldsymbol{x_n}, a_e)}{\rho_e(\boldsymbol{x_n}, a_o)}.$$

When $\bar{A}_{o,e}$ is small the non-optimal action a_o performs well even under the prior assumptions represented by π_e. Before observing the data, $\bar{A}_{o,e}(\boldsymbol{X_n})$ is a sequence of r.v. that converges in probability to zero, as n increases. In order to define a sample size criterion we focus on $e_n = \mathbb{E}_{m_d}\left[\bar{A}_{o,e}(\boldsymbol{X_n})\right]$, where $\mathbb{E}_{m_d}[\cdot]$ denotes the expected value with respect to the sample data distribution, $m_d(\boldsymbol{x_n}) = \int_\Theta f(\boldsymbol{x_n}|\theta)\pi_d(\theta)d\theta$, where π_d is the design prior. Hence, for a desired threshold γ, $n^\star = \min\{n \in \mathbb{N} : e_n \leq \gamma\}$ is the optimal sample size that depends on three priors (π_d, π_e, π_o).

3 One-sided Testing

Consider the set up of one-sided testing, i.e. $H_1 : \theta \leq \theta_t$ vs. $H_2 : \theta > \theta_t$, with $\theta_t \in \mathbb{R}$. Let $\mathscr{A} = \{a^{(1)}, a^{(2)}\}$ be the two terminal decisions, where $a^{(i)}$ denotes the choice of H_i, $i = 1, 2$, and

$$L(a^{(1)}, \theta) = b_2 \times 1_{\{\theta:\theta>\theta_t\}}(\theta) \quad \text{and} \quad L(a^{(2)}, \theta) = b_1 \times 1_{\{\theta:\theta\leq\theta_t\}}(\theta)$$

their loss functions ($b_i > 0$, $i = 1, 2$), with $1_A(\cdot)$ the indicator function of the set A. Then the posterior expected losses of $a^{(1)}$ and $a^{(2)}$ are

$$\rho_j(\boldsymbol{x_n}, a^{(1)}) = b_2\,(1 - \mathrm{F}_j(\theta_t|\boldsymbol{x_n})) \qquad \text{and} \qquad \rho_j(\boldsymbol{x_n}, a^{(2)}) = b_1 \mathrm{F}_j(\theta_t|\boldsymbol{x_n}),$$

where $F_j(\cdot|\boldsymbol{x_n})$ is the c.d.f. associated to $\pi_j(\theta|\boldsymbol{x_n})$, $j = o, e$. In this case it is easy to check that the optimal decision function $a_j(\boldsymbol{x_n})$ is

$$a_j(\boldsymbol{x_n}) = \arg\min_{a\in\mathscr{A}} \rho_j(\boldsymbol{x_n}, a) = \begin{cases} a^{(1)} \text{ if } \boldsymbol{x_n} \in \mathscr{Z}_j^{(1)} \\ a^{(2)} \text{ if } \boldsymbol{x_n} \in \mathscr{Z}_j^{(2)} \end{cases} \qquad j = o, e.$$

where

$$\mathscr{Z}_j^{(1)} = \{\boldsymbol{x_n} :\ \rho_j(\boldsymbol{x_n}, a^{(1)}) < \rho_j(\boldsymbol{x_n}, a^{(2)})\} = \{\boldsymbol{x_n} :\ b_2\,(1 - \mathrm{F}_j(\theta_t|\boldsymbol{x_n})) < b_1 \mathrm{F}_j(\theta_t|\boldsymbol{x_n})\}$$

and $\mathscr{Z}_j^{(2)}$ is its complement. The posterior expected loss of the decision function $a_j(\boldsymbol{x_n})$ w.r.t. π_e is

$$\rho_e(\boldsymbol{x_n}, a_j) = \begin{cases} b_2\,(1 - \mathrm{F}_e(\theta_t|\boldsymbol{x_n})) \text{ if } \boldsymbol{x_n} \in \mathscr{Z}_j^{(1)} \\ b_1\,\mathrm{F}_e(\theta_t|\boldsymbol{x_n}) \qquad\quad \text{if } \boldsymbol{x_n} \in \mathscr{Z}_j^{(2)} \end{cases} \qquad j = o, e.$$

Therefore, noting that $\rho_e(\boldsymbol{x_n}, a_e) = \min\{b_1 \mathrm{F}_e(\theta_t|\boldsymbol{x_n}), b_2\,(1 - \mathrm{F}_e(\theta_t|\boldsymbol{x_n}))\}$, we obtain

$$\bar{A}_{o,e}(\boldsymbol{x_n}) = \xi_e(\boldsymbol{x_n}) 1_{\mathscr{Z}_{o,e}}(\boldsymbol{x_n}) \tag{1}$$

where

$$\xi_e(\boldsymbol{x_n}) = 1 - \min\left\{\frac{b_1}{b_2}\frac{\mathrm{F}_e(\theta_t|\boldsymbol{x_n})}{1 - \mathrm{F}_e(\theta_t|\boldsymbol{x_n})}, \frac{b_2}{b_1}\frac{1 - \mathrm{F}_e(\theta_t|\boldsymbol{x_n})}{\mathrm{F}_e(\theta_t|\boldsymbol{x_n})}\right\}. \tag{2}$$

and

$$\mathscr{Z}_{o,e} = \{\boldsymbol{x_n} \in \mathscr{X}^n : a_o(\boldsymbol{x_n}) \neq a_e(\boldsymbol{x_n})\} = \left(\mathscr{Z}_o^{(1)} \cap \mathscr{Z}_e^{(2)}\right) \cup \left(\mathscr{Z}_o^{(2)} \cap \mathscr{Z}_e^{(1)}\right)$$

is the set of $\boldsymbol{x_n}$ leading to conflicting terminal decisions under π_e and π_o respectively. Now, note that $\mathscr{Z}_j^{(1)}$ can be rewritten in terms of the ε-quantile of the posterior distribution of θ, $q_\varepsilon^j(\boldsymbol{x_n})$ with $\varepsilon = \frac{b_2}{b_1+b_2}$, namely

$$\mathscr{Z}_j^{(1)} = \left\{\boldsymbol{x_n} \in \mathscr{Z} : \frac{1 - \mathrm{F}_j(\theta_t|\boldsymbol{x_n})}{\mathrm{F}_j(\theta_t|\boldsymbol{x_n})} < \frac{b_1}{b_2}\right\} = \left\{\boldsymbol{x_n} \in \mathscr{Z} : \theta_t > q_\varepsilon^j(\boldsymbol{x_n})\right\}. \qquad (3)$$

Therefore

$$\mathscr{Z}_o^{(1)} \cap \mathscr{Z}_e^{(1)} = \left\{\boldsymbol{x_n} \in \mathscr{Z} : q_\varepsilon^M(\boldsymbol{x_n}) < \theta_t\right\} \text{ and } \mathscr{Z}_o^{(2)} \cap \mathscr{Z}_e^{(2)} = \{\boldsymbol{x_n} \in \mathscr{Z} : q_\varepsilon^m(\boldsymbol{x_n}) > \theta_t\},$$

where $q_\varepsilon^m(\boldsymbol{x_n}) = \min\{q_\varepsilon^e(\boldsymbol{x_n}), q_\varepsilon^o(\boldsymbol{x_n})\}$ and $q_\varepsilon^M(\boldsymbol{x_n}) = \min\{q_\varepsilon^e(\boldsymbol{x_n}), q_\varepsilon^o(\boldsymbol{x_n})\}$. Hence we have

$$\mathscr{Z}_{o,e} = \left\{\boldsymbol{x_n} \in \mathscr{Z} : q_\varepsilon^m(\boldsymbol{x_n}) < \theta_t < q_\varepsilon^M(\boldsymbol{x_n})\right\}. \qquad (4)$$

Finally, from (1)

$$e_n = \int_{\mathscr{Z}} \bar{A}_{o,e}(\boldsymbol{x_n}) m_d(\boldsymbol{x_n}) d\boldsymbol{x_n} = \int_{\mathscr{Z}_{o,e}} \xi_e(\boldsymbol{x_n}) m_d(\boldsymbol{x_n}) d\boldsymbol{x_n}$$

that, in general, must be computed via Monte Carlo approximation. From the above expression we can note that e_n is a monotone function of the Lebesgue measure of $\mathscr{Z}_{o,e}$. An alternative sample size criterion could be based on the predictive probability p_n of the samples yielding conflict. Recalling that, $\forall \boldsymbol{x_n} \in \mathscr{Z}$, $\xi_e(\boldsymbol{x_n}) \leq 1$, it is easy to check that $p_n = \mathbb{P}_{m_d}[\mathscr{Z}_{o,e}] = \mathbb{E}_{m_d}[1_{\mathscr{Z}_{o,e}}(\boldsymbol{X_n})]$ is always smaller than or equal to $e_n = \mathbb{E}_{m_d}[\xi_e(\boldsymbol{X_n}) 1_{\mathscr{Z}_{o,e}}(\boldsymbol{X_n})]$. Therefore, for a given γ, e_n always yields a smaller sample size. The idea is that in e_n the contribution of each sample corresponding to a conflicting decision depends on the strength of the discrepancy in evidence it gives to the two hypotheses, whereas in p_n, it is invariably equal to one.

4 Results for the Normal mean

Let us now further assume that $X_i|\theta \sim \mathrm{N}(\theta, \sigma^2)$, $i = 1, 2, \ldots, n$ and that $\pi_j(\cdot)$ are conjugate priors, i.e. $\theta|\sigma^2 \sim \mathrm{N}(\mu_j, \sigma^2/n_j)$, $j = o, e$. When σ^2 is assumed to be known, the posterior distribution of θ is Normal with mean $\mu_j(\boldsymbol{x_n}) = \frac{n_j\mu_j + n\bar{x}_n}{n_j+n}$ and standard deviation $\sigma_j(\boldsymbol{x_n}) = \frac{\sigma}{\sqrt{n_j+n}}$. In this case $\bar{A}_{o,e}$ can be expressed in terms of Φ, z_ε and $W_j(\boldsymbol{x_n})$, where $\Phi(\cdot)$ is the standard normal c.d.f., z_ε its ε-quantile and

$$W_j(\boldsymbol{x_n}) = \frac{\mu_j(\boldsymbol{x_n}) - \theta_t}{\sigma_j(\boldsymbol{x_n})}, \qquad j = o, e.$$

First, from Equation (2) we have

$$\xi_e(\boldsymbol{x_n}) = 1 - \min\left\{\frac{b_1}{b_2}\frac{1-\Phi(W_e(\boldsymbol{x_n}))}{\Phi(W_e(\boldsymbol{x_n}))}, \frac{b_2}{b_1}\frac{\Phi(W_e(\boldsymbol{x_n}))}{1-\Phi(W_e(\boldsymbol{x_n}))}\right\},$$

Then, from (3), it follows that

$$\mathscr{X}_j^{(1)} = \left\{\boldsymbol{x_n} \in \mathscr{X} : W_j(\boldsymbol{x_n}) = \frac{\mu_j(\boldsymbol{x_n}) - \theta_t}{\sigma_j(\boldsymbol{x_n})} < z_{1-\varepsilon}\right\}$$

and finally

$$\mathscr{X}_{o,e} = \{\boldsymbol{x_n} \in \mathscr{X} : W_m(\boldsymbol{x_n}) < z_{1-\varepsilon} < W_M(\boldsymbol{x_n})\}, \qquad (5)$$

where $W_m(\boldsymbol{x_n}) = \min\{W_o(\boldsymbol{x_n}), W_e(\boldsymbol{x_n})\}$ and $W_M(\boldsymbol{x_n}) = \max\{W_o(\boldsymbol{x_n}), W_e(\boldsymbol{x_n})\}$.

4.1 Numerical example

Let us consider $\theta_t = 1$ and let the design prior be a Normal density of parameters $\mu_d = 1.5$, $n_d = 10$. Thus, π_d assigns to H_1 a prior probability as small as 0.056. Figure 1 shows the behavior of e_n as n increases, under two alternative choices of μ_e for different values of the prior sample sizes n_e and n_o. In the former case, we assume that there is a certain contrast between the two priors: π_e, centred on the threshold θ_t (e.g. $\mu_e = 1$), expresses a neutral attitude towards the two hypotheses, whereas π_o favors the null hypothesis (e.g. $\mu_o = 0$). In the left panel of Figure 1 for small values of n (due to the predominant role of the prior weights n_e and n_o) e_n increases up to a maximum value and then it definitively decreases, tending to zero more and more rapidly for smaller values of the prior sample sizes n_e and n_o. In the latter set-up, the conflict between π_e and π_o is emphasized, π_e supports the alternative hypothesis H_2 and μ_e is even larger than μ_d (i.e. $\mu_o = 0$ and $\mu_e = 2$). As shown in the right panel, e_n monotonically decreases as a function of n from 1 to 0. As before, when the two conflicting priors are more and more concentrated, the expected value of $\bar{A}_{o,e}$ is uniformly larger and, consequently, a larger number of observations is required for the conflict to be resolved.

Finally in Figure 2 we illustrate by examples the relationship that holds in general between e_n and p_n, that is $p_n < e_n$, as commented in the final remark of Section 3.

5 Future research directions

The article leaves open the possibility of further developments, such as the application to non-normal models and to more challenging (not necessarily one-dimensional) testing set-ups. Moreover instead of considering only one prior π_e, we could extend our approach by considering an entire class of priors Γ. In this case, we

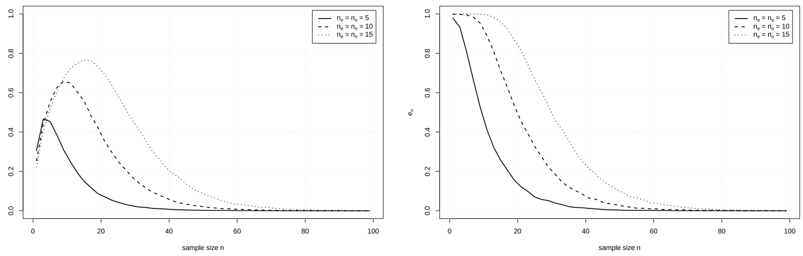


Fig. 1 e_n as a function of the sample size n, with $\mu_e = 1$ (*left panel*) and $\mu_e = 2$ (*right panel*) for different values of n_e and n_o, given $\theta_t = 1$, $\sigma = 1$, $\mu_d = 1.5$, $n_d = 10$, $\mu_o = 0$

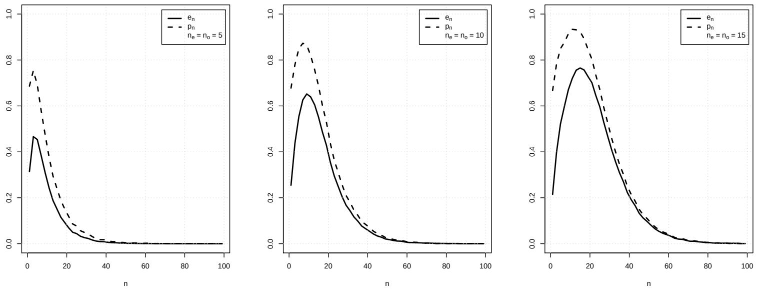


Fig. 2 Behavior of e_n and p_n for increasing values of n, with $\mu_e = 1$ (first row) and $\mu_e = 2$ (second row) for different values of n_o and n_e, given $\theta_t = 1$, $\sigma = 1$, $\mu_d = 1.5$, $n_d = 10$, $\mu_o = 0$.

would be interested in looking at the largest relative additional loss of a_o as π_e varies in Γ and the sample size is chosen by replacing e_n with $e_n^{\Gamma} = \mathbb{E}_{m_d}[\sup_{\pi_e \in \Gamma} \bar{A}_{o,e}]$.

References

1. Brutti P, De Santis F, Gubbiotti S. Predictive measures of the conflict between frequentist and Bayesian estimators. *Journal of Statistical Planning and Inference* 2014; **148**:111-122.
2. Brutti P, De Santis F, Gubbiotti S. Bayesian frequentist sample size determination: A game of two priors. *Metron* 2014; **72**(2):133-151.
3. De Santis F., Gubbiotti S. A decision-theoretic approach to sample size determination under several priors. *Applied Stochastic Models in Business and Industry* 2017; **33**(3):282-295.
4. Etzioni R, Kadane JB. Optimal experimental design for anothers analysis. *J. Am. Stat. Assoc.* 1993; **88**(424):1404-1411.
5. Kadane JB, Seidenfeld T. Randomization in a Bayesian perspective. *Journal of Statistical Planning and Inference* 1989; **25**:329-345.
6. Lindley DV, Singpurwalla N. On the Evidence Needed to Reach Agreed Action Between Adversaries, With Application to Acceptance Sampling. *Journal of the American Statistical Association* 1991; **86**(416):933-937.

Bayesian estimation of number and position of knots in regression splines

Stima Bayesiana del numero e della posizione dei nodi in spline di regressione

Gioia Di Credico, Francesco Pauli and Nicola Torelli

Abstract Regression splines, based on piecewise polynomials, are useful tools to model departures from linearity in the regression context. The number and location of the knots can be of interest in many contexts since they can detect possible change points in the relationship between the variables. This work is focused on the estimate of both number and location of knots in the simple case where linear truncated splines are chosen to represent the relationship, in this case, the position of the knot detects a change in the slope. In a Bayesian context, we propose a two-step procedure, to first determine the true number of knots and then to fit the final model estimating simultaneously location of knots and regression and spline coefficients.
Sommario *Le spline applicate a modelli di regressione con polinomi a tratti possono essere utili al fine di descrivere relazioni non lineari. In alcuni contesti e immaginando di limitare l'attenzione a splines con componenti lineari, può essere interessante conoscere il numero e la posizione dei nodi che sono quindi i cambi di pendenza della retta di regressione. Tuttavia, stimare numero e posizione dei nodi aggiunge una componente di stima non lineare al problema di ottimizzazione. Proponiamo una procedura in due passi che prima determina il numero ottimale dei nodi e poi ne stima la posizione, congiuntamente agli altri coefficienti del modello. La metodologia viene applicata qui ai modelli lineari adottando un approccio bayesiano.*

Key words: Regression splines, Stan, spike-and-slab priors, knot location

Gioia Di Credico
Department of Statistics, Padua University, Padua, Italy, e-mail: gioia.dicredico@studenti.unipd.it

Francesco Pauli and Nicola Torelli
Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Trieste, Italy e-mail: francesco.pauli@deams.units.it and e-mail: nicola.torelli@deams.units.it

1 Introduction

When modelling the relationship between a response and some (continuous) covariates the linearity assumption turns out to be too restrictive in many contexts. Naive solutions to overcome this limitation such as categorization of the predictor or its polynomial representation have well-known drawbacks.

A viable alternative is represented by spline functions. They are defined as piecewise polynomials with a fixed degree whose joint points are called knots. Splines are highly flexible, in fact, varying the number and position of knots may lead to extremely different shapes and a major risk is to overfit the data. A classical approach consists in using an optimizing criterion with a suitable penalization to control the roughness of the function. Other techniques proposed in the literature include the use of variable selection to choose basis function [6], or employing samplers that allow for varying dimension of the parameter [2, 3].

Assuming that the number and position of knots may have an important and substantial interpretation, here we consider their estimation following one of the most recent approaches to variable selection in a Bayesian context. Estimating the positions of the knots is not an easy task and, for a fixed degree, regression coefficients and locations of knots have to be estimated simultaneously, turning the standard estimation procedure into a nonlinear optimization problem. In the sequel, we propose a method to estimate the number and position of knots with a two-step procedure.

2 Methods

Consider the model

$$y_i = z_i^\intercal \alpha + f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

where z is the covariates vector that enters linearly in the model, α is the vector of regression coefficients, x is a continuous variable evaluated through a smooth function $f : \mathbb{R} \to \mathbb{R}$, described with a spline with few knots and ε is an i.i.d. Gaussian random error component.

We restrict our analysis to those situations in which a low number of knots can be adequate and their positions are directly interpretable and of specific interest for the analysis. This is the case, for example, when truncated power basis (TPB) of order one is used since in this case positions of knots represent changing points for the slope. One of the main drawbacks of truncated power basis representation is that the basis is not orthogonal, which can lead to numerical instability and slow convergence of the optimization algorithm. Keeping a low number of knots alleviates the issue [5].

Let then

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \gamma_k (x - \xi_k)_+, \tag{1}$$

where ξ_k is the position of the k-th knot and K is the total number of knots, and

$$(x-\xi)_+ = \begin{cases} x-\xi, & \text{if } x \geq \xi \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

is the truncated linear function. Given the number of knots, parameters estimation reduces to maximum likelihood estimate. A usual approach is to choose the knot locations using standard criteria (such as quantiles of the predictor distribution, uniformly distributed knots on the range of the independent variables and user-defined knots following a priori information [5]), estimate models with a different number and location of knots and compare them through standard criteria, such as AIC, BIC or GCV. This procedure often results in a not clear discrimination among all competing models.

In order to enhance the fit of the model, a possible extension is to consider locations of knots as parameters to be estimated along with other regression coefficients. In such a case, within a maximum likelihood approach, exploration of the objective function surface could locate local maxima leading to apparent solutions strongly dependent on starting values. A Bayesian specification of the model and exploring of the posterior distribution, possibly by using MCMC simulations, could prove in this case much more effective.

Our aim is to estimate both number and location of knots, thus a preliminary idea is to estimate several models with free knot locations and with increasing but fixed number of knots. Knots are constrained to be ordered and their prior distributions are Uniform on the range of the variable. Vague priors on the regression and spline coefficients are chosen (zero-centered normal with large variance). We will refer to this model as the no variable selection (NVS) model.

Models with an increasing number of knots were compared on the basis of diagnostic tools such as traceplots and $\hat{R}$ to check convergence of parameters and information criteria were used to choose the best model among the estimated ones. The main drawback of this procedure is a large number of models that one needs to consider and the implied computational effort in high dimensional problems. However, results obtained on simulated data on the basis of these diagnostic tools show that it can lead to reasonable estimates and that convergence of chains for knot related parameters is univocal only if the number of specified knots is lower or equal to the true one.

This prompted us to consider a two-step procedure:

- select the optimal number of knots considering a large, possibly, overparameterized model,
- fit the final model by simultaneously estimating locations of knots and regression and spline coefficients.

In the first step, we estimate a model having more knots than reasonably warranted. This leads to an overparameterized model where the posterior of some knot locations are expected to concentrate at the limits of the predictor range. To assess convergence of the spline parameters, our advice is to run several chains and look at the results of each chain separately. Indeed, overparameterizing the model may lead to chains which converge at different points. Since each knot location is uniquely linked to a spline coefficient, we evaluate the presence of a knot based on the analysis of the associated coefficient posterior distribution.

The concept underlying the proposed methodology is to perform variable selection on the basis functions, for this purpose we employ one of the most common approaches in Bayesian literature: that based on the definition of spike-and-slab priors. Several versions have been proposed in the literature [4] but, generally speaking, prior distributions for the regression coefficients are defined with a spike component, usually highly concentrated around zero, and a diffused slab part. This is the case of the stochastic search variable selection approach (SSVS), that defines a mixture distribution for each parameter that has to be selected [4]. This type of methodology gives us the opportunity to evaluate the presence of a variable through the marginal posterior distribution of the mixing proportion. Starting from the NVS model specification, we set a prior distribution on each spline parameter γ_k such that

$$\pi(\gamma_k|\lambda_k) = \lambda_k N(0,\sigma_{sl}) + (1-\lambda_k)N(0,\sigma_{sp}),$$

where the mixing proportion $\lambda_k \sim Beta(a,b)$, with $a = b$. Standard deviations of the two mixture components, σ_{sl} and σ_{sp}, are chosen to be respectively large and small. Appropriate values have to be evaluated taking into account the unit of measurement of dependent and independent variables.

Our method adapts the SSVS approach by assuming λ_k to be dependent on the knot location ξ_k. The prior distributions of the ordered knots remain defined as Uniform on the support of the variable X and independent from both the mixing proportion λ and the coefficient γ. Each coefficient γ_k, conditioned on the mixing parameter λ_k follows the same mixture distribution of two components specified in the SSVS approach described above, while each element of the mixing proportion vector λ is now defined as:

$$\lambda_k|\xi_k \sim Beta(a,b_k),$$

where a is a positive but very small value and $b_k : [\min(X);\max(X)] \to [a;1+a]$ is a U-shaped even function of the knot location which returns values close to $1+a$ when the knot is near the boundaries of the variable, while it is almost uniform and close to a elsewhere. In practice, the prior for the mixing parameter swings between a beta U-shaped distribution when the knot location is on plausible values and a beta distribution highly concentrated on zero when the knot is close to the boundaries. All the other prior distributions remain defined as in the previous model specification.

In the next section, we compare results from (i) the proposed method, named later on SSVSξ, with (ii) the ones obtained from the SSVS approach and (iii) the same model without a variable selection procedure, NVS.

3 Preliminary results

We simulate data from the linear regression model

$$y_i = 6 + 2x_i - 5(x_i - 2.7)_+ + 8(x_i - 4.3)_+ + \varepsilon_i, \quad i = 1, \ldots, 500,$$

where $\varepsilon_i \sim$ i.i.d. $N(0,3)$ and the predictor X is defined on the interval $[0;10]$. Two knots are placed respectively in 2.67 and 4.33. We set the parameter a of the mixing proportions λ for the SSVS and the SSVSξ models equal to 0.5. Moreover, we chose σ_{sl} equal to 100 and σ_{sp} equal to 0.1. Standard deviations of the prior distributions on spline coefficients and intercept were chosen equal to 100.

We run 10 chains with 2000 iterations each. Posterior inference is based on the last 1000 draws of each chain. To support the complete exploration of the posterior distribution, initial values for the location of the knots are chosen widely spread on the range of the predictor variable X. Spline coefficients and intercept are initialized at zero. The three models are fitted with a different number of knots (respectively with 2, 5 and 10 knots). The interest lies in the parameter estimates, both spline coefficients and knot locations, and in the analysis of the chains behavior.

The number of knots can be chosen in the SSVS and SSVSξ models looking at the plots in Fig 1. The x-axis represents the specified number of knots in the over-parameterized models, while the y-axis represents the posterior mean of the mixing proportion. Vectors of posterior means are sorted in descending order and each line corresponds to one chain. In both models performing variable selection, the selected number of relevant knots is always equal to 2, even if the SSVSξ approach makes a slightly clearer distinction with respect to the classic SSVS method.

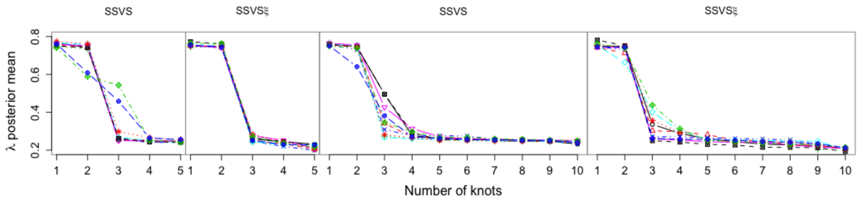


Fig. 1 Posterior means of the mixing parameters λ in the overparameterized SSVS and SSVSξ models with 5 and 10 knots.

The second step of the procedure is to estimate the models with the selected number of knots.

The three models are compared by means of diagnostic tools, such as traceplots, $\hat{R}$, effective sample size (n$_{eff}$) and analysis of marginal posterior distributions. Due to space constraints, in table 1 we report estimates only for the SSVSξ model, parameter estimates are close to the true parameter values. The greatest discrepancies among the model results are on the order of one decimal point. For the three models, $\hat{R}$ statistics equal to 1 suggest that the chains show good mixing, but differences in the n$_{eff}$ estimates highlight a lower estimate stability of SSVS model compared to the other two fitted models.

Table 1 Posterior distributions of the SSVSξ model parameters. The model is estimated with the true number of knots. $\hat{R}$ and n$_{eff}$ statistics.

Parameter	true	mean	sd	2.5%	50%	97.5%	Rhat	n$_{eff}^{SSVS\xi}$	n$_{eff}^{SSVS}$	n$_{eff}^{NVS}$
β_0	6	6.6	0.6	5.4	6.6	7.6	1.0	4644	762	3313
β_1	2	2.1	0.4	1.3	2.1	3.1	1.0	3039	667	2171
γ_1	-5	-4.5	0.7	-5.8	-4.5	-3.3	1.0	2290	287	3468
γ_2	8	7.4	0.6	6.3	7.3	8.5	1.0	2704	410	2690
ξ_1	2.7	2.4	0.2	1.9	2.4	2.8	1.0	3183	3105	2066
ξ_2	4.3	4.4	0.1	4.2	4.4	4.5	1.0	4634	597	5196
λ_1		0.7	0.3	0.1	0.8	1.0	1.0	9387	3717	
λ_2		0.7	0.3	0.1	0.8	1.0	1.0	8299	7027	

According to this limited evidence, SSVSξ approach should be chosen to perform the proposed procedure to estimate the number and location of the knots. Among the three tested models, SSVSξ gives us the best results in terms of estimation of the parameters and in terms of convergence of the algorithm.

Future developments involve (i) fitting more complex models considering also higher degree splines or multidimensional spline representation and (ii) comparing this procedure with alternative Bayesian approaches proposed in the literature (such as those mentioned in the Sec.1.

References

1. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language, J. Stat. Softw., **76**, 1–32 (2017)
2. Denison, D. G. T., Mallick, B. K., Smith, A. F. M.: Automatic Bayesian curve fitting, J. R. Stat. Soc. Ser. B, **60**, 333–350, (1998)
3. DiMatteo, I., Genovese, C. R., Kass, R. E. : Bayesian curvefitting with freeknot splines, Biometrika, **88**, 1055–1071 (2001)
4. O'Hara, R.B., Sillanp, M. J.: A review of Bayesian variable selection methods: what, how and which, Bayesian anal., **4**, 85–117 (2009)
5. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression, Cambridge Series in Statistical and Probabilistic Mathematics, Camb. Univ. Press (2003) doi: 10.1017/CBO9780511755453
6. Smith, M., Kohn, R.: Nonparametric regression using Bayesian variable selection, J. Econom., **75**, 317–343 (1996)

The importance of historical linkages in shaping population density across space

L'importanza della dimensione temporale nella modellizzazione spaziale della densità di popolazione

Ilenia Epifani and Rosella Nicolini

Abstract This study aims to investigate the extent to which history matters in shaping population distribution across space. In the wake of the current literature, the idea is to model individual location preferences by focusing on selected local determinants (neighborhood, education, income, amenities and distance from the CBD), while also taking into account temporal and spatial dependence in location choices. Our preliminary results revel the importance of segregation factors in shaping recent population density distribution.
Abstract *Il presente studio analizza l'importanza della dimensione temporale come determinante della distribuzione spaziale della densità di popolazione. In linea con i contenuti della letteratura contemporanea, la nostra idea è di modellizzare le scelte di localizzazione degli individui in base a delle preferenze strettamente relazionate con alcune determinanti territoriali specifiche (tipo di quartiere, educazione, reddito, amenità presenti nel territorio e distanza dal centro d'interessi -definito come CBD-) in aggiunta a condizioni di autocorrelazione spaziale e di dipendenza temporale del processo di decisione. Quest'ultima, in particolare, garantisce la condizione di coerenza intertemporale delle decisioni prese dagli individui. I nostri primi risultati rivelano che la combinazione delle precedenti determinanti genera un effetto segregazione che assume un ruolo di primo piano per approssimare la distribuzione spaziale della densità di popolazione.*

Key words: Bayesian inference, Hierarchical dynamic model, Population distribution, Spatial random effects

1 Introduction and Data Description

Individual location choices are often driven by environmental factors beyond subjective preferences or budget constraints. As widely discussed in the critical contribution by [8] the characteristics of a neighborhood play an extremely important role in shaping individual choices. Neighborhood features matter both at present but also in throught reputation effects. For instance, think of a neighborhood that consolidates its reputation as a ghetto versus another that recently started hosting people of different ethnic origins. The former is more likely to impact individual location decisions compared to the latter. This study develops around the interplay between the relevance of environmental factors and historical dimensions in location choices

Ilenia Epifani
Politecnico di Milano, Dip. di Matematica, P.zza L. da Vinci, 32, I-20133 Milano e-mail: ilenia.epifani@polimi.it

Rosella Nicolini
Departament d'Economia Aplicada, Universitat Autònoma de Barcelona, Edifici B – Campus UAB, 08193 Bellaterra e-mail: rosella.nicolini@uab.cat

of the population in Massachusetts in a monocentric framework. The idea is to define a dynamic statistical model that allows for accounting for the way past values of some selected location determinants (referring to both a specific place and surrounding territories) may influence the present individual location choice. Thus, we are able to tract potential changes in the relative weight of local determinants and spillover effects in location-choice decisions across time. As argued in [1], [2], the geographical structure of Massachusetts simplifies the setting and, the historical and consolidated attractiveness of Boston as a Central Business District makes it possible to focus on a monocentric distribution function.

We built a Bayesian dynamic Log-Normal Model with random spatial normal frailties to investigate census tract data coming from the NHGIS project [6], the US census and the American Community Survey (ACS) for the 14 counties in Massachusetts for the period 1970-2010. The comparability of the geographical units is based on the information obtained from the TIGER/Lines Finder project. A census tract is identified as an area in which 1500 to 8000 habitants live (with an optimal size of 4000 persons). The tract is often split into two (or more) subtracts when goes over that limit or when the spatial territory is affected by other structural changes. Unfortunately, there is no a clear mechanism that allows for tracking with precision these changes over time, whereas their number increased progressively: census tracts were 1950 in 1970, 1193 in 1980, 1323 in 1990, 1361 in 2000 and 1472 in 2010.

Let $Y_{ij}^{(t)}$ be the population density (per square mile) of the jth census tract within the ith county of Massachusetts at decade t, for $i = 1,\ldots,14$ and $t = 1970, 1980, 1990, 2000, 2010$. A preliminary explorative analysis of $Y_{ij}^{(t)}$ at county level reveals that the sample means per county exhibit temporal patterns different each from either county, but relatively stable within each county. Instead, the standard deviations proportionally shrink except for the most remote counties. These features suggest that the introduction of some spatial county random effects in the model can capture the heterogeneity among counties, whereas, the different patterns of the sample variances suggest a county-stratum variance heteroscedasticity.

In our analysis, the distance $D_{ij}^{(t)}$ of census tract i (in county j at decade t) from Boston is one of the key predictors of the density population and it is defined as a distance between two centroids: one being the centroid of the census tract with the highest population density (in a specific year) and the other being the centroid of any other remaining existing census-tract in the sample for that year. $D_{ij}^{(t)}$ is computed as the Euclidean value according to the geographical coordinates of the two centroids, as suggested by geographers. As our unit of reference (namely, census tracts) changes over time, distance $D_{ij}^{(t)}$ is a time-dependent predictor of the density population. Along with $D_{ij}^{(t)}$, we selected the ethnic composition, education and income as predictors for the population density. In particular, ethnic feature $M_{ij}^{(t)}$ of i th census tract is computed as the proportion of whites (over the total population). Income enters the statistical model via the income $I_{ij}^{(t)}$ per-capita per-census tract and via the Gini index $G_{ij}^{(t)}$ that measures the dispersion and inequality of income. Income $I_{ij}^{(t)}$ is not collected in 1970, so we can not use it in the regression for the first decade. Education in each census tract j enters the model via an index $E_{ij}^{(t)}$ obtained in the following way: first census-tracts were ranked according to the relative frequency of citizens having a high degree of education, then according to the relative frequency of persons having a low degree of education and hence the second value of ranking were substracted from the first. Finally, the presence of amenities is proxied by the proportion Z_i of free land (water areas) in county i: as discussed in [4], water can be considered a fundamental factor (or amenity) in creating recreational spaces for leisure time; moreover, as discussed in [2], it is reasonable to consider Z_i constant over time.

2 Bayesian Dynamic Hierarchical Log-Normal Models

Our empirical analysis is based on an augmented strategy. We begin with estimating a model without any kind of frailties. Then, we introduce some independent county-frailties. Finally, we elaborate a model with frailties having both a temporal and spatial dependence. In the light of previous remarks, we specified the following likelihood function

$$\begin{aligned}\log Y_{ij}^{(t)}|\boldsymbol{w}_t, \boldsymbol{\beta}_t, b_{0t}, b_{1,t}, \sigma_i^2 &\overset{\text{indep.}}{\sim} \\ &\overset{\text{indep.}}{\sim} \mathrm{N}\left(w_{it} + b_{0t} + b_{1t}Z_i + \beta_{1t}D_{ij}^{(t)} + \beta_{2t}M_{ij}^{(t)} + \beta_{3t}G_{ij}^{(t)} + \beta_{4t}E_{ij}^{(t)} + \beta_{5t}I_{ij}^{(t)} + \beta_{6t}I_{ij}^{(t)} \times D_{ij}^{(t)}\ ,\ \sigma_i^2\right), \quad (1)\end{aligned}$$

where $\boldsymbol{w}_t = (w_{1t}, \dots, w_{14t})$ is the vector of 14 county random effects at time t. The $\boldsymbol{w}_t$s' capture the common features shared by the census tracts (in the same county), but they are different from county to county; each w_{it} summarizes all the determinants of the population density, either unobservable or observable but neglected, common to all the census tracts in the ith county.

In order to disentangle the random county effects from the evolution of the coefficients associated with the distance from Boston as unique determinant of population distribution, estimations have been performed in different steps:

(a) running the regression including only distance without frailties (i.e. assuming $\beta_{2t} = \ldots = \beta_{6t} = 0$ and $\boldsymbol{w}_t = \boldsymbol{0}$),
(b) running the regression with all selected covariates without frailties (i.e. assuming only $\boldsymbol{w}_t = \boldsymbol{0}$),
(c) running the regression including only distance with frailties (i.e. assuming $\beta_{2t} = \ldots = \beta_{6t} = 0$),
(d) running the complete regression with frailties.

In all of these specifications, a temporal dependence has been introduced at the fixed effect level, by assuming a random walk for the dynamic regression parameters $\boldsymbol{B}_t := (b_{0t}, b_{1t}, \boldsymbol{\beta}_{1t})$ as in [9] and, we get

$$\boldsymbol{B}_0 \sim \mathscr{N}(\boldsymbol{0}, 10^4 \times I), \quad \boldsymbol{B}_t|\boldsymbol{B}_{t-1},\ \sigma_{b_0}^2, \dots, \sigma_{\beta_6}^2\ \sim \mathscr{N}(\boldsymbol{B}_{t-1}, diag\{\sigma_{b_0}^2, \sigma_{b_1}^2, \sigma_{\beta_1}^2, \dots, \sigma_{\beta_6}^2\}),\ \forall t.$$

This prior is quite comparable with the canonical way to deal with the classical idea of adaptive expectations. Instead, we accomodate the heteroscedasticity in the counties by adopting the following multiplicative structure of the population log-density variances borrowed by [5]:

$$\sigma_{y;i}^2 = \sigma^2 \times v_i, \quad \sigma^2 \sim \text{Inverse Gamma}(0.001, 0.001), \quad v_1, v_2, \dots, v_{14} \overset{iid}{\sim} \frac{\chi_r^2}{r}\,.$$

As for the county random frailties we test two alternative modelizations. To begin with, we assume a simple multiplicative hierarchical structure for $w_{1t}, \dots, w_{14t}$ across both time and counties:

$$w_{it} \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma_w^2), \quad \forall i,\ \forall t\,. \quad (2)$$

Alternatively, we model a spatio-temporal dependence between counties by including a space weight matrix A at a county-level to indirectly link the neighboring counties, via pseudo lagged Conditionally Autoregressive county random effects. The space weight matrix A is introduced as an adjacency matrix. Its cells are organized as follows: $A_{ij} = 1$ if i, j are neighbors, and $A_{ij} = 0$ otherwise. Then, A is used in the modeling of the county random effects:

$$w_{it}|\boldsymbol{w}_{t-1} \sim \mathrm{N}(w_{it-1} + \rho A\boldsymbol{w}_{t-1}, \sigma_{w;i}^2), \quad (3)$$

with initial conditions $w_{i0} \overset{iid}{\sim} \mathrm{N}(0, 10^6)\ \forall i$, and $\sigma^2_{w;i} = \lambda / n_i$ if i has $n_i \geq 1$ neighbors, $\sigma^2_{w;i} = \tilde{\sigma}^2_w$ if i is an isolated (or island) county. Our rational is that census-tracts belonging to a same county share some common features (e.g., urban regulation) while there may or may not be some dependence across counties in the way citizens form their expectations.

We complete the prior distribution for the remaining unknown parameters, adding a diffuse uniform prior distribution for σ_w on the range $(0, 10)$ in case of independent frailties (2). Instead, for dynamic-spatial frailties (3) we choose: $\tilde{\sigma}_w \sim Uniform(0, 10)$, $\lambda \sim$ Inverse Gamma$(0.5, 0.0005)$, $\rho \sim Uniform(1/l_{min}, 1/l_{max})$ where l_{min}, l_{max} are the minimum and maximum eigenvalues of the restricted spatial weight matrix A when excluding the two island-counties: Nantucket and Dukes. See [5] and the Manual of GeoBUGS, 2014.

3 Model Selection and Estimation

In order to assess the goodness-of-fit of our model, we computed the Bayesian p-values of the six alternative specifications on the base of the discrepancy measure by [3], for the density population on the logarithm scale: $\sum_{i,j,t} \Big(log(Y^{(t)}_{i,j}) - \mathrm{E}(\log(Y^{(t)}_{i,j})\,|\mathbf{Data}) \Big)^2 / \operatorname{Var}(\log(Y^{(t)}_{i,j})\,|\mathbf{Data})\,)$. Then, a Bayesian comparison of the six alternative specifications were performed by means of the DIC of each model and, for each model and each year, bu means of the percentage of the Bayesian census tracts outliers, i.e. of the census tracts whose actual population density falls into one of the two 2.5 percent tails of the marginal posterior predictive density. Overall, Bayesian-p value suggest that all our specifications perform relatively well by producing relatively good fits, whereas the DIC values reveal that the best empirical specification has to include ethnic composition, education, income along with the physical distance to Boston and the county random frailties. Both the "complete" models with county random frailties have quite similar DIC values, which are definitely better than the model with only distance and without frailties. Under this perspective, one can conclude that there is not a relevant difference in adopting a dynamic-spatial frailty-strategy rather than independent frailties. The Bayesian percentage outliers for each of the six models remains quite low (always under 5 percent) from 1980 to 2010. Instead, for the year 1970, results are at odds with the ones for the other decades.

For 1970, no covariate but distance to Boston seems to be able to model population density distribution. This result confirms that from 1970 to 1980 the land organization in Massachusetts passed through important structural changes, as anticipated by [2]. Figure 1 shows a clear tendency to identify the distance from Boston as the principal driver of population density distribution. Estimations referring to 1970 are a bit at odds with respect to the other decades. Instead, from 1980 onward, the results are consistent. The ethnic composition of the neighborhood and the level of income are two dominant and constant effects. High-income persons tend to settle in the area in which they can enjoy the possibility to rent or buy individual dwelling properties: these places are located far from Boston. In 2010 differences in education among the various census tracts also seem to become a statistically important discrimination factor in location choices, moving in the same direction as ethnic composition and income. County random frailties make the presence of natural amenities (here Z) ineffective. The interaction term between income and distance amplifies the attractiveness of each tract-unit, emphasizing the presence of a mass effect for a number of selected features (in this case, income). The idea is to capture the attractiveness of a destination focusing on some selected features that intervene in shaping individual preferences for location choices despite the physical distance from the CBD.

As far as the heteroscedasticity of the counties, we notice from the plots in Figure 2(a) that all of the econometric exercises provide consistent estimations of the clustered-per-counties heteroscedasticity of the census population density, with the highest variance in the most remote counties: Norfolk, Plymouth, Essex, Dukes, Suffolk, Middlesex and Barnstable are heteroscedastic with a low level of variance, Hampden,

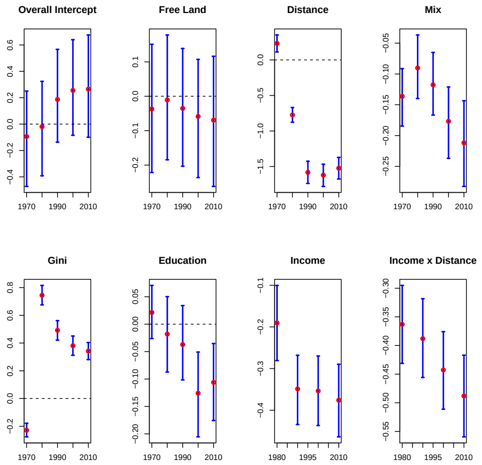


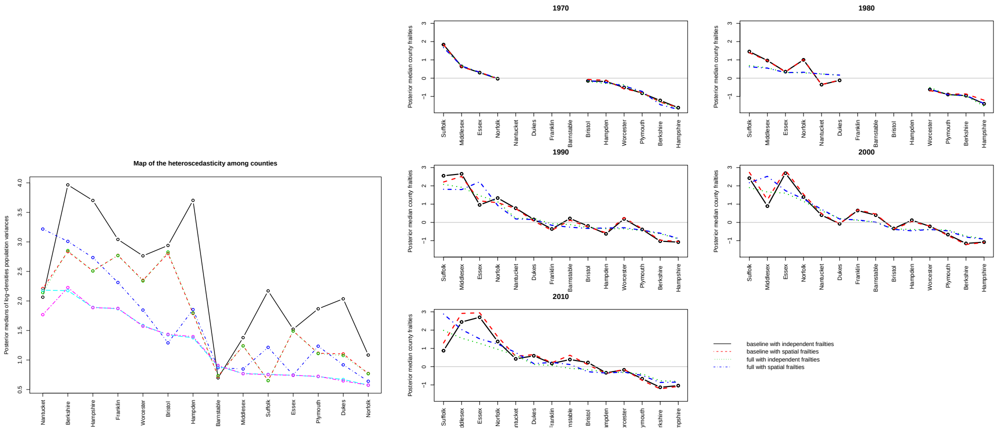
Fig. 1 Plot of the Bayesian estimates (posterior means) and posterior 90 percent Intervals of the regression coefficients in the *Complete Model with Independent Frailties* from 1970 to 2010. Data in 1970 for *Income* are not available. Legend: solid red circles: posterior mean.

Bristol and Worcester have common variance, whereas Hampshire, Franklin, Berkshire and Nantucket are heteroscedastic with a high level of variance.

With regard to the county random heterogeneity, the picture of the posterior median values of the 14 county random frailties w_{it} in Figure 2(b) reveal that, on one hand the fitted frailties are quite identical under the two specifications, on the other hand, adding new local determinants of the density population along with the distance from Boston, compresses the frailty-predicted value toward zero. Although, under a temporal perspective, each frailty follows an autonomous evolutionary path, still, it is possible to detect some regularities. A structural break is definitely present between the first two decades (i.e. 1970 and 1980) and the rest (i.e., 1990, 2000 and 2010). Referring to the second period, the most remote counties (with respect to Suffolk) show frailty values that continue to be negative across time: the core counties (namely, those closest to Boston) display a strictly negative estimated median value, and the remaining countries –all far from Boston– either approach zero or are strictly positive. Hence, this study confirms that the populations of towns belonging to remote counties far from Boston consolidated low attractiveness in relation to Boston as the CBD across time.

4 Concluding Remarks

In this paper we propose a novel study to assess the determinants of population density distribution across space in Massachusetts. Our empirical strategy relies on the definition of a Bayesian framework in which we take into account temporal and county-spatial dependence for the assessment of population distribution at a census tract level over the period 1970-2010. Our main results confirm that the physical distance from Boston is the relative dominant covariate in defining population distribution across time. However, we detect an increasing segregation effect in modeling population distribution across time: the ethnic and income characteristics of a neighborhood exhibit a relevant impact in the location choice of citizens. These discriminating features are strong enough to be robust and keep being relevant even in the presence of county random frailties aimed at capturing county features not directly embedded. Nevertheless, the county spatial dependence (built around the county frailty effects) does not improve the quality of the estimation if it is compared to the case of spatial-independent frailties. As for preliminary insights delivered from our results, one can conclude that the driving forces of population distribution are mostly associated with elements featuring segregation effects rather than spatial distance to a place of interest. Hence, potential actions aimed at smoothing population concentration (or congestion effects) should target the impact of segregation features rather than centering on accessibility issues. In policy terms, these new outcomes could deliver interesting suggestions for implementing policies aimed at reducing concentration or congestion effects as well as limiting the formation of spatial areas suffering from segregation effects.



(a) (b)

Fig. 2 Posterior Medians of the of the population log-density variances in Figure 2(a) and of the county random frailties in Figure 2(b) from 1970 to 2010. (Non available census-tracts in counties Barnstable, Franklin, Hampden and Norfolkfor for 1970, 1980). Counties are sorted horizontally in decreasing order of their posterior frailty medians under model (2) with all determinants.

References

1. Epifani, I., Nicolini, R.: On the density distribution across space: a probabilistic approach. J. Reg. Sci. **53**, 481-510 (2013)
2. Epifani, I., Nicolini, R.: Modelling population density over time: how spatial distance matters. Reg. Stud. **51**, 602-615 (2017)
3. Gelman, A., Carlin, J., Stern, H., Rubin, D.B.: Bayesian Data Analysis. CRC Press, Boca Raton, FL (1995)
4. Glaeser, E., Ward, B.A.: The causes and consequences of land use regulation: evidence from the Greater Boston. J. Urban Econ. **65**, 265-278 (2009)
5. LeSage J.P., Pace, K.R.: Introduction to Spatial Econometrics. CRC Press, Boca Raton, FL (2009)
6. Minnesota Population Center (2011). National Historical Geographic Information System: Version 2.0., Minneapolis, MN: University of Minnesota
7. Quigley, J.M.: Consumer Choice of Dwelling, Neighborhood and Public Services. Reg. Sci. Urban Econ. **5**, 41-63 (1985)
8. Topa, G., Zenou, Y.: Neighbourhood Effects versus Network Effects. In: Duranton, G., Henderson, J.V., Strange, W. (eds.) Handbook of Regional and Urban Economics **5**, pp. 561-624. Elsevier Publisher, Amsterdam (2015)
9. West, M., Harrison, P.J.: Bayesian Forecasting and Dynamic Models. 2nd edition, Springer, New York (1997)

Recent Developments in Sampling

Species richness estimation exploiting purposive lists: A proposal

Stima della ricchezza specifica utilizzando liste mirate. Una proposta

A. Chiarucci[1], R.M. Di Biase[2], L. Fattorini[3], M. Marcheselli[4] and C. Pisani[5]

Abstract The lists of species obtained by purposive sampling can be used to improve the sample-based estimation of species richness. A new estimator is proposed as a modification of the difference estimator in which the species inclusion probabilities are estimated by means of the species frequencies from incidence data. An asymptotically conservative estimator of the mean squared error is also provided.

Abstract *Le indagini mirate producono liste floristiche che possono essere usate per migliorare la stima della ricchezza specifica ottenuta tramite campionamento probabilistico. Viene qui proposto un nuovo stimatore della ricchezza specifica basato sullo stimatore per differenza, stimando le probabilità di inclusione sulla base dei dati di incidenza. Viene anche proposto uno stimatore asintoticamente conservativo dell'errore quadratico medio.*

Key words: Difference estimator, probabilistic sampling, purposive survey, supporting list, species richness.

[1] Alessandro Chiarucci, Department of Biological, Geological, and Environmental Sciences, University of Bologna; email: alessandro.chiarucci@unibo.it

[2] Rosa Maria Di Biase, Department for innovation in Biological, Agro-food and Forest systems, University of Tuscia; email: dibiase.rm@gmail.com

[3] Lorenzo Fattorini, Department of Economics and Statistics, University of Siena; email:lorenzo.fattorini@unisi.it

[4] Marzia Marcheselli, Department of Economics and Statistics, University of Siena; email:marzia.marcheselli@unisi.it

[5] Caterina Pisani, Department of Economics and Statistics, University of Siena; email:caterina.pisani@unisi.it

1 Introduction

Species richness, i.e. the number of species in a biological community, represents the simplest and most direct indicator of ecological diversity, largely used as the most convenient proxy for other components of biodiversity [6]. In the following, we will refer to plants, but analogous reasoning could be applied to limited mobility animals.

Species richness is an unknown parameter of the community under study, especially in large areas, and it can be evaluated by means of a purposive survey, as traditionally performed by ecologists, or estimated through probabilistic sampling.

In purposive sampling, species are recorded and listed by searching into specific sites expected to have a large number of species, high detection rates or high abundance of rare species. However, this is a subjective approach, so it does not allow any probabilistic statement about the accuracy and precision of species richness estimators.

On the other hand, in probabilistic sampling, species are identified and listed only if present in the selected samples. The estimates can be objectively evaluated through their sampling distributions, thus allowing for reliable comparisons across areas [e.g., 5]. Nevertheless, sample-based strategies *"are likely to miss the rare or unclassifiable habits that are likely to contribute most to regional diversity. ... Indeed, it is unlikely that such methods can outperform the guesses of experienced botanists"* [7, page 122]. Also in [7, page 122], the importance of probabilistic sampling in comparing species richness throughout time and space is recognized, even if *"it would be unwise to dismiss the efficient, yet subjective contributions of the expert botanist"*. However, at least to our knowledge, the species lists compiled by means of purposive surveys have never been used to improve species richness estimates arising from probabilistic sampling.

Therefore, we introduce a new species richness estimator, referred to as empirical difference (ED) estimator, exploiting both sources of information [2].

Section 2 contains some notations, while the ED estimator and a presumably asymptotic conservative estimator of its mean squared error are presented in Section 3. Section 4 contains a brief discussion on the improvement provided by list exploitation and concluding remarks.

2 Notation and setting

Consider a plant community within a delineated study area, so that each group of individual plants belonging to the same species may be viewed as a unit and the complete species list can be viewed as a population.

Referring to the species by their identifying numerical labels, the complete list of species can be represented by the set $\mathsf{C} = \{1, \ldots, K\}$. Since the complete list is usually unknown, the species richness K must be estimated. It should be noticed that species cannot be directly sampled because they constitute unknown assemblages of individual plants spread over the study area. Thus, the most effective way for sampling species is to sample individual plants, so that a species is sampled when at least one plant of that species is sampled.

The quantification of the first-order inclusion probabilities of species $\theta_1,\ldots,\theta_K$ entails the knowledge of all the units belonging to each species and their spatial distribution over the study area [5]. Consequently, even if the adopted sampling scheme ensures that the first-order inclusion probabilities can be determined directly or by some field measurements [e.g., 5] for (at least) the selected plants, the species inclusion probabilities cannot be quantified.

A study area cannot be adequately sampled by means of only one plot or transect and for this reason *n* independent replications of the sampling scheme [1] are usually performed, determining *n* samples of plants, which in turn give rise to *n* samples of species $\mathsf{G}_1,\ldots,\mathsf{G}_n$. The set of species observed in the whole survey is $\mathsf{G}_{(n)}=\bigcup_{i=1}^{n}\mathsf{G}_i$ and its size SO_n is the number of observed species.

For each replication *i*, $\mathbf{z}_i=\left[z_{i1},\ldots,z_{iK}\right]^{\mathsf{T}}$ is the *K*-vector in which the *j*th element z_{ij} is equal to 1 if the species *j* has been sampled and 0 otherwise. Usually, $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are organized into a 0-1 matrix of *n* columns and SO_n rows, the *presence-absence* or *incidence data*. These *n* vectors are independent realizations of the random vector $\mathbf{Z}=\left[Z_1,\ldots,Z_K\right]^{\mathsf{T}}$ with expectation $\boldsymbol{\theta}=\left[\theta_1,\ldots,\theta_K\right]^{\mathsf{T}}$. Let $\mathbf{x}=\sum_{i=1}^{n}\mathbf{z}_i$ be the realization of the random vector $\mathbf{X}=\left[X_1,\ldots,X_K\right]^{\mathsf{T}}$, with $X_j\sim B\left(n,\theta_j\right)$. Because x_j is the number of replications in which the species *j* has been sampled, $x_j=0$ for all the undetected species. Thus, even if virtually $\mathbf{x}$ is a *K*-vector, it contains an unknown number of zeros [3].

3 The empirical difference estimator

The most common way to exploit auxiliary information is the difference (D) estimator and its modifications, such as the widely used generalized regression and ratio estimators [8, chapter 6]. In this paper, the ED estimator is yet another modification of the D estimator which uses as auxiliary information the species list, henceforth referred to as supporting list L.

Consider *Y* and Y^0 dichotomous variables such that $y_j=1$ for each species $j\in\mathsf{C}$ and $y_j^0=1$ if $j\in\mathsf{L}$ and 0 otherwise, so that Y^0 can be adopted as a proxy for the survey variable *Y*. When the supporting list is accurate, Y^0 is a good proxy for *Y*, given that the errors $y_j-y_j^0=1-y_j^0$ are equal to 0 for any species $j\in\mathsf{L}$.

Exploiting the y_j^0 s, $K=\sum_{j\in\mathcal{C}} y_j$ can be rewritten [8, chapter 6] as

$$K=\sum_{j\in\mathcal{C}} y_j^0+\sum_{j\in\mathcal{C}}\left(y_j-y_j^0\right)=M+\sum_{j\in\mathcal{C}}\left(1-y_j^0\right)$$

where M is the number of species in the supporting list and the second member is unknown and needs to be estimated from the sample using the Horvitz-Thompson criterion

$$\hat{K}_D=M+\sum_{j\in G_{(n)}}\frac{1-y_j^0}{\tau_j}=M+\sum_{j\in G_{(n)}-\mathcal{L}}\frac{1}{\tau_j}$$

where $\tau_j=1-\left(1-\theta_j\right)^n$ is the probability that species j is detected during the whole sample survey. The estimator $\hat{K}_D$ is unbiased with a closed-form variance, which could be unbiasedly estimated from the sample. However, it cannot be calculated since the θ_j s (and consequently the τ_j s) are unknown. The frequencies x_j s in which the species enter the n samples can be adopted to estimate the θ_j s as $\hat{\theta}_j=\left(x_j+1\right)/\left(n+1\right)$ [4], from which $\hat{\tau}_j=1-\left(n-x_j\right)^n/\left(n+1\right)^n$.

The ED estimator turns out to be:

$$\hat{K}_E=M+\sum_{j\in G_{(n)}-\mathcal{L}}\frac{1}{\hat{\tau}_j}$$

$\hat{K}_E$ is biased with expectation and variance which cannot be expressed in closed form. However, as opposite to other species richness estimators, its realizations are never smaller than the cardinality of the set $G_{(n)}\cup\mathcal{L}$. Furthermore, if the supporting list is perfect, the ED estimator invariably estimates the true species richness without error. Hence, besides the uncertainty due to the estimation of the inclusion probabilities, the uncertainty of $\hat{K}_E$ is completely due to the species in the set $\mathcal{C}-\mathcal{L}$, i.e. the species lost in the supporting list which can be partially recovered by the sample survey. It should be noticed that if $G_{(n)}-\mathcal{L}$ is the empty set, the ED estimator coincides with M.

Because the estimator is biased, there is no sense in estimating its variance. Rather, we should estimate the relative root mean square error

$$\mathrm{RRMSE}=\frac{\sqrt{\mathrm{MSE}\left(\hat{K}_E\right)}}{K}=\frac{\sqrt{\mathrm{E}\left\{\left(\hat{K}_E-K\right)^2\right\}}}{K}$$

Because neither of the mean nor the variance of $\hat{K}_E$ can be expressed in a closed form, we derived an upper bound of $\mathrm{MSE}\left(\hat{K}_E\right)$ to be subsequently estimated from the available information, in such a way that the resulting estimator should be presumably asymptotically conservative. In [2, appendix] it is proved that

$$\mathrm{MSE}(\hat{K}_E)\leq 2K(4e^{-1}+1)\sum_{j\in\mathcal{C}-\mathcal{L}}\left\{1-\theta_j(1-e^{-1})\right\}^n$$

whose right side can be estimated by

$$\hat{\mathrm{MSE}}(\hat{K}_E) = 2\hat{K}_E(4e^{-1}+1)\sum_{j\in G_{(n)}-L}\frac{\left\{1-\hat{\theta}_j(1-e^{-1})\right\}^n}{\hat{\tau}_j}$$

Also in this case, if $G_{(n)}-L$ is the empty set, the $\hat{\mathrm{MSE}}$ turns out to be 0. Consequently,

$$\mathrm{RR}\hat{\mathrm{MSE}}(\hat{K}_E) = \frac{\sqrt{\hat{\mathrm{MSE}}(\hat{K}_E)}}{\hat{K}_E} = \sqrt{\frac{2(4e^{-1}+1)}{\hat{K}_E}\sum_{j\in G_{(n)}-L}\frac{\left\{1-\hat{\theta}_j(1-e^{-1})\right\}^n}{\hat{\tau}_j}}$$

It should be noticed that the ED estimator is a consistent estimator of *K* with bias and variance approaching 0 as the number of replications increases. Indeed, from the $\mathrm{MSE}(\hat{K}_E)$ inequality follows that $\hat{K}_E$ converges in quadratic mean (and hence also in mean) to *K*, since $\lim_{n\to\infty} 2K(4e^{-1}+1)\sum_{j\in C}\left\{1-\theta_j(1-e^{-1})\right\}^n(1-y_j^0)=0$.

4 Discussion

In order to check the improvement provided by the exploitation of floristic lists, a simulation study was performed. $\hat{K}_E$ was compared to other species richness estimators, using nonparametric estimators of species richness.

The results of the simulation (for the full explanation of the simulation process and its results see [2]) show noticeable improvement in bias and precision provided by the proposed estimator, especially when the supporting list is accurate. Practically speaking, the proposed estimator is likely to be effective when efforts in compiling lists are mostly directed toward habitats that are likely to host rare species. It should be noticed that lists missing the most common species are not unrealistic, because botanists are often focused on searching the rarest species, sometimes neglecting the most common ones [7].

If most of the rare species are included in the supporting list, the ED estimator can represent an efficient solution and the RRMSE estimator is presumably conservative, since some underestimations only occur when the supporting lists miss rare species. At least to our knowledge, this is the first attempt to estimate the MSE in species richness estimation, as most papers dealing with this problem propose estimators of the sampling variance. However, the estimation of variance in species richness estimation is irrelevant, because the major part of the sampling error is due to bias.

It is worth noting that it is important to work with updated floristic lists, so as to avoid the inclusion of species no longer in the area and other taxonomical problems (e.g. synonyms of species that were split into two or species that were merged). Thus, these drawbacks could deteriorate the performances of estimators exploiting floristic lists.

References

1. Barabesi, L., Fattorini, L.: The use of replicated plot, line and point sampling for estimating species abundances and ecological diversity. Environ. Ecol. Stat. (1998) doi: 10.1023/A:1009655821836
2. Chiarucci, A., Di Biase, R.M., Fattorini, L., Marcheselli, M., Pisani, C.: Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness. Ann. Appl. Stat. Forthcoming
3. D'Alessandro, L., Fattorini, L.: Resampling estimators of species richness from presence-absence data: Why they don't work. Metron, **60**, 5-19 (2002)
4. Fattorini, L.: Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. Biometrika (2006) doi: 10.1093/biomet/93.2.269
5. Fattorini, L.: Statistical inference on accumulation curves for inventorying forest diversity: A design-based critical look. Plant Biosyst. (2007) doi: 10.1080/11263500701401786
6. Gaston, K.J.: Species richness: measure and measurement. In: Gaston, K.J. (ed.) Biodiversity. A Biology of Numbers and Difference, pp. 77–113. Blackwell Science, Oxford (1996)
7. Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S., Wohlgemuth, T.: Quantitative tools for perfecting species lists. Environmetrics. (2002) doi: 10.1002/env.516
8. Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992)

Design-based exploitation of big data by a doubly calibrated estimator

Uno stimatore a calibrazione doppia per l'uso dei big data in un'ottica basata sul disegno

Maria Michela Dickson [1], Giuseppe Espa[2] and Lorenzo Fattorini[3]

Abstract Big data typically constitute masses of unstructured data, not always available for a whole population. When sampling only the sub-population where big data are available, but neglecting the remaining portion, this can be viewed as a fixed component of nonresponses, which sums the natural component of nonresponses present in each survey. In this paper, big data information is exploited to handle nonresponse, while a size variable available for the whole population is exploited to handle the neglected part of the population by means of a doubly calibrated estimation. Design-based expectation and variance are derived up to the first order approximation. A variance estimator is proposed. A Monte Carlo simulation exploring various scenarios demonstrates the efficiency of the strategy.

Abstract *I big data costituiscono una mole di dati non strutturata, non sempre disponibile per tutte le unità di una popolazione. Quando si campiona solo dalla sotto-popolazione per cui i big data sono disponibili, trascurando la restante parte, questo può essere visto come una ulteriore fonte di mancate risposte che si aggiunge a quella naturalmente presente in ogni indagine campionaria. Nel presente lavoro, viene proposto uno stimatore a doppia calibrazione, nel quale i big data vengono utilizzati per gestire le mancate risposte, mentre, per gestire la parte di popolazione trascurata nella selezione, viene utilizzata una variabile dimensionale disponibile per*

[1] Maria Michela Dickson, Department of Statistical Sciences, University of Padova; email: mariamichela.dickson@unipd.it

[2] Giuseppe Espa, Department of Economics and Management, University of Trento; email: giuseppe.espa@unitn.it

[3] Lorenzo Fattorini, Department of Economics and Statistics, University of Siena; email: lorenzo.fattorini@unisi.it

l'intera popolazione. Valore atteso e varianza approssimati sino al primo ordine sono derivati in un'ottica completamente basata sul disegno. Inoltre si propone uno stimatore della varianza. Infine, mediante simulazioni Monte Carlo, vengono studiati scenari differenti per dimostrare l'efficienza della strategia proposta.

Key words: Auxiliary variables; Big data; Calibration estimators; Simulation study.

1. Introduction

In the last twenty years the availability of data has hugely increased, making possible developments in many fields of research, primarily in statistical sciences. More recently, this increase in data availability is also characterized by an increase in size of the amount of information collected, opening an extended debate around the term *big data*.

This new and potentially infinite source of data is connoted, on one side, by a not definite frame and, on the other side, by a real-time updating. Clearly the mentioned features represent pros and cons that researchers and practitioners must consider when using big data for statistical analysis (Tam, 2015). While the large amount of information is an unquestionable positive issue, the lack of a frame makes difficult the definition of a population of interest. This matter became relevant in sampling theory and in the consequent inference that can be done under a design-based perspective.

Nevertheless, in many practical circumstances, the opportunity of exploiting big data information may be an advantage in surveys. At the same time, considering only units provided by these additional information, could lead to a missed observation of the other units, which remain excluded from the study. For instance, it happens in socio-economic surveys on people living conditions, which tend to exclude units that cannot be contacted from the population, or in environmental studies, in which remote sensing information are not available for some areas. In this framework, these units never enter the sample and can be viewed as a fixed component of nonresponses, which sums the natural component of nonresponses present in each survey.

The aim of the present paper is to reach the desirable chance to take advantage from big data, when available, but make inference on a whole population in the above mentioned practical circumstances. So that, we propose an estimator that may permit to achieve this goal, by means of calibration estimators (Deville and Särndal, 1992). First, a calibration is used to correct for nonresponses from the sample to the population provided by big data, and, then, a second calibration is implemented on the whole population, leading to a *doubly calibrated estimator*. For the proposed estimator, expectation and variance are derived and a variance estimator is proposed. The efficiency and applicability of the strategy is showed by a Monte Carlo simulation study on a population that mirrors real characteristics.

2. Preliminaries, notation and methods

Let $U = \{u_1, \dots, u_N\}$ a population of N units. We denote with $y_j, j \in U$ the value for unit j of a survey variable Y. The aim is to estimate the population total

$$T_Y = \sum_{j \in U} y_j.$$

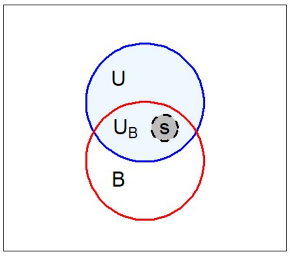
An auxiliary size variable Z is also available for each unit of U, such that the total $T_Z = \sum_{z \in U} z_j$ is known for any $z_j, j \in U$.

Moreover, a big data population B of M units intersects population U, so that we can define U_B as a sub-population of U composed by $N_B < N$ units. The total $T_{Y(B)} = \sum_{j \in U_B} y_j$ in U_B is unknown and it is the first quantity to be estimated in the procedure.

A sample S of size $n < N_B$ may be selected from the sub-population U_B by means of a fixed-size design having first and second order inclusion probabilities π_j, π_{jh} for any $h > j \in U_B$. As always happen in practice, the sample may be affected by nonresponses, so we define $R \subset S$ as the respondent sample. Note that the sampling scheme adopted to select S generates a sampling design on U_B but not on U.

To perform the first step of calibration, auxiliary information for U_B units are necessary. Let $\boldsymbol{x}_i = [x_{i1}, \dots, x_{iK}]$, with $i \in B$, the $\boldsymbol{X}$-vector of K auxiliary variables available in the population B of big data. The totals $\boldsymbol{T}_{X(B)} = \sum_{j \in U_B} \boldsymbol{x}_j$ and $T_{Z(B)} = \sum_{z \in U_B} z_j$ are known for all units in U_B. To better clarify the population setup, see Figure 1.

Figure 1: A stylized configuration of the population U_B, as the intersection between U and B, and sample S.



Because sample S is drawn from U_B, the H-T estimator (Horvitz and Thompson, 1953) of the total is $\hat{T}_{Y(B)} = \sum_{j \in S} \frac{y_j}{\pi_j}$ which is an unbiased estimator of $T_{Y(B)}$. Therefore, it would be a biased estimator of T_Y. However, considering nonresponses, the estimator

$$\hat{T}_{Y(B)/R} = \sum_{j \in R} \frac{y_j}{\pi_j} \neq \hat{T}_{Y(B)}$$

is a biased estimator even of $T_{Y(B)}$. In order to reduce the bias, following results proposed in Fattorini et al. (2013), it is possible to exploit auxiliary information $\boldsymbol{X}$ under a design-based point of view, obtaining the calibration estimator

$$\hat{T}_{Y(B)(cal)} = \hat{\boldsymbol{b}}_R^t \boldsymbol{T}_{X(B)}$$

where $\hat{\boldsymbol{b}}_R = \hat{\boldsymbol{A}}_R^{-1}\hat{\boldsymbol{a}}_R$ is the least-square coefficient vector of the regression of Y variable on $\boldsymbol{X}$ variables, performed on the respondent sample R, i.e. $\hat{\boldsymbol{A}}_R = \sum_{j\in R}\frac{x_j x_j^t}{\pi_j}$ and $\hat{\boldsymbol{a}}_R = \sum_{j\in R}\frac{y_j x_j}{\pi_j}$. The design-based properties of $\hat{T}_{Y(B)(cal)}$ were derived in Fattorini et al. (2013). In that paper, it has been demonstrated that the estimator is approximately unbiased and consistent if the relationship between Y and $\boldsymbol{X}$ is similar in both respondent and non-respondent sub-groups, and it has been derived variance estimation. So that, the authors provide a design-based solution to the problem of nonresponses.

The additional problem here is that we must estimate T_Y rather than $T_{Y(B)}$. However, because $\hat{T}_{Y(B)(cal)}$ is, at his best, an approximately unbiased and consistent estimator of $T_{Y(B)}$, it is a biased estimator of T_Y. Since the selection of S is only on U_B, the elements of $U - U_B$ cannot enter the sample. In this case, it is necessary to correct the estimator to reduce the bias due to the sample under-coverage. A choice may be to use the size variable Z, which is available for the whole population U. So that, the resulting *double calibration* estimator turns out to be

$$\hat{T}_{Y(\mathrm{d}cal)} = \frac{\hat{T}_{Y(B)(cal)}}{\hat{T}_{Z(B)}} T_Z = \frac{\hat{\boldsymbol{b}}_R^t \boldsymbol{T}_X}{\hat{T}_{Z(B)}} T_Z$$

where $\hat{T}_{Z(B)} = \sum_{j\in S}\frac{z_j}{\pi_j}$ is the H-T estimator of T_Z.

The double calibration estimator benefits of some desirable properties deriving from calibration but, for the sake of brevity, we do not report all proofs. However, following results of Fattorini et al. (2013), $\hat{T}_{Y(dcal)}$ is approximately unbiased if (i) the linear relationship between Y and $\boldsymbol{X}$ is approximately the same in the respondent and non-respondent sub-groups of U_B, and if (ii) the proportional relationship between Y and Z is approximately the same in the two sub-populations U_B and $U - U_B$. Under these conditions, following the consistency of the H-T estimator (see Isaki and Fuller, 1982) it is possible to derive that $\hat{T}_{Y(dcal)}$ converges in probability to T_Y.

Regarding variance and its estimation, following Särndal et al. (1992, p.175) the estimator $\hat{T}_{Y(dcal)}$ has an approximate variance equal to

$$V\left(\hat{T}_{Y(dcal)}\right) \approx \left(\frac{T_Z}{T_{Z(B)}}\right)^2 \sum_{h>j\in U_B} \left(\pi_j\pi_h - \pi_{jh}\right)\left(\frac{u_j}{\pi_j} - \frac{u_h}{\pi_h}\right)^2$$

Given that, the Sen-Yates-Grundy variance estimator (Sen, 1953; Yates and Grundy, 1953) is given by

$$\hat{V}_{SYG}\left(\hat{T}_{CAL}\right) = \left(\frac{T_Z}{\hat{T}_{Z(B)}}\right)^2 \sum_{h>j\in S} \frac{\left(\pi_j\pi_h - \pi_{jh}\right)}{\pi_{jh}}\left(\frac{\hat{u}_j}{\pi_j} - \frac{\hat{u}_h}{\pi_h}\right)^2$$

where $\hat{u}_j = \left(r_j y_j \boldsymbol{x}_j^t - r_j \hat{\boldsymbol{a}}_R^t \hat{\boldsymbol{A}}_R^{-1} \boldsymbol{x}_j \boldsymbol{x}_j^t - z_j \hat{T}_{Z(B)}^{-1} \hat{\boldsymbol{a}}_R^t\right)\hat{\boldsymbol{A}}_R^{-1}\boldsymbol{T}_x$, $j \in S$ are the empirical *influence values* (Davison and Hinkley, 1997) computed for each $j \in S$.

In order to expose the validity of the proposed methodology, in the next section a simulation study is presented.

3. Simulation study

A Monte-Carlo simulation is discussed in this section to investigate the performances of the proposed estimator. A population of $N = 10000$ units has been considered. No random model was adopted for generating nonresponses. The population covered by big data U_B constituted by 7500 units has been partitioned in respondent and non-respondent sub-groups. Therefore, the response pattern is a fixed characteristic of the units, just like the value of the survey variable. The size of respondent sub-group is equal to $N_R = 2250, 4500, 6750$ units, corresponding to 30%, 60% and 90% of the population units. We suppose available two auxiliary variables X_1 and X_2 for units included in U_B. These variables have been generated according to a normal distribution with mean equal to 1, variance equal to 1 and correlation coefficient equal to 0.20. The variable under estimation Y has been generated as

$$y_j = 1 + 0.5x_{1j} + 0.5x_{2j} + \varepsilon_j, \quad \forall j \in U$$

where ε_j is an error component with mean equal to 0 and variance constant, such that the model explains in one case, the 60% and, in another, the 90% of the variance of the y_js. In addition, a size variable Z is available for the whole population and it has been generated as $z_j = 2y_j + \gamma_j$, where γ_j is an error component with mean equal to 0 and variance proportional to $k|Y|$, with k set to assure correlation between Y and Z equal to 0.90. From the described populations, 10000 Monte-Carlo random samples of size $n = 100, 150, 250, 375, 500$ units have been selected by means of simple random sampling without replacement (SRSWOR). For each sample, relative Root Mean Square Error (rRMSE) and relative Bias (rB) have been computed. Table 1 and Table 2 report obtained results.

Table 1: Relative bias and relative RMSE for the population with model $R^2 = 0.60$ and correlation between Y and Z equal to 0.90

	$N_R = 2250$		$N_R = 4500$		$N_R = 6750$	
n	***rB***	***rRMSE***	***rB***	***rRMSE***	***rB***	***rRMSE***
100	0.0349	0.2831	0.0310	0.2252	0.0265	0.2026
150	0.0232	0.2186	0.0171	0.1717	0.0142	0.1509
250	0.0110	0.1579	0.0076	0.1273	0.0096	0.1136
375	0.0088	0.1267	0.0029	0.0998	0.0054	0.0908
500	0.0081	0.1086	0.0021	0.0855	0.0011	0.0764

Table 2: Relative bias and relative RMSE for the population with model $R^2 = 0.90$ and correlation between Y and Z equal to 0.90

	$N_R = 2250$		$N_R = 4500$		$N_R = 6750$	
n	***rB***	***rRMSE***	***rB***	***rRMSE***	***rB***	***rRMSE***
100	0.0266	0.1915	0.0254	0.1805	0.0231	0.1768
150	0.0185	0.1517	0.0153	0.1422	0.0128	0.1351
250	0.0709	0.1095	0.0077	0.1065	0.0097	0.1031
375	0.0058	0.0895	0.0033	0.0835	0.0061	0.0829
500	0.0057	0.0764	0.0030	0.0718	0.0021	0.0701

Conclusions

Results show that in both populations, both relative Bias and relative RMSE decrease as sample size response portion increases. We have explored the case in which the size variable and the target variable are strongly correlated, confirming results gathered by Fattorini et al. (2013). Clearly, when the value of R^2 of the model used to generate Y is higher, performances of the estimator are better. However, we have considered very low sampling fractions and, nevertheless, the relative Bias rapidly decreases, becoming very close to zero with a sampling fraction equal to 5%. The behaviour of rRMSE complies this result, decreasing when the sample size and the size of respondents sub-group increase.

References

1. Davison, A.C., Hinkley, D.V.: Bootstrap methods and their application. Vol. 1. Cambridge university press (1997).
2. Deville. J.-C., Särndal C.-E.: Calibration estimators in survey sampling. J. Am. Stat, Assoc. 87. 376–382 (1992).
3. Fattorini, L, Franceschi, S., Maffei, D.: Design-based treatment of unit nonresponse in environmental surveys using calibration weighting. Biom. J., 55, 925-943 (2013).
4. Horvitz, D. G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat, Assoc. 47. 663-685 (1952).
5. Isaki, C.T., Fuller, W.A.: Survey design under the regression superpopulation model. J. Am. Stat. Assoc. 77. 89-96 (1982).
6. Särndal, C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992).
7. Sen, A.R.: On the estimate of variance in sampling with varying probabilities. J. Indian Soc. Agric. Statist., 5, 119-127 (1953).
8. Tam, S.M.: A statistical framework for analysing big data. The Survey Statistician. 72. 36-51 (2015).
9. Yates, F., Grundy, P.M.: Selection without replacement from within strata with probability proportional to size. J. R. Statist. Soc. B, 15, 235-261 (1953).

Design-based mapping in environmental surveys

Mappe basate sul disegno nelle indagini ambientali

L. Fattorini, M. Marcheselli and C. Pisani

Abstract The estimation of the values of a survey variable throughout a continuum of points or in a finite population of spatial units is investigated when a sample of points or units is selected by a probabilistic sampling scheme. At each point or for each spatial unit, the value is estimated using an inverse distance weighting interpolator and conditions ensuring its design-based asymptotic unbiasedness and consistency are summarized.
Abstract *La stima dei valori di una variabile di interesse in un continuum di punti o in una popolazione finita di unità spaziali è considerata quando un campione di punti o unità è selezionato tramite uno schema di campionamento probabilistico. In ogni punto o per ogni unità spaziale, il valore è stimato usando un interpolatore spaziale e sono discusse le condizioni che ne assicurano la correttezza asintotica e la coerenza in un approccio basato sul disegno.*

Key words: Design consistency, Sampling, Spatial interpolation.

1 Introduction

The spatial pattern of natural resources is considerable in many environmental and ecological surveys and mapping is essential for a visual overview of the spatial

Lorenzo Fattorini
Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy e-mail: lorenzo.fattorini@unisi.it,

Marzia Marcheselli
Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy e-mail: marzia.marcheselli@unisi.it

Caterina Pisani
Department of Economics and Statistics, University of Siena, Piazza San Francesco 8, 53100 Siena, Italy e-mail: caterina.pisani@unisi.it

pattern of the variable of interest on the study region. For example, soil composition and mineral concentration are of interest in geology, soil and water pollution are crucial in ecology, species abundance and coverage are important in forestry.

Depending on the variable and the goals of the survey, the study region may be considered either a continuum of points or a finite population of spatial units, such as when the region is partitioned into a network of regular polygons or into a collection of irregular patches (e.g. Opsomer et al. 2007). Model-based estimation methods are adopted in most cases: uncertainty arises from the super-population probability model, which has been supposed to generate the surface or the population values, conditional on the sample of spatial units. By contrast, if any assumption about the super-population model generating the surface or the population values is avoided, uncertainty stems only from the sampling scheme and the surface or the population values are considered fixed.

Recently, in a design-based framework, Fattorini et al. (2018a) propose the use of the model-assisted inverse distance weighting interpolator (Bruno et al., 2013) for estimating a variable of interest when a finite population of spatial units is considered. Moreover Fattorini et al. (2018b) investigate the use of the inverse distance weighting interpolator to estimate the surface values in the continuous population setting. To render statistically sound a design-based map, conditions ensuring some sort of design-based asymptotic unbiasedness and consistency are obtained for both continuous and finite populations and asymptotically conservative estimators of the mean squared error are proposed (Fattorini et al., 2018a, 2018b).

2 Notation and setting

Consider a study region $\mathscr{A}$ of area A. Let $\mathscr{A}$ be a connected and compact set of $\mathbf{R}^2$. For $\mathbf{p},\mathbf{q} \in \mathscr{A}$, let $\|\mathbf{p}-\mathbf{q}\|$ be their distance, where $\|\cdot\|$ denotes a norm in $\mathbf{R}^2$.

If the value of the survey variable at $\mathbf{x}$, say $y(\mathbf{x})$, is defined for each point $\mathbf{x} \in \mathscr{A}$, the population is a continuum of points and the aim is the estimation of $y(\mathbf{x})$, at least ideally, for each $\mathbf{x} \in \mathscr{A}$ from a sample $(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ of n points selected from $\mathscr{A}$ by means of a suitable sampling scheme (Cordy, 1993).

On the other hand, if $\mathscr{A}$ is partitioned into N spatial units $\mathbf{a}_1,\ldots,\mathbf{a}_N$ of area $a_1,...,a_N$, we are dealing with a finite population. In this case let U denote both the set of spatial units and the set of indexes $\{1,\ldots,N\}$. Moreover let y_j be the amount of the survey variable within $\mathbf{a}_j$ in such a way that $f_j = y_j/a_j$ is the density and the goal is estimating the value y_j for each $j \in U$ using a probabilistic sample S of n units selected from U. Because in most situations the a_js are known for all the population units, it is equivalent to interpolate the y_js or the f_js but density estimation is more suitable for working in an asymptotic scenario in which the a_js decrease and then the y_js may approach zero.

In both frameworks a model-assisted estimation may be performed by exploiting Tobler's law (Tobler, 1970) and adopting the inverse distance weighting interpolator (Bruno et al., 2013).

In the continuous population setting, following Fattorini et al. (2018b) the inverse distance weighting interpolator turns out to be

$$\hat{y}(\mathbf{x}) = I_{\{\cup_{i=1}^{n}(\mathbf{X}_i=\mathbf{x})\}} y(\mathbf{x}) + [1 - I_{\{\cup_{i=1}^{n}(\mathbf{X}_i=\mathbf{x})\}}] \frac{\sum_{i=1}^{n} y(\mathbf{X}_i)\phi(||\mathbf{x}-\mathbf{X}_i||)}{\sum_{i=1}^{n}\phi(||\mathbf{x}-\mathbf{X}_i||)}, \qquad (1)$$

where I_E is the indicator function of the event E and $\phi : [0,\infty) \to \mathbf{R}^+$ is a non-increasing continuous function on $(0,\infty)$, with $\phi(0) = 0$, $\lim\limits_{d \to 0^+} \phi(d) = \infty$.

In the finite population setting, quoting Fattorini et al. (2018a) the inverse distance weighting interpolator is

$$\hat{f}_j = I_{\{j \in S\}} f_j + [1 - I_{\{j \in S\}}] \frac{\sum_{i \in S} f_i \phi(||\mathbf{c}_j - \mathbf{c}_i||)}{\sum_{l \in S}\phi(||\mathbf{c}_j - \mathbf{c}_l||)}, \qquad (2)$$

where $\mathbf{c}_j$ is the centroid of the spatial unit $\mathbf{a}_j$, $j = 1, \ldots, N$.

3 Asymptotic results

The design-based expectation and variance of (1) and (2) cannot be expressed in closed form, giving no insights about the bias and precision. Therefore, conditions providing asymptotic design-based unbiasedness and consistency are needed.

3.1 Continuous population

Suppose a sequence of fixed-size designs to select a sample of n_k points $\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_{n_k}^{(k)}$, with $n_k \to \infty$ as k increases. Moreover, from (1), let $\hat{y}_k(\mathbf{x})$ be the inverse distance weighting interpolator for the k-th design. The asymptotic design-based properties of $\hat{y}_k(\mathbf{x})$ are derived as the sample size increases but the surface remains fixed.

To achieve design consistency suppose that y is a bounded measurable function on $\mathscr{A}$ and $\lim_{d \to 0^+} d^2\phi(d) = \infty$. Fattorini et al. (2018b) prove that, under mathematical conditions ensuring an asymptotic spatial balance of the sampling scheme as the sample size increases, if y is continuous at $\mathbf{x}$, $\hat{y}_k(\mathbf{x})$ is point wise design consistent at $\mathbf{x}$ while if y is continuous on $\mathscr{A}$ it is uniformly design consistent. Particularly, under uniform random sampling point wise design consistency holds while under systematic grid sampling and tessellation stratified sampling, if the study area is partitioned into a sequence of polygonal grids, $\hat{y}_k(\mathbf{x})$ is uniformly design consistent.

Moreover if y is a Lipschitz function in a neighbourhood of $\mathbf{x}$ and $\phi(d) = d^{-\alpha}$, with $\alpha > 2$, under systematic grid sampling and tessellation stratified sampling, design consistency is ensured with a $O(n_k^{(2-\alpha)/(\alpha+1)})$ convergence rate, and the use of very large α values in the distance function ϕ seems to be advisable. However, a trade-off choice of α between 3 and 5 seems suitable for moderate sample sizes, ensuring a convergence rate of at least $O(n_k^{-1/4})$. In many real situations, the Lipschitz assumption within patches partitioning the study area is quite reasonable. Indeed, often there are parts of the study region in which the surface changes smoothly throughout space, well approaching the Lipschitz condition, while sudden variations only occur on borders, which may be realistically approximated by curves of measure 0. Therefore the surface shares the Lipschitz condition almost everywhere.

3.2 Finite population

Consider a sequence $\{U_k\}$ of partitions of $\mathscr{A}$. Each partition U_k is constituted of N_k spatial units $\mathbf{a}_1^{(k)},\ldots,\mathbf{a}_{N_k}^{(k)}$ of area $a_1^{(k)},...,a_{N_k}^{(k)}$ and centroids $\mathbf{c}_1^{(k)},\ldots,\mathbf{c}_{N_k}^{(k)}$. Denote by $y_j^{(k)}$ the amount of the survey variable within $\mathbf{a}_j^{(k)}$ and by $f_j^{(k)}$ its density. Suppose that $N_k \to \infty$, i.e. $\mathscr{A}$ is partitioned into an increasing number of spatial units: it is natural to suppose also that $\sup_{j\in U_k} \mathrm{diam}(\mathbf{a}_j^{(k)}) \to 0$. Finally suppose a sequence of designs to select a sample of size n_k from U_k. Since U_k is a sequence of finite populations of increasing sizes, each constituted by different spatial units, the sequence differs from those customarily adopted to achieve asymptotic results in finite population inference, which are constituted of nested populations with increasing totals.

Let $\hat{f}_j^{(k)}$ be the inverse distance weighting interpolator for the k-th design: the goal is to determine its asymptotic design-based behavior as the partition of the study area becomes thinner and thinner. It should be noticed that the unit j of the kth population is lost in the subsequent populations. This problem can be handled by introducing two sequences of functions from $\mathscr{A}$ onto R^+, say

$$f_k(\mathbf{x}) = \sum_{j\in U_k} f_j^{(k)} I_{\{\mathbf{x}\in\mathbf{a}_j^{(k)}\}}, \qquad \mathbf{x}\in\mathscr{A}$$

$$\hat{f}_k(\mathbf{x}) = \sum_{j\in U_k} \hat{f}_j^{(k)} I_{\{\mathbf{x}\in\mathbf{a}_j^{(k)}\}} \qquad \mathbf{x}\in\mathscr{A}.$$

Practically speaking, the N_k densities of the kth population are substituted by a piecewise constant function in $\mathscr{A}$, which equals $f_j^{(k)}$ onto $\mathbf{a}_j^{(k)}$.

Suppose that there exists a Riemann integrable function f from $\mathscr{A}$ onto $[0,L]$, which gives the density of the survey variable for any $\mathbf{x}\in\mathscr{A}$ and that the spatial units not only have diameters approaching to 0, but there is no excessively elongated unit. The assumption on the distance function already adopted in the continuous setting is also needed.

Fattorini et al. (2018a) prove that, under mathematical conditions ensuring an asymptotic spatial balance of the sampling scheme, if f is continuous at $\mathbf{x}$, $\hat{f}_k$ is point wise design consistent at $\mathbf{x}$ while, if f is continuous on $\mathscr{A}$, it turns out to be uniformly design consistent. Under simple random sampling without replacement and stratified sampling with proportional allocation point wise design consistency holds while one per stratum stratified sampling and systematic sampling ensures uniform design consistency.

4 Mean squared error estimation

Any estimator of the mean squared error of the interpolators should not be computationally demanding. Therefore time-consuming resampling procedures, such as the bootstrap or jackknife, should be avoided.

Owing to Tobler's law, the value of the sampled point nearest to $\mathbf{x}$ is likely to be a good proxy for $y(\mathbf{x})$ and the density of the sampled unit nearest to unit j is likely to be a good proxy for f_j. Then, in the continuous setting a simple and asymptotically conservative estimator for the mean squared error of $\hat{y}(\mathbf{x})$ is

$$\hat{V}(\mathbf{x}) = \left\{\hat{y}(\mathbf{x}) - y(\mathbf{X}_{\mathrm{near}(\mathbf{x})})\right\}^2, \tag{3}$$

where near$(\mathbf{x})$ is the index of the sample location that is nearest to $\mathbf{x}$. The finite population counterpart of (3)

$$\hat{V}_j = (\hat{f}_j - \hat{f}_{\mathrm{near}(j)})^2,$$

where near(j) is the label of the sampled unit that is nearest to unit j, can be adopted to estimate the mean squared error of $\hat{f}_j$.

5 Simulation study

As to the continuous population setting, in Fattorini et al. (2018b) a simulation is performed on three artificial surfaces on the unit square, referred to as surf1, surf2 and surf3, and, respectively, defined at any point $\mathbf{x} = (x_1, x_2)$ as

$$y(\mathbf{x}) = C_1(\sin^2 x_1 + \cos^2 x_2 + x_1), \quad y(\mathbf{x}) = C_2(\sin 3x_1 \sin^2 3x_2)^2,$$

$$y(\mathbf{x}) = \begin{cases} C_3 x_1 x_2, & \min(x_1, x_2) < 0.5, \\ C_3(1 + x_1 x_2), & \text{otherwise} \end{cases}$$

where the constants C_1, C_2, and C_3 ensure a maximum value of 10. The first two surfaces are continuous while surf3 shows a discontinuity at the edge of the upper-right quadrant of the unit square.

For each surface, $R = 10{,}000$ samples of size $n = 25, 100, 225, 400$ are independently selected by uniform random sampling, tessellation stratified sampling and systematic grid sampling: the last two schemes are performed by partitioning the unit square into quadrats of equal size and then selecting one point per quadrat. Estimation is performed for each of the $N = 10,000$ centroids of the equally-spaced 100×100 grid by means of (1) using $\phi(d) = d^{-\alpha}$ with $\alpha = 2, 2.5, 3, 4$.

The simulation results confirm the theoretical findings: for both the continuous surfaces, a sharp decrease in the minima, maxima and averages of the absolute bias and mean squared error occurs for $\alpha > 2$ as the sample size increases while the decreases are less marked for $\alpha = 2$. For surf3, uniform design consistency is precluded and pointwise design consistency is ensured for $\alpha > 2$ for only the set of continuity points. As the sample size increases, sharp decreases occur for only the minima and averages of the absolute bias and mean squared error for $\alpha > 2$ while the maxima remain approximately constant. The decreases are less marked for $\alpha = 2$.

In the finite population setting, Fattorini et al. (2018a) adopt surf2 and surf3 to define two densities on the unit square. For any density, five spatial populations of sizes N=100, 400, 1600, 6400, and 25,600 are constructed by partitioning the unit square into grids of 10×10, 20×20, 40×40, 80×80, and 160×160 quadrats, respectively, and then taking the integrals of the density onto the quadrats as population values. For any population, $R = 10,000$ samples of size $n = N/10$ are independently selected by means of simple random sampling without replacement, one-per-stratum stratified sampling and systematic sampling. One-per-stratum stratified sampling and systematic sampling are performed by partitioning the grids into blocks of 2×5 contiguous quadrats and selecting one quadrat per block. Once the samples are selected, (2) is adopted to estimate densities for all the quadrats in the populations by using the same distance function adopted in the continuous framework. Simulation results support the theoretical results and are analogous to those obtained for continuous population.The continuity of surf2 for $\alpha > 2$ ensures point-wise design consistency under simple random sampling without replacement and uniform design consistency under sytematic and one-per-stratum stratified sampling. Uniform design consistency is precluded for surf3 and point-wise design consistency is ensured only away from the discontinuity lines for $\alpha > 2$ and for all the sampling schemes.

References

1. Bruno F., Cocchi, D., Vagheggini, A.: Finite population properties of individual predictors based on spatial pattern. Environ. Ecol. Stat. **20** 467–494 (2013)
2. Cordy, C.B.: An extension of the Horvitz–Thompson theorem to point sampling from a continuous universe. Stat. Probabil. Lett. **18**, 353–362 (1993).
3. Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based maps for finite populations of spatial units. J. Am. Stat. Assoc. (2018a) DOI:10.1080/ 01621459.2016.1278174
4. Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based maps for continuous spatial populations. Biometrika (2018b) DOI:10.1093/biomet/asy012
5. Opsomer, J. D., Breidt, F. G., Moisen, G. G., Kauermann, G.: Model-assisted estimation of forest resources with generalized additive models. J. Am. Stat. Assoc. **102**, 400–416 (2007)
6. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Econ. Geogr. **46**, 234–240 (1970)

Testing for independence in analytic inference

Test di indipendenza nell'inferenza analitica

Pier Luigi Conti and Alberto Di Iorio

Abstract In analytic inference, data usually come from complex sampling designs, possibly with different inclusion probabilities, stratification, clustering of units. The effect of a complex sampling design is that sampling data are not *i.i.d.*, even if they are at a superpopulation level. This dramatically changes the probability distribution of usual test-statistics, such as Spearman's Rho. An approach based on a special form of resampling is proposed, and its properties are studied.
Abstract *Nell'inferenza analitica i dati generalmente provengono da disegni campionari complessi, che includono differenti probabilità di inclusione, stratificazione, grappoli di unità. L'effetto di un disegno di questo tipo è che i dati a livello campionario non sono* i.i.d.*, anche se lo sono a livello di superpopolazione. Di conseguenza, viene completamente modificata la distribuzione di probabilità delle statistiche-test comunemente utilizzate (come il Rho di Spearman). Nel presente lavoro viene studiato un approccio basato su una nuova forma di ricampionamento, di cui si studiano le proprietà.*

Key words: Independence tests, sampling design, asymptotics, empirical process, resampling.

1 Introduction

The use of superpopulation models in survey sampling has a long history, going back (at least) to [2], where the limits of assuming the population characteristics as *fixed*, especially in economic and social studies, are stressed. As clearly appears (cfr.

Pier Luigi Conti
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy, e-mail: pierluigi.conti@uniroma1.it

Alberto Di Iorio
Banca d'Italia; Via Nazionale, 91; 00184 Roma; Italy, e-mail: alberto.diiorio@uniroma1.it

[7], [5]), there are basically two forms of inference in a finite populations setting. The first one is *descriptive* or *enumerative* inference, namely inference about finite population parameters. This kind of inference is a static "picture" on the current state of a population, and does not take into account the mechanism generating the characters of interest of the population itself. The second one is *analytic* inference, and consists in inference on superpopulation parameters. This kind of inference is about the process that generates the finite population. In contrast with *enumerative* inference results, *analytic* ones are more general, and still valid for every finite population generated by the same superpopulation model.

In the present paper attention is focused on a special problem of analytic inference, namely testing for independence between two characters. The main consequence of using a sampling design with possibly different inclusion probabilities is that commonly used test-statistics based on ranks, such as Spearman Rho rank correlation or Gini G cograduation statistics are not distribution free under independence, and in general do not have the same distribution (neither finite sample, nor asymptotic) as in the case of *i.i.d.* sample data. This calls for the need of developing new test-statistics, suitable for data collected through a complex design. Since their distribution, both for a finite sample size and asymptotically, does have a complicate form depending on the superpopulation, the need of approximating their distribution arises. Unfortunately, the widespread Efron's bootstrap does not work in the present case, again because of the use of a complex sample design. In the sequel, a new resampling scheme will be proposed, and its properties studied. In particular, it will be shown that it is asymptotically correct.

Let $\mathscr{U}_N$ be a finite population of size N, and let X, Y be two characters of interest, defined on the population $\mathscr{U}_N$. Let further x_i, y_i be the values of characters X, Y for unit i $(= 1, \dots, N)$.

A sample s of size n is a subset of $\mathscr{U}_N$. The selection of s is performed according to a probabilistic sampling design. Formally speaking, for each unit i in $\mathscr{U}_N$, define a Bernoulli random variable (r.v.) D_i, such that the unit i is included in the sample if and only if (iff) $D_i = 1$, and let $\mathbf{D}_N = (D_1, \dots, D_N)$. A (unordered, without replacement) sampling design P is the probability distribution of $\mathbf{D}_N$. In particular, $\pi_i = E_P[D_i]$ $(\pi_{ij} = E_P[D_i D_j])$ is the first (second) order inclusion probability of unit i (pair of units i, j). The suffix P denotes the sampling design used to select population units.

The first order inclusion probabilities are frequently taken proportional to an appropriate function of the values of the *design variables*. The design variables may include strata indicator variables, as well as qualitative variables measuring cluster and unit characteristics (cfr. [5]); in what follows they are denoted by $\mathscr{T}_1, \dots, \mathscr{T}_L$, whilst $t_{i1}, \dots, t_{iL}$ are their values for unit i. As already said, we will assume that $\pi_i \propto z_i$, where $z_i = h(t_{i1}, \dots, t_{iL})$ is the value of $Z = h(\mathscr{T}_1, \dots, \mathscr{T}_L)$ for unit i.

Take now N real numbers $0 < p_i < 1$, $i = 1, \dots, N$, with $p_1 + \cdots + p_N = n$. The sampling design is a *Poisson design* with parameters $p_1, \dots, p_N$ if the r.v.s D_is are independent with $\pi_i = p_i$ for each unit i. The *rejective sampling*, or *normalized conditional Poisson sampling* ([4], [8]) corresponds to the probability distribution of the random vector $\mathbf{D}_N$, under Poisson design, conditionally on $n_s = n$.

The *Hellinger distance* between a sampling design P and the rejective design is

$$d_H(P, P_R) = \sum_{D_1,\dots,D_N} \left(\sqrt{Pr_P(D_N)} - \sqrt{Pr_R(D_N)} \right)^2 . \tag{1}$$

Our basic assumptions are listed below.

A1. $(\mathscr{U}_N;\ N \geq 1)$ is a sequence of finite populations of increasing size N.

A2. For each N, $(y_i, x_i, t_{i1}, \dots, t_{iL})$, $i = 1, \dots, N$ are realizations of a superpopulation $\{(Y_i, X_i, T_{i1}, \dots, T_{iL}),\ i = 1, \dots, N\}$ composed by *i.i.d.* $(L+2)$-dimensional r.v.s. In the sequel, the symbol $\mathbb{P}$ will denote the (superpopulation) probability distribution of r.v.s $(Y_i, X_i, T_{i1}, \dots, T_{iL})$s, and $\mathbb{E}$, $\mathbb{V}$ are the corresponding operators of mean and variance, respectively. Furthermore, if $Z_i = h(T_{i1}, \dots, T_{iL})$, the joint superpopulation d.f. of (Y_i, X_i, Z_i) will be denoted by

$$K(y, x, z) = \mathbb{P}(Y_i \leq y, X_i \leq x, Z_i \leq z), \tag{2}$$

and

$$H((y, x|z) = \mathbb{P}(Y_i \leq y, X_i \leq x|Z_i = z), \tag{3}$$

$$F(y|z) = \mathbb{P}(Y_i \leq y|Z_i = z),\ \ G(x) = \mathbb{P}(X_i \leq x|Z_i = z) \tag{4}$$

are the joint and marginal superpopulation d.f.s of Y_i and X_i (given Z).

A3. For each population $\mathscr{U}_N$, sample units are selected according to a fixed size sample design with positive first order inclusion probabilities $\pi_i \propto z_i$, with sample size $n = \pi_1 + \cdots + \pi_N$, and $z_i = h(t_{i1}, \dots, t_{iL})$, $i = 1, \dots, N$. It is assumed that

$$\lim_{N,n\to\infty} \mathbb{E}[\pi_i(1-\pi_i)] = d > 0. \tag{5}$$

Furthermore, the notation $x_N = (x_1, \dots, x_N)$ is used.

A4. The sample size n increases as the population size N does, with

$$\lim_{N\to\infty} \frac{n}{N} = f,\ \ 0 < f < 1.$$

A5. For each population $(\mathscr{U}_N;\ N \geq 1)$, let P_R be the rejective sampling design with inclusion probabilities $\pi_1, \dots, \pi_N$, and let P be the actual sampling design (with the same inclusion probabilities). Then

$$d_H(P, P_R) \to 0 \ \text{ as } N \to \infty,\ \ a.s. - \mathbb{P}.$$

A6. $\mathbb{E}[X_1^2] < \infty$, so that the quantity in (5) is equal to:

$$d = f\left(1 - \frac{\mathbb{E}[X_1^2]}{\mathbb{E}[X_1]^2}\right) + f(1-f)\frac{\mathbb{E}[X_1^2]}{\mathbb{E}[X_1]^2} > 0. \tag{6}$$

2 The problem

As already said, our goal is to construct an independence test for the two characters X, Y (conditionally on the design variables $T_j s$). For the sake of simplicity we will consider a single, discrete design variable T, taking values $T^1, \dots, T^k$. Hence, the hypothesis problem takes the form

$$H_0 : H(x,y|T) = F(x|T)G(y|T)$$
$$H_1 : H(x,y|T) \neq F(x|T)G(y|T)$$

A simple approach could consist in using a rank based test-statistic, such as the Spearman's Rho statistic or the Gini's cograduation statistic. Unfortunately, due to the use of the sample design, such statistics are not distribution free under H_0, neither exactly nor asymptotically. Hence, their use is inappropriate under general sampling designs.

In order to construct a test-statistic for the above problem, the general measure of monotone dependence proposed in [1] is extended to the present case. Given two continuous variables X,Y, let F and G be their marginal distributions, respectively, and let H be the joint distribution of the bivariate variable (X,Y). A general measure of the monotone dependence γ_g between X and Y, is a real-valued functional γ_g of the bivariate distribution $H(x,y)$ defined as follows

$$\gamma_g = \int_{\mathbb{R}^2} g(|F(x|T)+G(y|T)-1|) - g(|F(x|T)-G(y|T)|)\, dH(x,y|T), \quad (7)$$

where $g : [0,1] \to \mathbb{R}$ is a strictly increasing, continuous and convex function, with $g(0) = 0$ snd continuous first derivative. Under the null hypothesis of independence the latter quantity is equal to zero. If $g(s) = s^2$, then (7) reduces to the (un-normalized) Spearman's coefficient. If $g(s) = s$, then (7) reduces to the (un-normalized) Gini cograduation coefficient. In general, $\gamma_g = 0$ whenever X and Y are independent.

The basic idea is to estimate first H, F, G by their Hájek estimators

$$\widehat{H}(y, xvertt) = \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(x_i \leq x)} I_{(y_i \leq y)} I_{(T_i = t)} \Bigg/ \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(T_i=t)} \quad (8)$$

$$\widehat{F}(y) = \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(y_i \leq y)} I_{(T_i = t)} \Bigg/ \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(T_i=t)}, \quad (9)$$

$$\widehat{G}(x) = \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(x_i \leq x)} I_{(T_i = t)} \Bigg/ \sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(T_i=t)} \quad (10)$$

and then in estimating the quantity γ_g with a plug-in approach, by replacing the distribution function is the distributions functions in (7) with their Hájek estimators. In this way, the test-statistic

$$\widehat{\gamma}_g = \frac{\sum_{i=1}^{N} \frac{1}{\pi_i} \left(g(|\widehat{F}(x_i|T_i) + \widehat{G}(y_i|T_i) - 1|) - g(|\widehat{F}(x_i|T_i) - \widehat{G}(y_i|T_i)|) \right) D_i}{\sum_{i=1}^{N} \frac{1}{\pi_i} D_i}. \quad (11)$$

is obtained.

Proposition 1. *Suppose that the conditions* $A1$*-*$A6$ *are met, and assume that the null hypothesis* H_0 *holds true. Then, the r.v.*

$$\sqrt{n}\widehat{\gamma}_g \quad (12)$$

tends in distribution to a normal r.v. with zero mean and variance σ_0^2*, as N, n increase.*

The asymptotic variance σ_0^2 of (12) does have a complicate form, and cannot be estimated on the basis of sample data. The basic idea is to approximate the distribution of (11) under H_0 by resorting to resampling.

Define S_j as the subset of sample units having value T^j of the design variable, and let n_j be the size of S_j, $j = 1, \dots, k$.

0. Repeat M times steps 1-4 below.
1. Generate a pseudo-population of size N by selecting unit i in the sample with probability $\pi_i^{-1} / \sum \pi_i^{-1} D_i$. To each unit i^* of the pseudo-population a value $T_{i^*}^*$ is attached, such that $T_{i^*}^* = T_i$ whenever $i^* = i$.
2. If $T_{i^*}^* = T^j$, then sample independently from S_j, and with probability $\pi_i^{-1} / \sum_{k \in S_j} \pi_k^{-1}$, a X-value $X_{i^*}^*$ and a Y-value $Y_{i^*}^*$. As a result, a pseudo-population $\mathscr{U}_N^*$ is obtained, such that for each unit $i^* \in \mathscr{U}_N^*$ a triplet $(Y_{i^*}^*, X_{i^*}^*, T_{i^*}^*)$ is defined. Furthermore, $Y_{i^*}^*$ and $X_{i^*}^*$ are independent conditionally on $T_{i^*}^*$.
3. Draw a pseudo-sample of size n from the population $\mathscr{U}_N^*$, using a high entropy sampling design P^* with first order inclusion probabilities π_{i^*} according to $A3$.
4 Compute the value $\widehat{\gamma}_g^*$ of the statistic (11) for the pseudo-sample drawn at step 3.

In this way, the M replicates

$$\sqrt{n}\widehat{\gamma}_{g,m}^*, \;\; m = 1, \dots, M \quad (13)$$

are obtained.

Consider next the empirical distribution function (edf) constructed on the basis of the M replicates (13):

$$\widehat{R}(u) = \frac{1}{M} \sum_{m=1}^{M} I_{(\widehat{\gamma}_{g,m}^* \leq u)} \quad (14)$$

and let $\widehat{R}^{-1}$ be the corresponding quantile function

$$\widehat{R}^{-1}(p) = \inf\{u : \widehat{R}(u) \geq u\}. \quad (15)$$

Let further Φ_{0,σ_0} be a normal distribution function with zero expectation and variance σ_0^2. Proposition 1 establishes that, as both N, n increase, under the null hypothesis H_0 the probability distribution function $Pr_{H_0}(\sqrt{n}\widehat{\gamma}_g \leq u)$ tends to $\Phi_{0,\sigma_0}(u)$, uniformly w.r.t. u. Next proposition establishes that, as the number M of replicates increases, the edf (14) tends to the *same* limiting law. This proves the (asymptotic) validity of the proposed resampling technique.

Proposition 2. *Suppose that the conditions* A1*-*A6 *are met, and assume that the null hypothesis* H_0 *holds true. Then, as N, n, M increase, the following results hold (with probability 1).*

$$\sup_u \left| \widehat{R}(u) - Pr_{H_0}(\sqrt{n}\widehat{\gamma}_g \leq u) \right| \to 0;$$
$$\sup_u \left| \widehat{R}(u) - \Phi_{0,\sigma_0}(u)) \right| \to 0;$$
$$\widehat{R}^{-1}(p) \to \sigma_0 z_{1-p}$$

where z_{1-p} *satisfies the relationship* $\Phi_{0,1}(z_{1-p}) = p$.

As a consequence of Proposition 2, the region

$$\widehat{R}^{-1}(\alpha/2) \leq \sqrt{n}\widehat{\gamma}_g \leq \widehat{R}^{-1}(\alpha/2) \tag{16}$$

is an acceptance region of approximative size α for testing independence.

A comparison between the proposed test and the one based on the "usual" Spearman statistic has been performed *via* simulation. The proposed test performs better than the "usual" one in terms of power, as well as in terms of closeness of the actual size to the nominal significance level.

References

1. Cifarelli D M., Conti P L., Regazzini E.: On the asymptotic distribution of a general measure of monotone dependence. The Annals of Statistics, **24**, 1386–1399 (1996)
2. Cochran W G.: The use of the analysis of variance in enumeration by sampling. Journal of the American Statistical Association, **34**, 492–510 (1939)
3. Efron B.: Bootstrap methods: another look at the jackknife. The Annals of Statistics, **7**, 1–26 (1979)
4. Hájek J.: Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population. The Annals of Mathematical Statistics, **35**, 1491–1523 (1964)
5. Pfeffermann D.: The role of sampling weights when modeling survey data. International Statistical Review, **61**, 317–337 (1993)
6. Pfeffermann D., Sverchkov M.: Prediction of finite population totals based on the sample distribution. Survey Methodology, **30**, 79-92 (2004)
7. Särdnal C -E., Swensson B., Wretman, J H.: Model Assisted Survey Sampling. Springer-Verlag, New York (1992)
8. Tillé, Y.: Sampling Algorithms. Springer Verlag, New York (2006)

On the aberrations of two-level Orthogonal Arrays with removed runs

Sulle aberrazioni di Orthogonal Arrays binari con punti rimossi

Roberto Fontana and Fabio Rapallo

Abstract We consider binary Orthogonal Arrays and we analyze the aberrations of the fractions obtained by the deletion of $p = 1, 2$ or 3 design points. Some explicit formulae are given for $p = 1$ and some examples are presented in the other cases.
Abstract *A partire da Orthogonal Arrays a due livelli, studiamo le aberrazioni delle frazioni ottenute tramite rimozione di* $p = 1, 2$ *o* 3 *punti sperimentali. Deriviamo alcune formule esplicite nel caso* $p = 1$*, e per gli altri casi presentiamo alcuni esempi.*

Key words: Fractional factorial designs, Word-Length Pattern

1 Introduction

The theory of Orthogonal Arrays (OAs) has a long history and represents a major research topic for both methodology and applied Statistics. The need for efficient experimental designs has led to the definition of several criteria for the choice of the design points. All such criteria aim to produce the best estimates of the relevant parameters for a given sample size. As general references for OAs, the reader can refer to [5].

In particular, we consider only binary designs under the Minimum Aberration (MA) criterion, but we focus on the following problem. In several situations, it is hard to define *a priori* a fixed sample size. For example, budget constraints or time limitations may occur after the definition of the design, or even when the experiments are running, thus leading to an incomplete design. In such a situation, it is relevant not only to choose an OA with good properties, but also to define an order

Roberto Fontana
DISMA, Politecnico di Torino, e-mail: roberto.fontana@polito.it

Rapallo Fabio
DISIT, Università del Piemonte Orientale e-mail: fabio.rapallo@uniupo.it

of the design points, so that the experimenter can stop the sequence of runs and loose as little information as possible. While OAs with added runs are well studied, see for instance [2], less has been done in the case of OAs with removed runs. The order of the runs of an OA and the OAs with removed runs are both interesting topics, but the first one is mainly studied with the aim of minimizing the changes of the factor levels, see e.g. [8], while the second one is only considered in the framework of D-optimality, see [1]. In [7] and [8] several papers, describing practical problems where OAs with missing runs play a major role, are listed.

In this work, we consider the MA criterion for binary OAs and we study the behavior of the aberrations when p points are removed from an OA, for small values of p. The MA criterion is based on the Word-Length Pattern (WLP) introduced in [4]. In practice, the WLP is used to discriminate among different designs $\mathscr{F}_1, \ldots, \mathscr{F}_d$ by looking at the lexicographic minimum of the vector

$$A_{\mathscr{F}_i} = (A_0(\mathscr{F}_i) = 1, A_1(\mathscr{F}_i), \ldots, A_m(\mathscr{F}_i)) , \quad i = 1, \ldots, d.$$

2 Orthogonal Arrays and aberrations

Let us consider an experiment with m 2-level factors. The full factorial design is $\mathscr{D} = \{-1, 1\}^m$. We briefly recall here the basic definitions concerning Orthogonal Arrays and aberrations. For details refer to [3].

Definition 1. A fraction $\mathscr{F}$ is a multiset $(\mathscr{F}_*, f_*)$ whose underlying set of elements $\mathscr{F}_*$ is contained in $\mathscr{D}$ and f_* is the multiplicity function $f_* : \mathscr{F}_* \to \mathbb{N}$ that for each element in $\mathscr{F}_*$ gives the number of times it belongs to the multiset $\mathscr{F}$.

We recall that the underlying set of elements $\mathscr{F}_*$ is the subset of $\mathscr{D}$ that contains all the elements of $\mathscr{D}$ that appear in $\mathscr{F}$ at least once. We denote the number of elements of the fraction $\mathscr{F}$ by $\#\mathscr{F}$, with $\#\mathscr{F} = \sum_{x \in \mathscr{F}_*} f_*(x)$.

To describe the counting function of a fraction, we follow the theory in [3]. The simple terms of the form X_j, i.e., the j-th component function which maps a point $x = (x_1, \ldots, x_m)$ of $\mathscr{D}$ to its j-th component,

$$X_j \colon \mathscr{D} \ni (x_1, \ldots, x_m) \longmapsto x_j \in \{-1, 1\}$$

and the interactions $X^\alpha = X_1^{\alpha_1} \cdot \ldots \cdot X_m^{\alpha_m}$, $\alpha \in L = \{0, 1\}$ i.e., the monomial functions of the form

$$X^\alpha : \mathscr{D} \ni (x_1, \ldots, x_m) \mapsto x_1^{\alpha_1} \cdot \ldots \cdot x_m^{\alpha_m}$$

are a basis of all the real functions defined over $\mathscr{D}$. We use this basis to represent the counting function of a fraction according to the following definition.

Definition 2. The counting function R of a fraction $\mathscr{F}$ is a polynomial defined over $\mathscr{D}$ so that for each $x \in \mathscr{D}$, $R(x)$ equals the number of appearances of x in the fraction. A $0-1$ valued counting function is called an indicator function of a single-replicate fraction $\mathscr{F}$. We denote by c_α the coefficients of the representation of R on $\mathscr{D}$ using the monomial basis $\{X^\alpha,\ \alpha \in L\}$:

$$R(x) = \sum_{\alpha \in L} c_\alpha X^\alpha(x),\ x \in \{-1,1\}^m,\ c_\alpha \in \mathbb{R}.$$

Among the fractions of a full factorial design 2^m, we consider Orthogonal Arrays.

Definition 3. A fraction $\mathscr{F}$ factorially projects onto the I-factors, $I = \{i_1,\ldots,i_k\} \subset \{1,\ldots,m\}$, $i_1 < \ldots < i_k$, if the projection $\pi_I(\mathscr{F})$ is a full factorial design or a multiple of a full factorial design, i.e., the multiset $(\{-1,1\}^m, f_*)$ where the multiplicity function f_* is constant over $\{-1,1\}^m$.

Definition 4. A fraction $\mathscr{F}$ is an Orthogonal Array (OA) of strength t if it factorially projects onto any I-factors with $\#I = t$.

The connections between the OAs and the counting function are given in the proposition below.

Proposition 1. *A fraction is an OA of strength t if and only if all the coefficients $c_\alpha,\ \alpha \neq 0 \equiv (0,\ldots,0)$ of the counting function up to the order t are* 0.

Definition 5. The Word-Length Pattern (WLP) of a fraction $\mathscr{F}$ of the full factorial design $\mathscr{D}$ is the vector $A_\mathscr{F} = (A_0(\mathscr{F}), A_1(\mathscr{F}), \ldots, A_m(\mathscr{F}))$, where

$$A_j(\mathscr{F}) = \sum_{|\alpha|_0 = j} a_\alpha \quad j = 0,\ldots,m,$$

$$a_\alpha = \left(\frac{c_\alpha}{c_0}\right)^2,$$

$|\alpha|_0$ is the number of non-null elements of α, and $c_0 := c_{(0,\ldots,0)} = \#\mathscr{F}/\#\mathscr{D}$.

In the definition above, the number a_α is the aberration of the term X^α.

3 The effect on the WLP of the removal of one, two or three points

In this section we study the effect on the WLP of the removal of one, two or three points from an OA of strength t. With respect to the removal of one point outlined in Sect. 3.1 we obtain an analytical expression for the first $t+1$ terms of the WLP.

3.1 One removed point

Let us consider an orthogonal array $\mathscr{F}$ with n runs, m 2-level factors and strength t. The WLP of $\mathscr{F}$ is

$$A_{\mathscr{F}} = (A_0(\mathscr{F}) = 1, A_1(\mathscr{F}) = 0, \ldots, A_t(\mathscr{F}) = 0, A_{t+1}(\mathscr{F}), \ldots, A_m(\mathscr{F}))\,.$$

Let us consider a point $e = (e_1, \ldots, e_m) \in \mathscr{F}$ and the fraction $\mathscr{F}_e$ that contains all the points of $\mathscr{F}$ apart from e, $\mathscr{F}_e = \mathscr{F} \setminus \{e\}$. Let us denote by $R = \sum_\alpha c_\alpha X^\alpha$ the counting function of $\mathscr{F}$ and by $R_{\{e\}} = \sum_\alpha c_\alpha^{(e)} X^\alpha$ the counting function of the fraction made by the single point e. We can write

$$R_{\{e\}}(x_1, \ldots, x_m) = \frac{1}{2^m}(1 + e_1 x_1) \cdot \ldots \cdot (1 + e_m x_m)\,.$$

The aberration $a_\alpha^{(e)}$ of $R_{\{e\}}$ corresponding to the term α is

$$a_\alpha^{(e)} = \frac{(c_\alpha^{(e)})^2}{(c_0^{(e)})^2} = \frac{(\frac{1}{2^m} e_1^{\alpha_1} \cdot \ldots \cdot e_m^{\alpha_m})^2}{(\frac{1}{2^m})^2} = 1$$

because $(e_1, \ldots, e_m) \in \{-1, 1\}^m$. It follows that the aberration $a_\alpha^{(\mathscr{F}_e)}$ of the fraction $\mathscr{F}_e$ corresponding to the term α, for $1 \le |\alpha|_0 \le t$ is

$$a_\alpha^{(\mathscr{F}_e)} = \frac{(c_\alpha - c_\alpha^{(e)})^2}{(c_0 - c_0^{(e)})^2} = \frac{(c_\alpha^{(e)})^2}{(n/2^m - 1/2^m)^2} = \frac{1}{(n-1)^2}$$

because $c_\alpha = 0$ for $1 \le |\alpha|_0 \le t$. Consequently the terms $A_k(\mathscr{F}_e), k = 1, \ldots, t$ of the WLP of $\mathscr{F}_e$ are

$$A_k(\mathscr{F}_e) = \frac{\binom{m}{k}}{(n-1)^2}, k = 1, \ldots, t\,.$$

This means that $A_k(\mathscr{F}_e), k = 1, \ldots, t$ does not depend on the point e which has been removed. In Sect. 3.2 we study some real examples.

3.2 *Examples with one, two or three points*

Now we consider the effect on the WLP of the removal of one, two or three points from some OAs. The OAs are taken from the repository publicly available at http://pietereendebak.nl/oapage/, which has been created by Pieter Eendebak and Eric Schoen, see [6].

Let us consider one of the best OAs, with respect to the WLP criterion, in the class of OAs with $m = 5$ 2-level factors, $n = 12$ runs and strength $t = 2$. Let us denote this OA by $\mathscr{F}$ and its points by $e_1, e_2, \ldots, e_{12}$. Writing the runs as columns and the factors as rows, the fraction $\mathscr{F}$ is

$$\begin{array}{c} \begin{array}{cccccccccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 & e_9 & e_{10} & e_{11} & e_{12} \end{array} \\ \mathscr{F} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 \end{pmatrix} \end{array}$$

The WLP of $\mathscr{F}$ is $A_{\mathscr{F}} = (1,0,0,A_3(\mathscr{F}) = 1.111, A_4(\mathscr{F}) = 0.5556, A_5(\mathscr{F}) = 0)$.

Now we remove each of the twelve points from $\mathscr{F}$ and we compute the corresponding WLPs. The results are reported in Table 1. We observe that, according to the results of Sect. 3.1, $A_1(\mathscr{F}_e) = 5/(12-1)^2 = 0.041$ and $A_2(\mathscr{F}_e) = 10/(12-1)^2 = 0.083$. It is worth noting that there are two different WLPs. More specifically there are 10 fractions $\mathscr{F}_e$ with $A_3(\mathscr{F}_e) = 1.140$ and 2 fractions $\mathscr{F}_e$ with $A_3(\mathscr{F}_e) = 1.405$.

Table 1 WLPs of the twelve 11-run subsets of $\mathscr{F}$.

point[a]	$A_1(\mathscr{F}_e)$	$A_2(\mathscr{F}_e)$	$A_3(\mathscr{F}_e)$	$A_4(\mathscr{F}_e)$	$A_5(\mathscr{F}_e)$
e_1	0.041	0.083	1.140	0.636	0.008
e_2	0.041	0.083	1.140	0.636	0.008
e_3	0.041	0.083	1.405	0.372	0.008
e_4	0.041	0.083	1.140	0.636	0.008
e_5	0.041	0.083	1.140	0.636	0.008
e_6	0.041	0.083	1.140	0.636	0.008
e_7	0.041	0.083	1.140	0.636	0.008
e_8	0.041	0.083	1.140	0.636	0.008
e_9	0.041	0.083	1.140	0.636	0.008
e_{10}	0.041	0.083	1.405	0.372	0.008
e_{11}	0.041	0.083	1.140	0.636	0.008
e_{12}	0.041	0.083	1.140	0.636	0.008

[a] this column specifies the removed run.

Now, we remove all the possible 66 pairs of points $\{e_i, e_j\}, i, j = 1, \ldots, 12$, $i < j$ from $\mathscr{F}$. We denote the fraction obtained by removing the points e_i, e_j from $\mathscr{F}$ by $\mathscr{F}_{e_i,e_j}$. We obtain 7 different WLPs which are reported in Table 2.

Finally, we remove all the possible 220 subsets of three points $\{e_i, e_j, e_k\}$, $i, j, k = 1, \ldots, 12$, $i < j < k$ from $\mathscr{F}$. We denote the fraction obtained by removing the points e_i, e_j, e_k from $\mathscr{F}$ by $\mathscr{F}_{e_i,e_j,e_k}$. We obtain 12 different WLPs which are reported in Table 3.

The results show that WLPs depend on the *points* which are removed from a given OA and not simply on *their number*. Table 1 demonstrates that even when just one-single point is deleted there are two different WLPs. From Table 1 we can also conclude that the worst fractions in terms of WLPs are $\mathscr{F}_{e_3}$ and $\mathscr{F}_{e_{10}}$. However if we remove two points simultaneously, further investigations revealed that the best fraction in terms of WLP was $\mathscr{F}_{e_3,e_{10}}$ because $A_1(\mathscr{F}_{e_3,e_{10}}) = 0$.

Table 2 The different WLPs obtained by removing 2 points e_i and e_j from $\mathscr{F}$.

N^a	$A_1(\mathscr{F}_{e_i,e_j})$	$A_2(\mathscr{F}_{e_i,e_j})$	$A_3(\mathscr{F}_{e_i,e_j})$	$A_4(\mathscr{F}_{e_i,e_j})$	$A_5(\mathscr{F}_{e_i,e_j})$
1	0	0.4	1.6	0.2	0
10	0.04	0.24	1.2	0.68	0.04
20	0.08	0.16	1.2	0.76	0
10	0.08	0.16	1.52	0.44	0
10	0.12	0.16	1.12	0.76	0.04
10	0.12	0.16	1.44	0.44	0.04
5	0.16	0.24	1.12	0.68	0

[a] N is the number of fractions with the given WLP

Table 3 The different WLPs obtained by removing 3 points e_i, e_j and e_k from $\mathscr{F}$.

N^a	$A_1(\mathscr{F}_{e_i,e_j,e_k})$	$A_2(\mathscr{F}_{e_i,e_j,e_k})$	$A_3(\mathscr{F}_{e_i,e_j,e_k})$	$A_4(\mathscr{F}_{e_i,e_j,e_k})$	$A_5(\mathscr{F}_{e_i,e_j,e_k})$
30	0.062	0.321	1.309	0.852	0.012
10	0.062	0.321	1.704	0.457	0.012
10	0.062	0.42	1.21	0.753	0.111
10	0.062	0.519	1.704	0.259	0.012
30	0.16	0.222	1.21	0.951	0.012
50	0.16	0.222	1.605	0.556	0.012
10	0.16	0.321	1.111	0.852	0.111
10	0.16	0.321	1.506	0.457	0.111
20	0.16	0.42	1.21	0.753	0.012
10	0.259	0.222	1.407	0.556	0.111
20	0.259	0.321	1.111	0.852	0.012
10	0.259	0.321	1.506	0.457	0.012

[a] N is the number of fractions with the given WLP

References

1. Butler, N.A., Ramos, V.M.: Optimal additions to and deletions from two-level orthogonal arrays. J. R. Stat. Soc. Ser. B. Stat. Methodol. **69**(1), 51–61 (2007)
2. Chatzopoulos, S.A., Kolyva-Machera, F., Chatterjee, K.: Optimality results on orthogonal arrays plus p runs for s^m factorial experiments. Metrika **73**(3), 385–394 (2011)
3. Fontana, R., Pistone, G., Rogantin, M.P.: Classification of two-level factorial fractions. J. Statist. Plann. Inference **87**(1), 149–172 (2000)
4. Fries, A., Hunter, W.G.: Minimum aberration 2^{k-p} designs. Technometrics **22**(4), 601–608 (1980)
5. Hedayat, A.S., Sloane, N.J.A., Stufken, J.: Orthogonal Arrays: Theory and Applications. Springer New York, New York (2012)
6. Schoen, E.D., Eendebak, P.T., Nguyen, M.V.: Complete enumeration of pure-level and mixed-level orthogonal arrays. Journal of Combinatorial Designs **18**(2), 123–140 (2010)
7. Street, D.J., Bird, E.M.: D-optimal orthogonal array minus t run designs. J. Stat. Theory Practice pp. 1–20 (2018). DOI 10.1080/15598608.2018.1441081
8. Wang, P., Jan, H.: Designing two-level factorial experiments using Orthogonal Arrays when the run order is important. The Statistician **44**(2), 379–388 (1995)

Recent Developments in Statistical Modelling

Quantile Regression Coefficients Modeling: a Penalized Approach

Modelli Quantili Parametrici: un Approccio Penalizzato

Gianluca Sottile, Paolo Frumento and Matteo Bottai

Abstract Modeling quantile regression coefficients functions permits describing the coefficients of a quantile regression model as parametric functions of the order of the quantile. This approach has numerous advantages over standard quantile regression, in which different quantiles are estimated one at the time: it facilitates estimation and inference, improves the interpretation of the results, and is statistically efficient. On the other hand, it poses new challenges in terms of model selection. We describe a penalized approach that can be used to identify a parsimonious model that can fit the data well. We describe the method, and analyze the dataset that motivated the present paper. The proposed approach is implemented in the `qrcmNP` package in R.
Abstract *I coefficienti di una regressione quantilica sono funzioni iniettive dell'ordine del quantile. L'approccio standard è quello di stimare i quantili uno alla volta. Un metodo alternativo è quello di esprimere la forma funzionale dei coefficienti usando un modello parametrico. Questo approccio ha numerosi vantaggi: semplifica le procedure di stima e inferenza, migliora l'interpretazione dei risultati, e risulta statisticamente efficiente. Al tempo stesso, pone nuove sfide in termini di selezione del modello. La nostra proposta è quella di usare un metodo penalizzato che permetta di identificare un modello parsimonioso che rappresenti correttamente la funzione quantilica. In questo articolo descriviamo il metodo, e analizziamo il dataset che ha motivato il lavoro. L'approccio proposto è stato implementato nel pacchetto R `qrcmNP`.*

Key words: Lasso penalty, Penalized integrated loss minimization, Penalized quantile regression coefficients modeling, Inspiratory capacity

Gianluca Sottile
University of Palermo, Department of Economic, Business and Statistical sciences, Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Paolo Frumento and Matteo Bottai
Karolinska Institutet, Institute of Environmental Medicine, Unit of Biostatistics, Stockholm, Sweden, e-mail: paolo.frumento@ki.se, matteo.bottai@ki.se

1 Introduction

Quantiles fully describe the conditional distribution of a response variable given covariates. Quantile regression (QR; [6]) and its generalizations (e.g., [3]) are the standard tools for quantile modeling. In QR, the conditional quantile function is usually written as

$$Q(p \mid \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}(p), \tag{1}$$

where $\boldsymbol{x}$ is a q-dimensional vector of covariates, and $\boldsymbol{\beta}(p)$ is a vector of unknown coefficients describing the relationship between $\boldsymbol{x}$ and the p-th quantile of the response variable, $p \in (0,1)$. In standard quantile regression, different quantiles are estimated one at the time. When a grid of quantiles is computed, e.g., $p = 0.01, 0.02, \ldots, 0.99$, results can only be summarized graphically. The estimated coefficients are generally non-smooth functions of p and may suffer from high volatility, which can hinder their interpretability.

Recently, [5] suggested modeling the quantile regression coefficient functions, $\boldsymbol{\beta}(p)$, by using parametric functions. Model (1) is reformulated as follows:

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T \boldsymbol{\beta}(p \mid \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\theta}$ is a vector of model parameters. This approach is referred to as *quantile regression coefficients modeling* (QRCM) and permits modeling the entire quantile function, while keeping the quantile regression structure expressed by equation (1). Consider, for example, describing $\boldsymbol{\beta}(p \mid \boldsymbol{\theta})$ by k-th degree polynomial functions:

$$\beta_j(p \mid \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1} p + \ldots + \theta_{jk} p^k, j = 1, \ldots, q.$$

Each covariate has $(k+1)$ associated parameters, for a total of $q \times (k+1)$ model coefficients. When either q or k are large, estimation may become difficult and the model may be poorly identified, causing the variability to grow out of control.

Among different approaches discussed in literature, the least absolute shrinkage and selection operator (LASSO; [9]) is the most used method to perform model selection. This procedure requires selecting a tuning parameter. In the literature, traditional criteria include cross-validation (CV), Akaike's information criterion (AIC), and Bayesian information criterion (BIC).

Numerous papers (e.g., [1, 10]) have investigated the estimation of penalized quantile regression models in high-dimensional setting using the L_1-norm of the coefficients, denoted by L_1-QR [1, 8]. These approaches, however, focus on model selection when estimating one quantile at a time. Generally, this is inefficient and makes it difficult to interpret the results, because some coefficients could be only significant at some quantiles.

We propose applying the L_1-penalty to the integrated loss function described by [5], which is minimized to estimate the unknown parameter $\boldsymbol{\theta}$ in model (2). We refer to this procedure as *penalized quantile regression coefficients modeling* (QRCMPEN).

The paper is structured as follows. We introduce a penalized estimator in Section 2, and propose criteria to select the tuning parameter in Section 3. Section 4 concludes the paper with the analysis of the dataset that motivated this research.

2 The estimator

We assume that model (2) holds, and adopt the following parametrization: $\boldsymbol{\beta}(p \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p)$, where $\boldsymbol{b}(p) = [b_1(p), \dots, b_k(p)]^{\mathrm{T}}$ is a set of k known functions of p, and $\boldsymbol{\theta}$ is a $q \times k$ matrix with entries θ_{jh} such that $\beta_j(p \mid \boldsymbol{\theta}) = \theta_{j1}b_1(p) + \ldots + \theta_{jk}b_k(p)$, $j = 1, \dots, q$. The conditional quantile function is then rewritten as $Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T\boldsymbol{\theta}\boldsymbol{b}(p)$. The choice of the vector $\boldsymbol{b}(p)$, is something arbitrary when the model is not known in advance, indeed, an intuitive approach could be to choose functions as flexible as possible. Moreover, as discussed by [5], the values of $\boldsymbol{b}(0)$ and $\boldsymbol{b}(1)$ should reflect the assumptions about the support of the outcome, and the interpretation of parameters may be highly dependent on the model specification. Although all outcomes are bounded in practice, including unbounded functions facilitates modeling the tails of the distribution. As shown by [5], estimation is carried out by minimizing

$$\bar{L}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(p \mid \boldsymbol{\theta}))\mathrm{d}p, \tag{3}$$

where $L(\boldsymbol{\beta}(p))$ is the loss function of standard quantile regression given by $L = \sum_{i=1}^{n} (p - I(y_i \leq \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(p)))(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(p))$. This estimation procedure is referred to as *integrated loss minimization* (ILM), and implemented in the `qrcm` package in R.

This modeling approach is very flexible, and usually provides a good fit of the data. However, it tends to generate large models, causing overparametrization and loss of efficiency. To implement an automatic procedure for model selection, we modify the loss function (3) by introducing a L_1-norm penalizing factor:

$$\bar{L}_{\mathrm{PEN}}^{(\lambda)}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(p \mid \boldsymbol{\theta})) + \lambda \sum_{j=1}^{q}\sum_{h=1}^{k} \mid \theta_{jh} \mid \mathrm{d}p, \tag{4}$$

where $\lambda \geq 0$. We refer to this estimation approach as *penalized integrated loss minimization* (PILM). To minimize $L_{\mathrm{PEN}}^{(\lambda)}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we use a pathwise coordinate descent algorithm [4]. The described PILM estimator is implemented in the `qrcmNP` package in R.

3 Tuning parameter selection

With a given set of data, the true model is not known. Having adequate criteria for model selection is therefore crucial. In penalized regression, the tuning parameter

λ balances the trade-off between goodness of fit and efficiency. We denote by $\widehat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\theta}}^{(\lambda)}$ the estimator of $\boldsymbol{\theta}$ obtained by minimizing (4) at a given value of λ. AIC- and BIC-type selectors are grid-search criteria that minimize $\text{Dev}^{(\lambda)} + c_n \cdot \text{df}^{(\lambda)}$, where $\text{Dev}^{(\lambda)}$ is the explained deviance of the model (a measure of goodness-of-fit defined below) corresponding to $\widehat{\boldsymbol{\theta}}$, c_n is a constant that could depend on the sample size n, and $\text{df}^{(\lambda)}$ reflects the number of nonzero elements of $\widehat{\boldsymbol{\theta}}$. To improve efficiency and computation, we propose standardizing both $\boldsymbol{x}$ and $\boldsymbol{b}(p)$. Following [7], we define $\text{Dev}^{(\lambda)} = \log \bar{L}^{(\lambda)}_{\text{PEN}}(\widehat{\boldsymbol{\theta}})$, i.e., the logarithm of the minimized loss function given by (4). The AIC and BIC criteria are given by

$$\text{AIC}^{(\lambda)} = \log \bar{L}^{(\lambda)}_{\text{PEN}}(\widehat{\boldsymbol{\theta}}) + n^{-1}\text{df}^{(\lambda)}, \tag{5}$$

$$\text{BIC}^{(\lambda)} = \log \bar{L}^{(\lambda)}_{\text{PEN}}(\widehat{\boldsymbol{\theta}}) + (2n)^{-1}\log(n)\text{df}^{(\lambda)}C_n, \tag{6}$$

where C_n is some positive constant, that diverges to infinity as n increase. The value $C_n = 1$ corresponds to the ordinary BIC.

4 Variables selection for inspiratory capacity

We applied the PILM estimator to a subset (n = 2201) of the data analyzed in [2]. The data refer to a study carried out in 1988-1991 in Northern Italy, and included 1063 males and 1138 females. The study aimed to estimate percentiles of inspiratory capacity (IC), a measure of lungs function. The following nine predictors were available: age, height, body mass index (BMI), sex, and indicators for current smoking, occupational exposure, cough, wheezing, and asthma.

We model the intercept using a linear combination of $\log(p)$ and $\log(1-p)$, that together define the quantile function of the asymmetric Logistic distribution, a very flexible model used to describe possibly skewed random variables with heavy tails, while the coefficients associated with the covariates were described by a shifted Legendre polynomial up to third degree, inclusive of an intercept. The maximal model had $3+4\times 9 = 39$ parameters. We used AIC and BIC to assess model fit. As shown by simulations, AIC criterion tends to select overparametrized models, while BIC criterion is more parsimonious with a higher ability to discard irrelevant covariates ($\boldsymbol{x}$) and basis functions ($\boldsymbol{b}(p)$). Results are reported in Table 1.

Table 1 Model selection based on AIC and BIC criteria. We report the number of parameters, the number of selected covariates, the optimal λ value, the value of the minimized loss function, and the p-value of a Kolmogorov-Smirnov goodness-of-fit test.

Criterion	n. of parameters	n. of covariates	λ	Loss	P-value KS
AIC	31/36	7/9	20.79	293.31	.77
BIC	19/36	4/9	60.47	294.01	.53

We used the model selected by BIC and estimated it again using unpenalized QRCM. The model is represented graphically in Figure 1. Because we were mostly interested in the low quantiles of IC, in Table 2 we only report the estimated quantile regression coefficients, $\widehat{\boldsymbol{\beta}}(p) = \boldsymbol{\beta}(p \mid \widehat{\boldsymbol{\theta}})$, at $p = 0.01$, $p = 0.05$, and $p = 0.50$. Age, height, BMI and sex were statistically significant. Figure 1 shows the regression coefficient functions for all covariates over the interval $p \in (0,1)$. Age had a negative effect at all quantiles, and the associated coefficient function showed an increasing linear trend. Per each one-year increase in age, the 1st and 5th percentile of IC decreased by about 0.014 and 0.013 liters respectively, while its median decreased by about 0.01 liters. Height and BMI both had a positive effect. Quantile regression coefficients were increasing, but had a non linear trend. For each one-centimeter increase in height, quantiles below the median increased by approximately 0.03 liters. For each unit increase of BMI, IC increased by 0.033 and 0.037 at the 1st and 5th percentile, respectively, and by 0.056 at the median. The coefficient function associated with the indicator of male gender was positive and increasing. This indicated that the distribution of IC in males was shifted towards upper values and had a longer right tail than that of females.

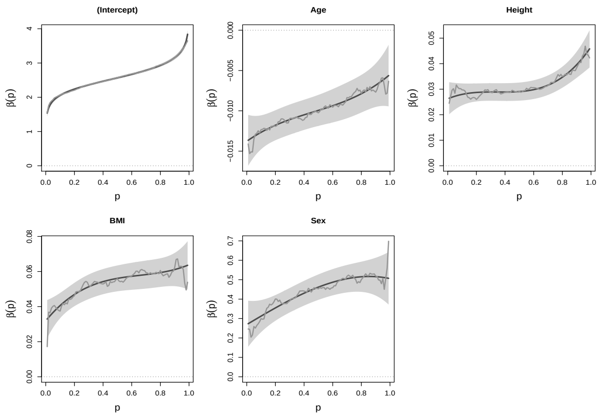


Fig. 1 ILM estimates of $\boldsymbol{\beta}(p)$ under the model selected by BIC. Confidence bands are displayed as shaded areas. The broken lines connect the coefficients of ordinary quantile regression estimated at a grid of quantiles. The dashed line indicates the zero.

Finally, comparing our proposal with standard penalized quantile regression [1, 8] we could observe that it is inefficient and makes it difficult to interpret results, as already mentioned in the introduction. Indeed, different variables are discarded for each percentile, i.e., sex and wheeze for $p = 0.01$, none for $p = 0.05$ and, smoke, occupational exposure, cough and wheeze for $p = 0.50$.

Table 2 Estimated quantile regression coefficients at $p = 0.01$, $p = 0.05$ and $p = 0.50$, obtained from the model selected by BIC. Estimated standard errors in brackets. The asterisk (*) denotes significance less than 0.05.

	$p = 0.01$	$p = 0.05$	$p = 0.50$
Intercept	$1.539(0.049)^*$	$1.893(0.029)^*$	$2.567(0.018)^*$
Age	$-0.014(0.002)^*$	$-0.013(0.001)^*$	$-0.010(0.001)^*$
Height	$0.026(0.003)^*$	$0.027(0.003)^*$	$0.029(0.002)^*$
BMI	$0.033(0.005)^*$	$0.037(0.004)^*$	$0.056(0.004)^*$
Male	$0.273(0.061)^*$	$0.291(0.051)^*$	$0.462(0.033)^*$

5 Discussion

We described a penalized approach that can be applied to the QRCM framework introduced by [5]. Modeling the conditional quantile function parametrically can be more efficient than estimating quantiles one at a time, as in ordinary quantile regression. Moreover, it permits performing model selection directly on the parameters that describe conditional quantiles, instead of proceeding quantile-by-quantile, as the penalized methods for quantile regression proposed so far.

Using this approach has the disadvantage that, as each covariate has multiple associated parameters, the number of model coefficients tends to be large. The PILM estimator demonstrated to select the correct model with a high probability. A computationally efficient algorithm is implemented in the `qrcmNP` package in R.

References

1. Belloni, A., Chernozhukov, V.: L1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics **39**(1), 82–130 (2011)
2. Bottai, M., Pistelli, F., Di Pede, F., Baldacci, S., Simoni, M., Maio, S., Carrozzi, L., Viegi, G.: Percentiles of inspiratory capacity in healthy nonsmokers: a pilot study. Respiration **82**(3), 254–262 (2011)
3. Chaudhuri, P.: Global nonparametric estimation of conditional quantile functions and their derivatives. Journal of multivariate analysis **39**(2), 246–269 (1991)
4. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1), 1–22 (2010)
5. Frumento, P., Bottai, M.: Parametric modeling of quantile regression coefficient functions. Biometrics **72**(1), 74–84 (2016)
6. Koenker, R., Bassett Jr, G.: Regression quantiles. Econometrica: journal of the Econometric Society pp. 33–50 (1978)
7. Lee, E., Noh, H., Park, B.: Model selection via bayesian information criterion for quantile regression models. Journal of the American Statistical Association **109**, 216–229 (2014)
8. Li, Y., Zhu, J.: L1-norm quantile regression. Journal of Computational and Graphical Statistics **17**(1), 163–185 (2008)
9. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B **58**, 267–288 (1996)
10. Wu, Y., Liu, Y.: Variable selection in quantile regression. Statistica Sinica pp. 801–817 (2009)

Simultaneous calibrated prediction intervals for time series

Intervalli di previsione simultanei calibrati per serie storiche

Giovanni Fonseca, Federica Giummolè and Paolo Vidoni

Abstract This paper deals with simultaneous prediction for time series models. In particular, it presents a simple procedure which gives well-calibrated simultaneous predictive intervals with coverage probability equal or close to the target nominal value. Although the exact computation of the proposed intervals is usually not feasible, an approximation can be easily obtained by means of a suitable bootstrap simulation procedure. This new predictive solution is much simpler to compute than those ones already proposed in the literature based on asymptotic calculations. An application of the bootstrap calibrated procedure to first order autoregressive models is presented.
Abstract *Questo lavoro riguarda la costruzione di intervalli di previsione simultanei per serie storiche. In particolare, presenta una semplice procedura per ottenere intervalli di previsione simultanei calibrati con probabilità di copertura uguale o molto vicina al valore nominale. Sebbene il calcolo esatto di questi intervalli non sia sempre possibile, essi si possono approssimare tramite un'opportuna procedura bootstrap. Le approssimazioni così ottenute hanno il vantaggio di essere molto più semplici da calcolare delle soluzioni asintotiche già note. Viene infine presentata un'applicazione della procedura di calibrazione bootstrap per la previsione in modelli autoregressivi del primo ordine.*

Key words: Calibration, coverage probability, prediction intervals, time series.

Giovanni Fonseca and Paolo Vidoni
Dipartimento di Scienze Economiche e Statistiche, Università di Udine. e-mail: giovanni.fonseca@uniud.it and paolo.vidoni@uniud.it

Federica Giummolè
Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari Venezia. e-mail: giummole@unive.it

1 Introduction

In the statistical analysis of time series, a key problem concerns prediction of future values. Although, in the literature, great attention has been received by pointwise predictive solutions, in this paper we deal with the notion of prediction intervals, which explicitly takes account of the uncertainty related to the forecasting procedure. In particular, we assume a parametric statistical model and we follow the frequentist viewpoint, with the aim of constructing prediction intervals having good coverage accuracy.

It is well-known that the estimative or plug-in solution, though simple to derive, is usually not adequate. In fact, it does not properly take account of the sampling variability of the estimated parameters, so that the (conditional) coverage probability of the estimative prediction intervals may substantially differ from the nominal value.

Improved prediction intervals based on complicated asymptotic corrections have been proposed in a general framework by Barndorff-Nielsen and Cox (1996) and Corcuera and Giummolè (2006) and, for the case of time series models, by Giummolè and Vidoni (2010) and Vidoni (2004). A calibrating approach has been suggested by Beran (1990) and applied, for example, by Hall *et al.* (1999), using a suitable bootstrap procedure. Fonseca *et al.* (2011, 2014) extended this proposal and presented applications to one-step ahead prediction intervals for stationary time series models. Simulation-based prediction intervals for autoregressive processes are also considered by Kabaila and Syuhada (2007). Finally, there is an extensive literature on non-parametric bootstrap prediction intervals for autoregressive time series (see, for example, Clements and Kim, 2007 and references therein).

In this paper we extend the bootstrap calibration procedure proposed in Fonseca *et al.* (2011, 2014) to the multidimensional case. In particular this work is dedicated to the construction of joint prediction regions which are expected to contain a future sequence of observations with the required coverage probability. Although the specification of a multivariate prediction region may be quite general, we restrict our attention to joint regions of rectangular form, which are usually considered for forecasting future paths of time series observations. Recently, Wolf and Wunderli (2015) introduce a similar system of simultaneous prediction limits. In the last section, an application to simultaneous prediction within AR(1) models is presented.

2 Simultaneous calibrated prediction intervals

Given a discrete-time stochastic process $\{Y_t\}_{t\geq 1}$, we assume that $Y=(Y_1,\dots,Y_n)$, $n>1$, is observable, while $Z=(Z_1,\dots,Z_m)=(Y_{n+1},\dots,Y_{n+m})$, $m\geq 1$, is a future or not yet available random vector, corresponding to an m-dimensional sequence of future observations. The vector (Y,Z) is continuous with $g(z|y;\theta)$ and $G(z|y;\theta)$, $\theta\in\Theta$, the conditional multivariate density and distribution function of vector Z given

$Y = y$. In the presence of a transitive statistic U, y is substituted by the observed value u of U.

Given the observed sample $y = (y_1, \dots, y_n)$, a system of simultaneous α-prediction limits for vector Z is a set of functions $c^j_\alpha(y)$, $j = 1, \dots, m$, such that, exactly or approximately,

$$P_{Y,Z}\{Z_j \leq c^j_\alpha(Y), j = 1, \dots, m; \theta\} = \alpha, \tag{1}$$

for every $\theta \in \Theta$ and for any fixed $\alpha \in (0,1)$. In the presence of a finite dimensional transitive statistic, we usually consider the conditional coverage probability

$$P_{Y,Z|U}\{Z_j \leq c^j_\alpha(Y), j = 1, \dots, m|U = u; \theta\} = \alpha. \tag{2}$$

An α-level joint prediction region of rectangular form is readily obtained by specifying two suitable systems of lower and upper simultaneous prediction limits.

As we can see, for example in Ravishanker *et al.* (1991) and Alpuim (1997), simultaneous prediction limits for vector Z in a time series context are usually defined as

$$z_{j,\alpha} = z_{j,\alpha}(Y;\theta) = P_j + h_\alpha(\theta) se_j(\theta), \quad j = 1, \dots, m, \tag{3}$$

evaluated at $\theta = \hat{\theta}$, where $\hat{\theta} = \hat{\theta}(Y)$ is the maximum likelihood estimator for θ, or an asymptotically equivalent alternative. Here $P_j = P_j(Y;\theta)$ is a suitable unbiased point predictor for Z_j, such that $E_{Y,Z_j}(Z_j - P_j) = 0$, with prediction standard error $se_j(\theta) = \sqrt{V_{Y,Z_j}(Z_j - P_j)}$. Indeed, $h_\alpha(\theta)$ is a quantity satisfying

$$P_{Z|Y}\{\mathscr{E}_j \leq h_\alpha(\theta), j = 1, \dots, m|Y = y; \theta\} = F\{h_\alpha(\theta), \dots, h_\alpha(\theta)|y;\theta\} = \alpha,$$

with $\mathscr{E}_j = (Z_j - P_j)/se_j(\theta)$, $j = 1, \dots, m$, the standardized forecast errors with joint distribution function $F(e_1, \dots, e_m|y;\theta)$, conditional on $Y = y$.

In order to compute the prediction limits specified by relation (3), we need a vector of unbiased point predictors $P = (P_1, \dots, P_m)$, the associated vector of prediction standard errors $se(\theta) = \{se_1(\theta), \dots, se_m(\theta)\}$ and the quantity $h_\alpha(\theta) = \varphi^{-1}(\alpha|y;\theta)$, where $\varphi^{-1}(\cdot|y;\theta)$ is the inverse of function $\varphi(x|y;\theta) = F(x, \dots, x|y;\theta)$, which corresponds to the conditional distribution function $F(e_1, \dots, e_m|y;\theta)$ constrained to $\{(e_1, \dots, e_m) \in \mathbf{R}^m | e_1 = \dots = e_m = x\}$. For stationary linear models, we usually consider the optimal predictors $P_j = E(Z_j|Y)$, $j = 1, \dots, m$. Indeed, with this choice for the point forecasts, provided that we have a linear or a Gaussian process, the vector of the (standardized) forecasts errors $(\mathscr{E}_1, \dots, \mathscr{E}_m)$ is independent of Y.

The (unconditional) coverage probability of the estimative simultaneous prediction limits $\hat{z}_{j,\alpha} = z_{j,\alpha}(Y;\hat{\theta})$, $j = 1, \dots, m$, corresponds to

$$\begin{aligned}
P_{Y,Z}\{Z_j \leq \hat{z}_{j,\alpha}, j = 1, \dots, m; \theta\} &= E_Y[P_{Z|Y}\{Z_j \leq \hat{z}_{j,\alpha}, j = 1, \dots, m|Y;\theta\};\theta] \\
&= E_Y[P_{Z|Y}\{\mathscr{E}_j \leq (\hat{z}_{j,\alpha} - P_j)/se_j(\theta), j = 1, \dots, m|Y;\theta\};\theta] \\
&= E_Y[F\{a_1 + h_\alpha(\hat{\theta})b_1, \dots, a_m + h_\alpha(\hat{\theta})b_m|Y;\theta\};\theta] = D(\alpha, \theta),
\end{aligned}$$

where $a_j = a_j(Y,\theta) = (\hat{P}_j - P_j)/se_j(\theta)$ and $b_j = b_j(Y,\theta) = se_j(\hat{\theta})/se_j(\theta)$, $j = 1,\dots,m$, with $\hat{P}_j = P_j(Y;\hat{\theta})$.

Following the calibrating procedure proposed in Fonseca *et al.* (2014) for univariate prediction limits, we may consider function

$$\varphi_c(x|y;\hat{\theta},\theta) = D\{F(x,\dots,x|y;\hat{\theta}),\theta\} = D\{\varphi(x|y;\hat{\theta}),\theta\} \tag{4}$$

instead of $\varphi(x|y;\hat{\theta})$, in order to specify the quantity

$$h^c_\alpha(\hat{\theta},\theta) = \varphi_c^{-1}(\alpha|y;\hat{\theta},\theta) = \varphi^{-1}(D^{-1}(\alpha,\theta)|y;\hat{\theta}) = h_{D^{-1}(\alpha,\theta)}(\hat{\theta}),$$

with $\varphi_c^{-1}(\cdot|y;\hat{\theta},\theta)$ and $D^{-1}(\cdot,\theta)$ the inverse functions of $\varphi_c(\cdot|y;\hat{\theta},\theta)$ and $D(\cdot,\theta)$, respectively. It is easy to show that the calibrated simultaneous prediction limits thus obtained, namely

$$z^c_{j,\alpha}(Y;\hat{\theta},\theta) = \hat{P}_j + h^c_\alpha(\hat{\theta},\theta)\, se_j(\hat{\theta}), \quad j = 1,\dots,m, \tag{5}$$

present a coverage probability equal to the target nominal value α. Indeed, the specification of quantities $h^c_\alpha(\hat{\theta},\theta)$ from (4) determines simultaneous prediction limits satisfying relation (1) exactly for all $\alpha \in (0,1)$.

Whenever a closed form expression for $\varphi_c(x|y;\hat{\theta},\theta)$ is not available, we may consider a suitable parametric bootstrap estimator. Let y^b, $b = 1,\dots,B$, be parametric bootstrap samples generated from the estimative distribution of the data and let $\hat{\theta}^b$, $b = 1,\dots,B$, be the corresponding maximum likelihood estimates. Since $D(\alpha,\theta)$ is defined as an expectation, we define the following bootstrap estimator for (4)

$$\varphi^b_c(x|y;\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} F\{\hat{a}^b_1 + h_\alpha(\hat{\theta}^b)\hat{b}^b_1,\dots,\hat{a}^b_m + h_\alpha(\hat{\theta}^b)\hat{b}^b_m|Y;\hat{\theta}\}|_{\alpha=\varphi(x|y;\hat{\theta})},$$

where $\hat{a}^b_j = (\hat{P}^b_j - \hat{P}_j)/se_j(\hat{\theta})$ and $\hat{b}^b_j = se_j(\hat{\theta}^b)/se_j(\hat{\theta})$, $j = 1,\dots,m$, with $\hat{P}^b_j = P_j(Y;\hat{\theta}^b)$. In this case, the associated α-level quantity permits the definition of a system of simultaneous prediction limits with coverage probability equal to α, apart from an error term depending on the efficiency of the bootstrap procedure.

3 Example: AR(1)

Let $\{Y_t\}_{t\geq 1}$ be a first-order Gaussian autoregressive process with

$$Y_t = \mu + \rho(Y_{t-1} - \mu) + \varepsilon_t, \qquad t \geq 1,$$

where μ and ρ are unknown parameters and $\{\varepsilon_t\}_{t\geq 1}$ is a sequence of independent Gaussian random variables with zero mean and unknown variance σ^2. We assume $|\rho| < 1$ so that the process is stationary. The observable random vector is

$Y = (Y_1,\dots,Y_n)$ and the next m realizations of the process are $Z = (Y_{n+1},\dots,Y_{n+m})$. The conditional distribution of Z given $Y = y$ is Gaussian with mean $\mu_{Z|Y} = (\mu_{n+1},\dots,\mu_{n+m})$, where $\mu_{n+1} = \mu + \rho(y_n - \mu)$, $\mu_{n+k} = \mu + \rho(\mu_{n+k-1} - \mu)$, $k = 2,\dots m$, and variance-covariance matrix Σ where $\Sigma_{ij} = \sigma^2\rho^{|i-j|}$. Indeed, Y_n is a transitive statistic and we evaluate the performance of simultaneous prediction limits by means of their coverage probability conditioned on the observed value y_n of Y_n, as in (2).

A simulation study shows the performance of the proposed predictive solution (5). Conditional coverage probabilities for estimative and bootstrap calibrated prediction limits of level $\alpha = 0.9$ are calculated by means of the simulation technique presented in Kabaila (1999), keeping the last observed value fixed to $y_n = 1$. The results for $m = 2,5$ future variables are collected in Table 1 and show that the bootstrap solution remarkably improve on the estimative one.

Table 1 AR(1) Gaussian model. Conditional coverage probabilities for simultaneous estimative and bootstrap calibrated prediction limits of level $\alpha = 0.9$, conditioned on $y_n = 1$; $m = 2,5$ future observations are predicted; μ is considered to be known and equal to 0; $\rho = 0.5, 0.8$, $\sigma^2 = 1$ and $n = 20$. Estimation is based on 1,000 Monte Carlo replications. Bootstrap procedure is based on 500 bootstrap samples. Estimated standard errors are always smaller than 0.011.

m	ρ	Estimative	Calibrated
2	0.5	0.872	0.898
	0.8	0.859	0.895
5	0.5	0.859	0.893
	0.8	0.857	0.884

Acknowledgements This research was partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003 and by Udine University under the PRID 2017 research grants. Simulation studies were performed on SCSCF, a multiprocessor cluster system owned by Università Ca' Foscari Venezia.

References

1. Alpuim, M.T.: One-sided simultaneous prediction intervals for AR(1) and MA(1) processes with exponential innovations. Journal of Forecasting, **16**, 19–35 (1997).
2. Barndorff-Nielsen, O.E., Cox, D.R.: Prediction and asymptotics. Bernoulli, **2**, 319–340 (1996).
3. Beran, R.: Calibrating prediction regions. Journal of the American Statistical Association, **85**, 715–723 (1990).
4. Clements, M.P., Kim, J.H.: Bootstrap prediction intervals for autoregressive time series. Computational Statistics & Data Analysis, **51**, 3580–3594 (2007).
5. Corcuera, J.M., Giummolè, F.: Multivariate Prediction. Bernoulli, **12**, 157–168 (2006).
6. Fonseca, G., Giummolè, F., Vidoni, P.: Bootstrap Calibrated Predictive Distributions For Time Series. In: Proceedings of S.Co. 2011: Complex Data Modeling and Computationally Inten-

sive Statistical Methods For Estimation and Predictions, Padova, CLEUP, September 19-21, 2011.

7. Fonseca, G., Giummolè, F., Vidoni, P.: Calibrating predictive distributions. Journal of Statistical Computation and Simulation, **84**, 373–383 (2014).
8. Giummolè, F., Vidoni, P.: Improved prediction limits for a general class of Gaussian models. Journal of Time Series Analysis, **31**, 483–493 (2010).
9. Hall, P., Peng, L., Tajvidi, N.: On prediction intervals based on predictive likelihood or bootstrap methods. Biometrika, **86**, 871–880 (1999).
10. Kabaila, P.: An efficient simulation method for the computation of a class of conditional expectations. Australian & New Zeland Journal of Statistics, **41**, 331–336 (1999).
11. Kabaila, P., Syuhada, K.: Improved prediction limits for AR(p) and ARCH(p) processes. Journal of Time Series Analysis, **29**, 213–223 (2007).
12. Ravishanker, N., Wu, L.S.Y., Glaz, J.: Multiple prediction intervals for time series: comparison of simultaneous and marginal intervals. Journal of Forecasting, **10**, 445–463 (1991).
13. Vidoni, P.: Improved prediction intervals for stochastic process models. Journal of Time Series Analysis, **25**, 137–154 (2004).
14. Wolf, M., Wunderli, D.: Bootstrap joint prediction regions. Journal of Time Series Analysis, **36**, 35–376 (2015).

Reversibility and (non)linearity in time series
Reversibilità e (non)linearità nelle serie storiche

Luisa Bisaglia and Margherita Gerolimetto

Abstract The recent literature has proposed a (limited) number of approaches to test for time reversibility, that is one of the main hypotheses in time series. A very interesting proposal is a Gini-based framework that, among other things, includes a test for time reversibility focussing on possible differences between backward and forward autocorrelations. This feature is indeed useful to identify models with underlying heavy tailed and non-normal innovations. In this paper we intend to shed some more light on this and investigate, via Monte Carlo simulations, on the possibility that this test can effectively have power in detecting some form of nonlinearity.

Key words: Reversibility, long memory, bootstrap

1 Introduction

Time reversibility is one of the main features of strictly stationary Gaussian linear stochastic processes. From an intuitive point of view, a stochastic process is said to be time-reversible if its probabilistic structure is invariant with respect to the reversal of the time indices. In an applications perspective, a check for time reversibility is a useful addition to existing diagnostics for stationary data since the absence of this feature (so the process is irreversible) signals the exclusion of serially independent or gaussian processes as candidate models.

Starting from the late Nineties the literature has discussed the issue of testing for time reversibility (Ramsey and Rothman, 1996; Hinich and Rothman, 1998; Chen *et al.*, 2000; Chen, 2003). Among the most recent contributes, Racine and Maasoumi (2007) proposed an approach based on an entropy measure of symmetry. Psaradakis

Luisa Bisaglia
Dept. of Statistical Science, University of Padova, e-mail: luisa.bisaglia@unipd.it

M. Gerolimetto
Dept. of Economics, University of Ca' Foscari, Venice, e-mail: margherita.gerolimetto@unive.it

(2008) introduces a sample index of the deviation from zero of the median of the one-dimensional law of differenced data. Shelef and Schechtman (2016) develop a framework for time series anaylsis based on a set of Gini-based equivalents that, among other things, introduces a test for time reversibility focussing on possible differences between backward and forward (in time) autocorrelation. This feature is indeed useful to identify models with underlying heavy tailed and non-normal innovations. One of the nice advantages of this method is that is based on only first-moment assumption.

While there are several finite sample experiment documenting the capability of this class of test for time reversibility to recognize situations that depart from gaussianity, in particular heavy tailed innovations, there is nothing to our knowledge about the possibility that this test can effectively have power in dectecting some form of nonlinearity. Indeed, although nonlinearity does not necessarily imply time-irreversability, (see, e.g., Lewis *et al.*, 1989) time reversible nonlinear processes appear to be the execption rather than the rule (Tong, 1990). With this in mind, in this paper we propose a Monte Carlo experiment, where time reversibility is tested for a variety of the most common non-linear models. For some type of models, in particular TAR and Markov Switching, the results are very promising and require further investigation that is in the future research lines.

The paper is organizes as follows. In section 2 we recall the Gini-based time reversibility test. In section 3 we propose our Monte Carlo experiment and some selected results. Section 4 concludes.

2 Gini-based reversibility test

2.1 Overview

From a formal point of view, a strictly stationary discrete-parameter stochastic process Y_t is said to be time reversible if $(Y_{t1},...,Y_{th})$ and $(Y_{-t1},...,Y_{-th})$ have the same joint distributions for every $h \in \mathcal{N}$ and any h-tuple $(t_1,...,t_h)$ such that $-\infty < t_1 < ... < t_h < \infty$. Put it differently, Y_t is time reversible if looking forward and backward at the time series result in similar probabilistic structure. Weiss (1975) showed that time reversibility for finite order ARMA processes is a typically Gaussian property.

Shelef and Schechtman (2016) shows that time reversibility can be associated to some form of symmetry. This, in turn, can be investigated by observing the behaviour of the two Gini autocovariances at lag s.

$$\gamma^{G1}_{(t,t_{t-s})} = COV(Y_t, F(Y_{t-s})) \qquad \gamma^{G2}_{(t,t_{t-s})} = COV(Y_{t-s}, F(Y_t)) \tag{1}$$

The above expressions can be viewed as Gini autocovariances looking backward and forward. Under strictly stationarity conditions, the following equality hold for all t and s

$$\gamma^{G_1}_{(s)} = COV(Y_t, F(Y_{t-s})) = COV(Y_{t-j}, F(Y_{t-j-s})) \tag{2}$$

and

$$\gamma^{G_2}_{(s)} = COV(Y_{(t-s)}, F(Y_t)) = COV(Y_{(t-j-s)}, F(Y_{t-j})) \tag{3}$$

where $\gamma^{G_1}_{(s)}$ and $\gamma^{G_2}_{(s)}$ are time independent. Note that $\gamma^{G_1}_{(s)}$ and $\gamma^{G_2}_{(s)}$ are equal in case of time reversibility.

When stationarity holds, a Gini version of the autocorrelation function (ACF) between Y_t and Y_{t-s} can also be defined

$$\rho^{G_1}_{(s)} = \frac{\gamma^{G_1}_{(s)}}{\gamma^{G_1}_{(s=0)}} \qquad \rho^{G_2}_{(s)} = \frac{\gamma^{G_2}_{(s)}}{\gamma^{G_2}_{(s=0)}} \tag{4}$$

It is interesting to observe that for an $AR(1)$ process, $Y_t = \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t$, we have that $\gamma^{G_1}_{(s)} = \phi^s \gamma^{G_1}_{(s=0)}$, then $\rho^{G_1}_{(s)} = \phi_1^s = \rho(s)$ and this indicates that the first Gini-ACF is equal to the traditional ACF, denoted by ρ_s.

Consistent Gini-ACFs estimates are the following (Shelef and Schechtman, 2016)

$$\hat{\rho}^{G_1}_{(s)} = \frac{\sum_{t=1}^{T-s}(Y_{t+s} - \bar{Y})(R(Y_t) - \bar{R}(Y_{1:(T-s)}))}{\sum_{t=1}^{T}(Y_t - \bar{Y})(R(Y_t) - \bar{R}(Y_{1:T}))} \tag{5}$$

and

$$\hat{\rho}^{G_2}_{(s)} = \frac{\sum_{t=1}^{T-s}(Y_t - \bar{Y})(R(Y_{t+s}) - \bar{R}(Y_{(s+1):T}))}{\sum_{t=1}^{T}(Y_t - \bar{Y})(R(Y_t) - \bar{R}(Y_{1:T}))} \tag{6}$$

where $R(Y_t)$ is the rank of Y_t and $\bar{R}(Y_{i:j}) = \sum_{t=i}^{j} R(Y_t)/(j-i+1)$.

The Gini-based framework by Shelef and Schechtman (2016) also includes Gini PACF, defined as the last coefficient of a partial Gini autoregression equation of order s

$$Y_t = \phi^{G_1}_{s1} Y_{t-1} + \phi^{G_1}_{s2} Y_{t-2} + ... + \phi^{G_1}_{ss} Y_{t-s} + \varepsilon_t$$

hence

$$\rho^{G_1}_{(j)} = \phi^{G_1}_{s1} \rho^{G_1}_{(j-1)} + ... + \phi^{G_1}_{ss} \rho^{G_1}_{(j-s)}$$

Plugging the second Gini ACF ($\rho^{G_2}_{(s)}$) in place of the first, leads to a second version of the Gini PACF, that can be called second Gini-PACF. As for the estimation, the two Gini-PACFs can be estimated solving for $\phi^{G_1}_{ss}$ and $\phi^{G_2}_{ss}$, $s = 1, 2, ...$ the implied two systems of equations.

2.2 Testing for time reversibility

Implied by the definition of time reversibility itself, a crucial feature of the Gini autocorrelations is that if the series is time reversible, the Gini-ACFs at the remarkable lags are equal. Hence, generally, if the Gini-ACFs differ, this should indicate that Y_t

and Y_{t-s} are not exchangeable and this in turn implies time irreversibility. In order to capture this feature in case of moving average, MA(q), processes whose ACFs cuts off after q lags, Gini-PACFs should also be taken into consideration. This leads (Shelef and Schechtman, 2016) to the following system of hypotheses at each lag s for the null hypothesis of time reversibility

$$H_{01} : \rho_{(s)}^{G_1} = \rho_{(s)}^{G_2} \quad and \quad H_{01} : \phi_{ss}^{G_1} = \phi_{ss}^{G_2} \tag{7}$$

The alternative hypothesis is that at least one of the two equation is violated (for further detail, see the original article by Shelef and Schechtman, 2016).

The logic followed by the authors aims at identifying large absolute differences between the sample Gini ACFs and PACFs, denoted by, respectively, $\hat{\theta}_{Gini-ACF(s)} = \hat{\rho}_{(s)}^{G_1} - \hat{\rho}_{(s)}^{G_2}$ and $\hat{\theta}_{Gini-PACF(s)} = \hat{\phi}_{ss}^{G_1} - \hat{\phi}_{ss}^{G_2}$. The test statistic then is

$$\sqrt{T}\left|\hat{\theta}_{Gini-ACF(s)} - \theta_{Gini-ACF(s)}, H_0\right| \quad and \quad \sqrt{T}\left|\hat{\theta}_{Gini-PACF(s)} - \theta_{Gini-PACF(s)}, H_0\right|$$

where under the null hypothesis the differences are equal to zero. In practice, large value of the test statistics support rejection of the null hypothesis.

Due to the complicated sampling distribution of the Gini-based estimators, that also involve additional restrictive assumption on the time series, critical values for this test tests are obtained via moving block bootstrap. All details about the algorithms are in Shelef and Schechtman (2016).

3 Reversibility and (non)linearity: preliminary Monte Carlo evidence

In their original paper, Shelef and Schechtman (2016) conduct a Monte Carlo experiment showing that at least at the first lags the proposed Gini-based time reversibility test reaches a reasonably high power when the innovations of the ARMA models are not Gaussian, but Pareto, lognormal and α–stable.

Here we intend to study the power of this test under a different setting, *i.e* in case the data generating process (DGP) is nonlinear. This is a very preliminary study and at this beginning stage we consider only some nonlinear DGPs. They are listed below, innovations are distributed as $N(0,1)$:

1. TAR(1,1), where

$$X_t = \begin{cases} -0.5X_{t-1} + a_t & X_{t-1} \leq 1 \\ 0.4X_{t-1} + a_t & X_{t-1} > 1 \end{cases}$$

$$X_t = \begin{cases} 2 + 0.5X_{t-1} + a_t & X_{t-1} \leq 1 \\ 0.5 - 0.4X_{t-1} + a_t & X_{t-1} > 1 \end{cases}$$

$$X_t = \begin{cases} 1 - 0.5X_{t-1} + a_t & X_{t-1} \leq 1 \\ 1 + a_t & X_{t-1} > 1 \end{cases}$$

2. MS(1), where

$$X_t = \begin{cases} -0.5X_{t-1} + a_t & s_t = 1 \\ 0.4X_{t-1} + a_t & s_t = 2 \end{cases}$$

with $p_{11} = p_{22} = 0.5,\ 0.9$.
3. $BL(0,0,1,1)$, where $X_t = a_t + 0.5X_{t-1}a_{t-1}$
4. $BL(0,0,2,1)$, where $X_t = a_t + 0.8X_{t-1}a_{t-1} + 0.5X_{t-2}a_{t-1}$

The number of Monte Carlo simulations is 2000, the number of bootstrap replications for the moving block bootstrap is 500 and the block size is 30 (following the findings by Shelef and Schechtman, 2016). The considered sample sizes are $T = 200, 500, 1000$.

Table 1 Percentages of rejection ($m = 1,2$ number of lags, nominal level 0.05)

	T=200		*T=500*		*T=1000*	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
DGP1: TAR(1,1)	37.7	20.0	72.8	35.3	96.8	50.3
DGP2: TAR(1,1)	42.8	18.8	88.8	41.2	100	61.4
DGP3: TAR(1,1)	45.8	24.8	79.5	49.6	91.6	50.4
DGP4: BIL(0,0,1,1)	22.4	12.4	45.6	17.4	62.8	35.5
DGP5: BIL(0,0,2,1)	27.6	17.4	57.6	28.7	73.6	43.6
DGP6: MS(1)	100	62.4	100	82.6	100	92.3
DGP7: MS(1)	100	90.8	100	93.8	100	95.7

Our empirical power results are shown in table 1. They clearly reveal a very interesting capability of the test to detect nonlinearity with the increase of T. As expected, the percentage of rejections is very high for the largest sample size ($T = 1000$), especially for TAR and MS model, but also for smaller values of T the behaviour is fairly good. In particular, for the MS DGPs the test detects successfully the nonlinear feature even at $T = 200$. Moreover, results are in line with the performance of the majority of nonlinearity test in literature (see Bisaglia and Gerolimetto (2014) for a recent survey on nonlinearity tests and a comparative Monte Carlo experiment). It should be remarked that, as emphasized by Shelef and Schechtman (2016), the performance tends to deteriorate with the increase of m.

As said at the beginning of this section, this is only a preliminary Monte Carlo experiment. Yet, we find these results promising, in particular the test appears to perform well for Markov Switching models (followed by TAR models). We reckon that this could be effectively an alternative route to check for nonlinearity. Some other investigations are in order both in terms of Monte Carlo simulations (e.g. comparison with analogous test provided in the literature) and possible improvement of the performance of the test, for instance by considering other resampling methods to obtain the critical values.

References

1. Bisaglia, L. and Gerolimetto M. (2014) “Testing for (non)linearity in economic time series: a Montecarlo comparison” *Quaderni di Statistica*, **16**, 5–32.
2. Chen, Y.-T. (2003) “Testing serial independence against time irreversibility” *Studies in Nonlinear Dynamics and Econometrics* **7**, 1-28.
3. Chen, Y.-T., Chou, R. Y. and Kuan, C.-M. (2000) “Testing time reversibility without moment restrictions” *Journal of Econometrics* **95**, 199-218.
4. Lewis, P. A. W., McKenzie, E. and Hugus, D. K. (1989) “Gamma processes” *Communications in Statistics Stochastic Models* **5**, 1-30.
5. Hinich, M. J. and Rothman, P. (1998) “A frequency domain test of time reversibility” *Macroeconomic Dynamics* **2**, 72-88.
6. Racine, J. S. and Maasoumi, E. (2007) “A versatile and robust metric entropy test of time-reversibility, and other hypotheses” *Journal of Econometrics* **138**, 547-67.
7. Ramsey, J. B. and Rothman, P. (1996) “Time irreversibility and business cycle asymmetry” *Journal of Money, Credit, and Banking* **28**, 1-21.
8. Psaradakis, Z. (2008) “Assessing time-reversibility under minimal assumptions” *Journal of Time Series Analysis* **29**, 881-905
9. Shelef, A. and Schechtman, E. (2016) “A Gini-based time series analysis and test for reversibility” *Stat Papers*, DOI: https://doi.org/10.1007/s00362
10. Tong, H. (1990) “Non-linear Time Series: A Dynamical System Approach” *Oxford University Press*, Oxford
11. Weiss, G. (1975) “Time-reversibility of linear stochastic processes” *Journal of Applied Probability* **12**, 831-836

Heterogeneous effects of subsidies on farms' performance: a spatial quantile regression analysis

Effetti eterogenei dei sussidi sulle performance delle aziende agricole: un'analisi basata sulla regressione quantilica spaziale

Marusca De Castris and Daniele Di Gennaro1

Abstract Italian agricultural sector is characterized by a wide heterogeneity which can affect the effectiveness of rural policies and, by consequence, economic performances. Indeed, wide differences arise both at farm (i.e. sector, dimension, etc.) and regional levels. In particular, Giannakis and Bruggeman (2015) show how agricultural policies can provide enlarge regional disparities between advanced and lagged regions. In this paper, we analyse the differential impact of the policies by considering Italian lagged regions. The introduction of a Spatial Autoregressive Quantile model allows to take into account both spatial and farm-specific characteristic. Evidences are found in favour of significant and positive spatial spillovers of the policies, especially for the less performing farms.

Abstract *L'eterogeneità presente nel settore agricolo Italiano può influenzare l'efficacia delle politiche rurali e, conseguentemente, le performance economiche delle aziende. In tal senso, disparità possono emergere in relazioni ai fattori propri delle aziende agricole (settore, dimensione, ecc.) e del contesto regionale. Giannakis e Bruggeman (2015) mostrano come le politiche rurali incrementino le disparità tra aree avanzate ed arretrate. Limitando l'analisi ad alcune regioni del sud, abbiamo analizzato l'impatto differenziale degli incentivi considerando le caratteristiche spaziali delle aziende agricole attraverso una regressione quantilica spaziale. I risultati mostrano esternalità significative e positive delle politiche, soprattutto per le aziende con performance più basse.*

Key words: Spatial Quantile Regression, Agricultural Policies, Policy efficacy

[1] Marusca De Castris, Department of Political Sciences, University Roma Tre; email: marusca.decastris@uniroma3.it

Daniele Di Gennaro, Department of Political Sciences, University Roma Tre; email: daniele.digennaro@uniroma3.it

1 Introduction

This paper aims to evaluate the efficacy of Common Agricultural Policy (CAP) to improve performances by focusing on Italian lagged regions. Agricultural sector is deeply rooted in place-based production processes. The presence of spatial dependence produces biased estimates of the performances. This paper, using data on subsidies and economic results of farms from the RICA dataset, which is part of the Farm Accountancy Data Network (FADN), proposes a spatial Augmented Cobb-Douglas Production Function to evaluate the effects of subsidies on farm's performances. The major innovation of our study is the implementation of a micro-founded quantile version of a spatial lag model (Kim and Muller, 2004) to examine how the impact of the subsidies may vary across the conditional distribution of agricultural performances. Results show a significant decreasing shape along the distribution of the subsidies which becomes negligible for higher quantiles.

2 Data and Methodology

In 2008, EU-27 countries deal with a contraction of agricultural production, in real terms, and a deflationary trend on prices. Additionally, the volatility on both energy and fertilizer markets boosts input prices and contributes to an overall reduction of added value per worker and employment. Under this perspective, Italian case is of particular interest. Italian added value at factor cost increased by 2.4%, while the sectoral share of the GDP remains stable at the 2.3%. However, the good economic performances are not sufficient to reduce the gap between agriculture and the other economic sectors[1].

Indeed, Italy is characterized by several structural problems which affect agricultural performances. These issues include the presence of systematic differences between North and South, the lack of young farmers (only 13,2 % has less than 44 years) and the land abandonment on marginal areas, especially for high altitude zones.

The gap between North and South appear clear in terms of added value per worker unit. Although Southern regions grow more than the ones in the North (3.5% vs 0.6%), the average added value per worker unit is still well below Italian average (19300 vs 22000 €). In this paper, we consider how the structural weakness of Southern agriculture[2] can affect the efficacy of public support by focusing on the

[1] Agricultural added value at factor cost per worker unit is 24316 € (44% of the average of Italian economy).

[2] The development gap between North and South is not limited to primary sector. Indeed, regions located in the South of Italy are recognized, by European Commission, as less developed and transition regions. This classification is based on the levels of GDP and employment. Less developed (resp. transition) regions include the areas where GDP per head is less than 75% (resp. between 75 and 90%) of the EU average. The Italian less developed regions are Campania, Apulia, Calabria, Sicily and Basilicata, while transition regions are Abruzzo, Molise and Sardinia. However, we exclude Campania from our analysis

impact of agricultural policies on economic performances. For the year 2008, we exploit information from RICA dataset[1] by introducing five different variables in our analysis. The final dataset is composed by 1298 farms.

Variable	*Label*	*Unit*	*Description*
Value Added	VA	€	Total Revenues-Current Expenses
Labour	L	Unit	Full time worker
Capital Stock	K	€	Land+Agricutltural Fixed Capital
Land	G	Hectares	Utilised Agricultural Area
Subsidies	S	€	Total amount subsidies per farm

In this table we resume all the major determinants on value added formation in the primary sector: Labour, Fixed Capital, Land and Subsidies. The differential impact of all the different variables in determining and stimulating value added is considered by estimating a Cobb-Douglas APF. The dependent variable, value added, is a proxy of economic performances, while subsidies is a composite variable obtained by adding all the amount of the different public instruments allocated to every farm (i.e. we do not distinguish between National or European fund or between policies devoted to current activities, rural development or capital subsidies) and it can be considered as a global indicator of the public capital. In this sense, our baseline takes the form in:

$$\ln(\mathrm{Y}) = \ln \mathrm{A} + \alpha * \ln \mathrm{L} + \beta * \ln \mathrm{K} + \gamma * \ln \mathrm{G} + \delta * \ln \mathrm{S} \qquad (1)$$

Checking for the presence of spatial dependence, we found evidences of a significant spatial autocorrelation. In this way, our final model becomes:

$$ln(Y) = \rho \mathrm{W} * \ln(\mathrm{Y}) ln\, A + \alpha * ln\, L + \beta * ln\, K + \gamma * \ln G + \delta * ln\, S \qquad (2)$$

This equation takes the traditional form of a so-called Spatial Autoregressive Model[2] (SAR). Equation (2) introduces a spatial weight matrix, W. This matrix is based on a cut-off distance (33 km) ensuring the presence of at least one neighbour for every single farm. In this paper, considering the wide heterogeneity between different farms we make use of a Spatial Quantile version of a traditional SAR.

2.1 Spatial Quantile Regression

Quantile regression is an important method for including heterogeneous effects of covariates on a response variable (Koenker and Hallock, 2001). To include the

for a lack of comparability with the other southern regions, while Abruzzo and Molise are not considered for a lack of information about the farms located in these regions.

[1] Rica is part of the European Farm Accountancy data network (FADN) and it represents the only harmonized survey to collect micro-economic data on firms operating in agricultural. Italian RICA collects information on 11000 farms sampled at regional level. RICA's field of observation considers only the farm with at least 1 hectare of UAA or a production value greater than 2500 Euros.

[2] Lesage and Pace (2009) provide an in-depth analysis on the estimation of a SAR model and its decomposition of the marginal impacts.

presence of interactions, the quantile regression generalisation of the (linear) spatial lag model can be written as:

$$\mathrm{Y} = \rho(\tau)\mathrm{WY} + \mathrm{XB}(\tau) + \mathrm{u} \qquad (3)$$

where $Y = Q_{(\tau)}(Y|X)$ is the conditional quantile function of Y, τ refers to the selected quantile and $B(\tau)$ is the vector of the sensitivity coefficients of the conditional quantile on changes in value of the covariates X. Estimating spatial quantile regression for different quantiles allows to predict the distribution of the outcome variable at given values of the explanatory variables (McMillen and Shimizu,2017). Equation (3) underlines that the spatial parameter, ρ, is dependent from the considered quantile τ, allowing for different degree of spatial dependence across the conditional distribution.

In this paper we follow the two-stage estimation procedure in Kim and Muller(2004)[1].This approach were initially developed to control for endogeneity in "traditional" quantile regression model, but adjustments to deal with the spatial endogeneity in a Spatial Autoregressive quantile model were straightforward.

On the first step, a variable constituted by the spatial lag of Y (in our case Added Value) is regressed over a set of instruments, as in Equation (4):

$$\widehat{\mathrm{WY}} = \mathrm{Z}\theta(\tau) + \mathrm{u}, \qquad Z = [X, WX] \qquad (4)$$

Instruments are selected following the intuition in Kelejian and Prucha (1998). Low order interactions are needed to avoid linear dependence and retain full column rank of the set of instruments (Baltagi et al. ,2014). At the second stage, the variable $\widehat{\mathrm{WY}}$ is added on a quantile regression of Y on the X's.

$$\mathrm{Y} = \rho(\tau)\widehat{\mathrm{WY}} + \mathrm{XB}(\tau) + \mathrm{u} \qquad (5)$$

Clearly, τ refers to the same quantile in both equations (4) and (5). The consistency in this approach is guaranteed by estimating differentiated first stages for every quantile considered, while inference based solely on the second-stage of the procedure can be invalid. For this reason, standard errors for the overall two-stage procedure are bootstrapped.

3 Results

Results of the spatial autoregressive quantile regression are presented in Table 2 and Figure 1. While in Table 2 we report only the estimates for the tails (0.1 and 0.9) and the median of the distribution, Figure 1 represent the entire conditional distribution of the parameters (i.e. every percentile between 0.01 and 0.99).

In overall, results show that labour (Figure 5) is the major components in fostering economic performances with an elasticity of 0.8. The distribution across the quantiles is pretty stationary, highlighting the independence of labour from the level of economic performances (i.e. homogeneous effects).

[1] Chernozhukov and Hansen (2006) propose an alternative approach based on a generalisation of the instrumental variables framework to allow for estimation of quantile models.

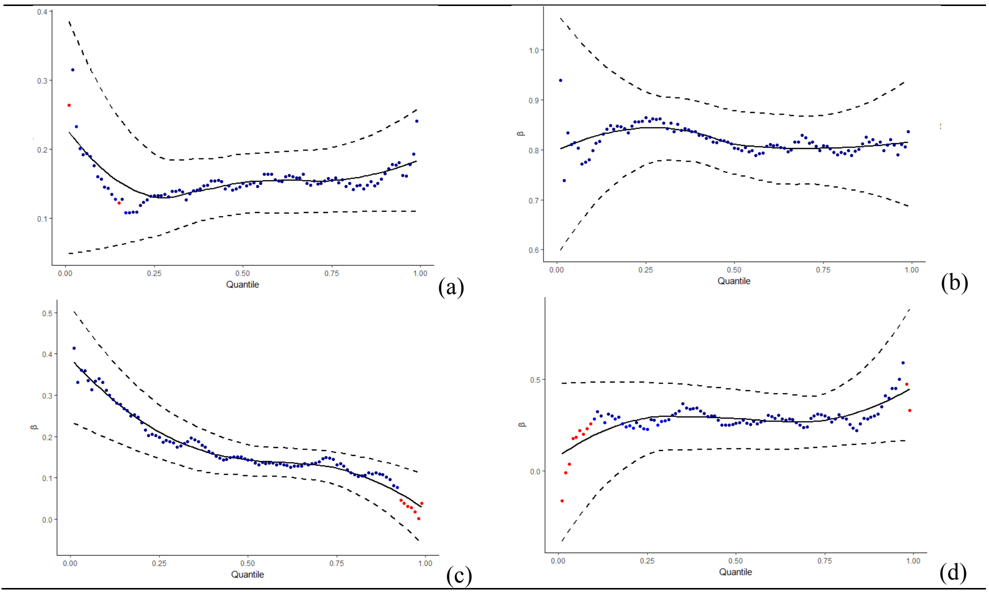
Interestingly, fixed capital has a stronger impact in the extremes of the distribution, presents a decreasing shape with a maximum in lowest quantile which turn to increase at the first quartile (i.e. low and high levels of fixed capital influence more economic performances). However,Subsidies shows major evidences in favour of the heterogeneity of the effects. This component shows a decreasing shape across all the distribution, with an inflection point in the neighbourhood of the median. Surprisingly, lower levels of subsidies have a greater impact on farm's performances (+1% of public funding contributes to an increase of 0.4%in added value), while for the upper tail decrease to less than 0.1 and switch to be not significant. Land follows an increasing distributional shape, but it is not significant across all the distribution.

Table 2: Spatial Quantile Regression Estimation

	Coeff	Z-value	P-Value	Quantile
K	0.16	3.82	0.00	
L	0.80	13.65	0.00	
G	-0.05	-0.89	0.37	0.1
S	0.31	6.61	0.00	
P	0.28	2.26	0.02	
K	0.15	7.19	0.00	
L	0.80	23.98	0.00	
G	0.05	2.55	0.01	0.5
S	0.14	7.79	0.00	
P	0.26	3.78	0.00	
K	0.17	5.55	0.00	
L	0.81	16.88	0.00	
G	0.10	2.95	0.00	0.9
S	0.10	3.66	0.00	
P	0.31	2.98	0.00	

Note: Estimates are reported in terms of elasticities, while the cut-off distance considered is 33 km.

Figure 1: Spatial Quantile Regression Estimation



Note: Figure 1 shows the estimates of the Spatial AR model for every quantile. Panel (a) reports estimates for variable K, (b) for L, (c) for S and (d) for ρ. The graph for the ground is not reported because of a lack of significance. Solid line represents the smoothed function of the estimates, while dashed lines are the confidence interval at 95%. Statistical significance is reported by different colours: Dark Blue= 0.01, Light Blue= 0.05, Red no significance

Lastly, evidences of significant spillover effects are found. Distributional shape of ρ (Figure 1-d) parameter shows a positive and significant effect on economic performances, even if both lower and upper tails are not meaningful. These results provide clear evidences in favour of the existence of spatial patterns on agricultural activities in Italian lagged regions. However, this parameter is not sufficient to estimate the intensity of the spillover effects. In this sense, we decompose the marginal impacts by following the traditional procedure presented in LeSage and Pace (2009).

Table 3: Marginal Impacts

Marginal Impact	K	L	G	S	*Quantile*
D	0.16***	0.80***	-0.05	0.31***	0.1
	[3.33]	[15.11]	[-1.00]	[7.55]	
I	0.06	0.31	-0.02	0.12	
	[1.56]	[1.64]	[-0.84]	[1.60]	
T	0.22**	1.12***	-0.07	0.44***	
	[3.05]	[5.22]	[-1.00]	[4.45]	
D	0.15***	0.81***	0.05°	0.14***	0.5
	[7.07]	[24.41]	[1.93]	[6.56]	
I	0.05*	0.28**	0.02	0.05**	
	[2.26]	[2.50]	[1.41]	[2.69]	
T	0.21***	1.08***	0.06°	0.19***	
	[5.43]	[9.46]	[1.85]	[6.85]	
D	0.17***	0.81***	0.10**	0.10***	0.9
	[5.87]	[17.41]	[2.97]	[4.06]	
I	0.07°	0.36*	0.04°	0.04°	
	[1.93]	[2.22]	[1.72]	[1.84]	
T	0.24***	1.18***	0.14**	0.14***	
	[4.13]	[7.01]	[2.75]	[3.43]	

Note: Table 3 presents the results of the decomposition in direct (D), indirect (I) and total effects (T). Q indicates the considered quantile. Estimates are considered in terms of elasticities, while z-values are in square brackets. The z-values and p-values are estimated by Bootstrap.

Statistical significance: *** <0.001, ** 0.01, * 0.05,°0.1

Spatial quantile regression requires an in-depth analysis for every quantile considered. Table 3 resume the decomposition of the marginal effects for the tails and the median of the conditional distribution. Direct and total effects estimates are positive and significant across all the conditional distribution for all the variables, while results on land are ambiguous and negligible. Nonetheless labour and fixed capital provide evidences of homogeneous effect on the outcome variable conditional to different level of the covariates, we provide evidences of heterogeneous effects for the subsidies.

In detail, the effect of the policies slightly declines for higher quantiles. The wider extension of significant positive effects at the lower quantiles suggests that there is an inverse relationship between subsidies and economic performances. Lower levels of subsidies are devoted to farmers which benefit of agricultural policies as an income maintenance instrument and for which, consequentially, economic performances are mainly affected. Looking at the indirect effects, fixed capital becomes less effective and seems to be not linked to neighbouring characteristics, while positive and significant spillover effects are found in terms of human and public capital. The indirect effects on labour can be explained in terms of favouring qualified labour mobility between neighbouring areas, while the impact on the subsidies is mainly related to structural, environmental and administrative conditions which are in deeply rooted in provincial and regional structure and which can be shared between neighbouring farms.

Conclusions

The heterogeneity arising in Italian agricultural sector can be deeply analysed by a spatial quantile regression model which highlights how both spatial and individual characteristics can influence the performance. Our analysis shows homogeneous effect of farm-specific factors (i.e. labour, fixed capital and land) on economic performance, while heterogeneous impacts of the policies are found, in particular for less performing farms. Evidences of positive and significant spillovers are limited to the inner part of the distribution. In this sense, this paper confirms the hypothesis that actual rural policies are designed as an income support instrument for the farmers.

References

1. Baltagi, B. H., Fingleton, B., and Pirotte, A. (2014). Spatial lag models with nested random effects: An instrumental variable procedure with an application to english house prices. Journal of Urban Economics, 80(Supplement C):76 – 86.
2. Chernozhukov, V. and Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. Journal of Econometrics, 132(2):491 – 525.
3. Giannakis, E. and Bruggeman, A. (2015). The highly variable economic performance of European agriculture. Land Use Policy, 45(Supplement C):26 – 35.
4. Kelejian, H. H. and Prucha, I. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. The Journal of Real Estate Finance and Economics, 17(1):99–121.
5. Kim, T.-H. and Muller, C. (2004). Two-stage quantile regression when the first stage is based on quantile regression. Econometrics Journal, 7(1):218–231.
6. Koenker, R. and Hallock, K. (2001). Quantile Regression. Journal of Economic Perspectives, 15(4):143–156.
7. LeSage, J. and Pace, R. K. (2009). Introduction to Spatial Econometrics. CRC Press.
8. McMillen, D. and Shimizu, C. (2017). Decompositions of spatially varying quantile distribution estimates: The rise and fall of Tokyo house prices. Urbana, 51:61801.

On the estimation of high-dimensional regression models with binary covariates

Sulla stima dei modelli di regressione ad alta dimensionalità con covariate binarie

Valentina Mameli[1], Debora Slanzi[1,2] and Irene Poli[1,3]

Abstract In this paper we address the problem of estimating the parameters of high dimensional regression models characterized by binary covariates. We suggest a new procedure which combines particular clustering for the binary covariates and group penalized regression for estimating the model parameters. The good performance of the methodology is shown in a simulation study.

Abstract *Questo lavoro affronta il tema della stima di modelli di regressione ad alta dimensionalità con covariate binarie. In particolare, si propone un procedura di stima per questa classe di modelli che combina tecniche di cluster analisi e modelli di regressione penalizzata di gruppo. La metodologia proposta viene valutata con uno studio di simulazione.*

Key words: Binary covariates, Clustering techniques, High-dimensional regression models.

1 Introduction

In many scientific fields of research, recent advances in technology have allowed to gather data sets characterized by a very high number of variables. The sample size of these data can be small compared to the number of variables and only a small number of these variables can be relevant to the study. Moreover, in several contexts binary variables are present to express the presence or absence of particular

[1] European Centre for Living Technology, Ca' Foscari University of Venice, Italy.
[2]Department of Management, Ca' Foscari University of Venice, Italy.
[3]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy.
e-mail: valentina.mameli@unive.it, e-mail: debora.slanzi@unive.it, e-mail: irenpoli@unive.it

elements or features. For this structure of the problem we would like to provide a contribution in developing a procedure to select influential variables and estimate model parameters. Several different methodologies have been suggested in literature for variable selection in high-dimensional models, among these penalized regression models have gained popularity over the last few decades; see among others [4, 7, 1, 11, 6, 2]. Penalized procedures are designed with the aim of both selecting the most relevant variables and estimating the parameters of the models. In a general regression setting, models with continuous explanatory variables have been extensively studied, while models with binary explanatory variables have received much less attention.

The aim of this paper is to derive a new procedure to estimate high-dimensional models by combining the class of penalized regression models with binary variables clustering techniques. We propose to estimate penalized regressions based on the information obtained from the introduction of a grouping structure of covariates. More specifically, we propose a two-step procedure: in the first step, we group the covariates into non-overlapping clusters (or groups) using an approach able to deal with the binary nature of the covariates; in the second step, we select the most relevant clusters and covariates by using a penalized regression procedure in which the information obtained in the clustering phase is embedded.

The paper is organized as follows. In Section 2 we review some variable selection procedures, for individual, for group and for bi-level selection; we then present a new inferential procedure for high-dimensional regression models with binary covariates based on penalized regression models and clustering techniques. In Section 3 we conduct a simulation study to evaluate the performance of the new procedure.

2 Methodology

2.1 Model set-up

Le us consider the multiple linear regression model

$$y_i = X_i\beta + \varepsilon_i, \quad i = 1,\dots,n \tag{1}$$

where $X_i = (x_{i1},\dots,x_{ip})^T$ is a p-dimensional vector of explanatory variables (or covariates), y_i is the response value for the i-th observation, ε_i is the error term, n is the sample size and $\beta = (\beta_1,\dots,\beta_p)$ is the vector of parameters. The vector β is unknown and has to be inferred from the data. When the number of variables, p, is much larger than the sample size n the model is usually referred as a high-dimensional regression model. If only a small number of variables affects the response, the model results to be characterized by the *sparsity* condition [7]. To estimate the vector of regression coefficients β we consider penalized regression models and minimize the following function

$$Q(\beta) = \frac{1}{2n}(y - X\beta)^T (y - X\beta) + P(\beta|\lambda), \tag{2}$$

where $y = (y_1, \ldots, y_n)^T$ is the $n \times 1$ vector of response values and $X = (X_1, \ldots, X_n)^T$ is the $n \times p$ design matrix. The function $P(\cdot)$ is defined as a penalty on the regression coefficient parameters β and λ is a tuning parameter. The most used methods for choosing λ are cross-validation criteria or information criteria (see [9] among others). A number of penalized regression methods have been proposed in literature; see among others [6]. They include procedures for individual variable selection, group variable selection and bi-level variable selection. The Least Absolute Shrinkage Selection Operator (LASSO) proposed by [7] is one of the most famous procedure for individual variable selection. If the main interest is in selecting relevant groups of covariates and not individual ones, it is possible to take account of a grouping structure among the covariates as in group penalized regression procedures which include the group LASSO ([10]), the group Minimax Concave Penalty method [6] and the group Smoothly Clipped Absolute Deviation [6]. If, on the other hand, the focus is on selecting both the important groups of covariates as well as variables within these groups, bi-level selection procedures can be considered as the composite Minimax Concave Penalty ([1]), and the group exponential LASSO ([2]). These selection procedures have been introduced to overcome some limitations of the LASSO estimator and present a number of appealing properties in terms of both estimation accuracy as well as variable selection properties. Although a large amount of work has been done in the literature on the selection of continuous covariates in the high-dimensional framework, less attention has been given to the binary covariates case. To address the problem of estimating high-dimensional regression models with binary covariates we develop a methodology based on combining clustering techniques with penalized regression models. In this procedure a crucial step is the selection of binary variables groups to be embedded into the penalized regression model. The group selection can lead to a more effective variable selection and accurate predictions.

2.2 *Estimating the parameters of the clustering structure regression*

For a given clustering structure, the model (1) can be specified as

$$y = \sum_{k=1}^{K} \tilde{X}_k \tilde{\beta}_k + \varepsilon,$$

where $\tilde{X}_k$ is the $n \times d_k$ design matrix representing the d_k covariates belonging to the k-th cluster, $\tilde{\beta}_k = (\beta_{k1}, \ldots, \beta_{kd_k}) \in \mathbb{R}^{d_k}$ is the vector of regression coefficients of the k-th cluster and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is the error vector. Let x_{ij} be the value of X_j at the i-th observation. Assume that $x_{ij} = 1$ if X_j is present in the i-th observation and $x_{ij} = 0$ otherwise, for $i = 1,\dots,n$ and $j = 1,\dots,p$. Let c_j denote a latent cluster label for X_j, with $c_j = k$ if X_j is allocated to the k-th cluster, $k = 1,\dots,K$. The scope of the clustering analysis in this context is to associate to each covariate X_j a unique label c_j with $j = 1,\dots,p$.

Among the clustering techniques the most suitable for binary data include the agglomerative hierarchical clustering methods with binary dissimilarity matrices, the Bayesian non-parametric approach for binary data and the K-modes. Our purpose is to exploit the use of clustering techniques to identify non-overlapping groups of covariates. In order to estimate the grouping structure, in this paper we consider the following clustering methods:

- The standard agglomerative hierarchical clustering algorithms produce a nested sequence of clusters, which initially considers each observation as a single cluster, then at each stage the two least dissimilar clusters are combined. The process is repeated until only one cluster will contain all the observations. The dissimilarity between clusters can be measured by linkage methods: average, complete and single. Moreover, among the main distance measures between objects proposed for binary data we consider the Jaccard and the Tanimoto distances; see [3].
- The Bayesian non-parametric approach, recently proposed by [8], assumes that the data x_{ij} are independent draws from a mixture of infinite Bernoulli distributions whose parameters are distributed according to a Beta distribution. Clustering of data is obtained by calculating the posterior probability of the latent clusters labels c_j for $j = 1,\dots,p$.
- The K-modes approach is a generalization of the K-means procedure suitable for categorical data; see [5]. This approach has two key differences with the classical K-means. First, it assumes that the representative point of the clusters (also known as centroid) is the modal value of a cluster. Second, the distance between objects is the Hamming distance. K-modes tries to find a partition of the objects into K groups by minimizing the distance between each observation and the group centroid.

We propose to estimate the model parameters by clustering the covariates according to a procedure above described and then introducing a group penalty in the estimating function (2). More specifically, the procedure involves the following two-steps:

1. cluster the covariates into non-overlapping groups by using a clustering technique suitable for binary-data;
2. regress the response variable y on the set of grouped covariates using a penalized regression procedure based on embedding the information gained from the preliminary clustering.

3 A simulation study

We conduct a simulation study to evaluate the performance of the clustering effects on the variable selection procedures. Among the several penalties that can be used here we focus on the composite Maximum Concave Penalty (cMCP), the group exponential LASSO (gel), the group Maximum Concave Penalty (gMCP), the group LASSO (gLASSO) and the group Smoothly Clipped Absolute Deviation (gSCAD). For each of these group penalties, we consider different clustering methods in order to introduce a grouping structure into the model. For the purpose of this analysis, we consider the hierarchical clustering with three linkage methods: complete, single, average links, and three dissimilarity measures: Tanimoto and Jaccard distances. Moreover, we consider the K-modes and the Bayesian non-parametric methods. We compared group penalized regression models with the LASSO model. The simulation is based on the linear regression model $y_i = X_i\beta + \varepsilon_i$, $i = 1,\dots,n$, where $\varepsilon_i \sim N(0,\sigma^2)$ as introduced in Equation 1. The standard deviation σ is assumed to be 1 and covariates were generated from a Bernoulli distribution. We consider the following setup: $p = 200$ covariates but just 4 of these covariates have non-zero coefficients. We randomly split the data into training and testing datasets. In this simulation the size of the data set is $n = 100$ and the size of the training set is 80. The number of clusters has been fixed to 10 for all the clustering methods considered. To evaluate the performance of the various group penalization methods combined with different clustering procedures, we calculate some measures of prediction accuracy and variable selection efficiency. In the simulation study, 1000 replicated data sets were generated from the model. For each of these datasets, we compute the Predictive Mean Square Error (PMSE), the Sensitivity (the ratio between the number of selected important variables and the number of important variables), and the Specificity (the ratio between the number of removed unimportant variables and the number of unimportant variables). The results of this simulation are presented in Tables 1 and 2. From Table 1 we can notice that all the penalized regression procedures considered yield satisfactory results in terms of PMSE for all the clustering techniques. In particular, the cMCP penalty provides almost the same good results regardless of the clustering algorithm considered. Moreover, we can notice the very good results achieved in terms of PMSE for the K-modes clustering and the Bayesian non parametric clustering for all penalties considered. Among group selection approaches, the gLASSO achieves the uppermost Sensitivity, especially when we consider the hierarchical approach with Tanimoto distance combined with average and single links. Both the bi-level selection penalties achieve low levels of Sensitivity, but high levels of Specificity, as we expected. In Table 2 we report the results for the LASSO model to allow a comparison with the results of group penalized regressions. We notice that LASSO shows good PMSE and Specificity values but a low value for Sensitivity. From this simulation study we can notice the good performance of this approach which embedded clusters of binary covariates in a group penalized regression model. Further simulation studies and analysis will be developed to evaluate conditions for better performances of this new approach.

Acknowledgements We would like to acknowledge the European Centre for Living Technology (`http://www.unive.it/pag/23664/`) for useful discussions of the research.

Table 1 Simulation results on the performance of the clustering effects on the variable selection procedures over 1000 replicates: Sensitivity, Specificity, PMSE (standard errors between brackets).

		Penalties				
		Bi-level		**Group**		
Clustering methods	**Measures**	cMCP	gel	gLASSO	gMCP	gSCAD
Hierarchical average Jaccard	Sensitivity	0.210 (0.002)	0.140 (0.002)	0.275 (0.009)	0.064 (0.003)	0.255 (0.008)
	Specificity	0.978 (0.001)	0.997 (0.000)	0.650 (0.010)	0.892 (0.004)	0.671 (0.009)
	PMSE	1.659 (0.017)	1.855 (0.018)	2.046 (0.017)	2.058 (0.017)	2.035 (0.017)
Hierarchical complete Jaccard	Sensitivity	0.212 (0.002)	0.104 (0.005)	0.404 (0.016)	0.002 (0.001)	0.008 (0.003)
	Specificity	0.978 (0.001)	0.978 (0.004)	0.591 (0.014)	0.965 (0.002)	0.945 (0.003)
	PMSE	1.662 (0.017)	2.045 (0.022)	2.121 (0.017)	2.083 (0.019)	2.060 (0.018)
Hierarchical single Jaccard	Sensitivity	0.211 (0.002)	0.097 (0.004)	0.443 (0.015)	0.145 (0.011)	0.167 (0.011)
	Specificity	0.977 (0.001)	0.974 (0.004)	0.546 (0.015)	0.848 (0.011)	0.829 (0.012)
	PMSE	1.671 (0.018)	2.061 (0.022)	2.152 (0.018)	2.143 (0.019)	2.131 (0.019)
K-modes	Sensitivity	0.187 (0.002)	0.129 (0.002)	0.238 (0.008)	0.151 (0.000)	0.155 (0.001)
	Specificity	0.985 (0.001)	0.999 (0.000)	0.888 (0.010)	0.997 (0.000)	0.994 (0.000)
	PMSE	1.518 (0.017)	1.811 (0.017)	1.439 (0.015)	1.371 (0.015)	1.390 (0.015)

		Penalties				
		Bi-level		**Group**		
Clustering methods	**Measures**	cMCP	gel	gLASSO	gMCP	gSCAD
Hierarchical average Tanimoto	Sensitivity	0.209 (0.002)	0.081 (0.002)	0.611 (0.015)	0.278 (0.013)	0.311 (0.013)
	Specificity	0.977 (0.001)	0.994 (0.002)	0.422 (0.015)	0.766 (0.013)	0.737 (0.013)
	PMSE	1.698 (0.018)	1.974 (0.019)	2.056 (0.017)	2.122 (0.017)	2.087 (0.017)
Hierarchical complete Tanimoto	Sensitivity	0.211 (0.002)	0.101 (0.002)	0.203 (0.007)	0.087 (0.002)	0.145 (0.003)
	Specificity	0.977 (0.001)	0.999 (0.001)	0.840 (0.007)	0.946 (0.001)	0.897 (0.003)
	PMSE	1.664 (0.017)	1.893 (0.017)	1.957 (0.017)	1.898 (0.016)	1.916 (0.016)
Hierarchical single Tanimoto	Sensitivity	0.208 (0.002)	0.090 (0.004)	0.696 (0.015)	0.456 (0.016)	0.448 (0.016)
	Specificity	0.980 (0.001)	0.984 (0.003)	0.315 (0.014)	0.552 (0.015)	0.558 (0.015)
	PMSE	1.672 (0.017)	2.033 (0.021)	2.101 (0.018)	2.163 (0.018)	2.147 (0.018)
BNP	Sensitivity	0.201 (0.002)	0.165 (0.002)	0.504 (0.008)	0.213 (0.005)	0.462 (0.008)
	Specificity	0.979 (0.001)	0.998 (0.000)	0.642 (0.007)	0.877 (0.003)	0.674 (0.007)
	PMSE	1.614 (0.017)	1.796 (0.017)	1.668 (0.016)	1.700 (0.016)	1.670 (0.016)

Table 2 Simulation results on the performance of the LASSO procedure over 1000 replicates: Sensitivity, Specificity, PMSE (standard errors between brackets).

LASSO		
Sensitivity	Specificity	PMSE
0.236	0.963	1.698
(0.002)	(0.001)	(0.017)

References

1. Breheny, P., Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, **2**, 369–380.
2. Breheny, P. (2015). The Group Exponential Lasso for Bi-Level Variable Selection. *Biometrics*, **71**, 731–740.
3. Everitt, B., Landau, S., Leese, M., Stahl, D. (2011). Cluster analysis. 5th edn, Wiley, Chichester.
4. Galimberti, G., Montanari, A., Viroli, C. (2009). Penalized factor mixture analysis for variable selection in clustered data, *Computational statistics & data analysis*, **53**, 4301–4310.
5. Huang, Z (1998). Extensions to the v-means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, **2**, 28—304.
6. Huang, J., Breheny, P., Ma, S. (2012) A Selective Review of Group Selection in High-Dimensional Models. *Statistical Sciences*, **27**, 481–499.
7. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.
8. Santra, T. (2016) A Bayesian non-parametric method for clustering high-dimensional binary data. `https://arxiv.org/pdf/1603.02494`.
9. Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.
10. Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.
11. Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38**, 894–942.

Social Indicators

Can a neighbour region influence poverty? A fuzzy and longitudinal approach

Può una regione confinante influenzare la povertà? Un approccio longitudinale e sfocato

Gianni Betti, Federico Crescenzi and Francesca Gagliardi

Abstract One of the most important goals of the 2030 UN Agenda for Sustainable Development is to *"…eradicate poverty, in all its forms and dimensions …"*. In order to give a comprehensive answer to such needs, in this paper we propose to adopt a longitudinal measure recently proposed by Verma *et al.* (2017), which is based on the fuzzy set approach to multidimensional poverty: the "Fuzzy At-persistent-risk-of-poverty rate"; then we propose to estimate this measure at regional level via small area estimation techniques, by introducing a spatial correlation model. In this way we are able to take into account whether a neighbour region can influence poverty in all its forms and dimensions, namely, the multidimensional dimension, the regional dimension and the longitudinal dimension.

Abstract Uno dei più importanti *goal* dell'Agenda 2030 per lo Sviluppo Sostenibile delle Nazioni Unite è *"…eradicate poverty, in all its forms and dimensions …"*. Con lo scopo di tentare di rispondere a questa necessità, nel presente lavoro proponiamo di utilizzare una misura longitudinale recentemente proposta da Verma *et al.* (2017), basata sull'approccio multidimensionale e sfocato per la misura della povertà: il cosiddetto "*Fuzzy At-persistent-risk-of-poverty rate*"; inoltre, proponiamo di stimare tale misura a livello regionale, tramite l'introduzione di un modello con correlazione spaziale tra gli errori. In questo modo prevediamo di catturare l'influenza che ha una regione confinante nella misura della povertà, in ogni sua dimensione, ovvero quella multidimensionale, quella regionale, ed infine quella longitudinale.

Key words: small area estimation, fuzzy longitudinal poverty, SILC, Spain.

[1] Gianni Betti, Department of Economics and Statistics, University of Siena, email: gianni.betti@unisi.it

Federico Crescenzi, PhD student, University of Bologna; email: federico.crescenzi2@unibo.it

Francesca Gagliardi, Department of Economics and Statistics, University of Siena, email: gagliardi10@unisi.it

1 Introduction

One of the most important goals of the 2030 UN Agenda for Sustainable Development is to *"…eradicate poverty, in all its forms and dimensions* …" (UN, 2015). This is particularly necessary after the global crisis started in 2008 and the failure of meeting the Millennium Development Goal of halving extreme poverty in the world by 2015.

The need of reducing poverty and launching anti-poverty programmes and policies has also been expressed by the European Union by the Europe 2020 Strategy (European Commission, 2010). This consists in a series of policy objectives called "headline targets", which should be reached by 2020. Among these targets, there is the reduction of the at-risk-of-poverty rate (ARPR, known in the literature as head count ratio, or FGT(0) in the family of Foster *et al*., 1984), and of the at-persistent-risk-of-poverty rate in a longitudinal context for monitoring poverty over time. Moreover, poverty measures are most useful to policy-makers and researchers when they are finely disaggregated, that is when they want to represent geographic units smaller than whole countries; this is exactly the purpose of DG Regional Policy of the European Commission, aiming to use sub-national/ regional level data (NUTS 2) for the social indicators used for monitoring the "headline targets" at the regional level.

In order to give a comprehensive answer to the needs reported above, in this paper we propose to adopt a longitudinal measure recently proposed by Verma *et al.* (2017), which is based on the fuzzy set approach to multidimensional poverty: the "Fuzzy At-persistent-risk-of-poverty rate"; then we propose to estimate this measure at regional level via small area estimation (SAE) techniques, by introducing a spatial correlation model. In this way we are able to take into account whether a neighbour region can influence poverty in all its forms and dimensions, namely, the multidimensional dimension, the regional dimension and the longitudinal dimension.

2 Longitudinal measures of fuzzy and multidimensional poverty

In this section we describe the construction of the fuzzy longitudinal poverty measures, which aim at estimating occasional, persistent or chronic concepts of poverty. In the fuzzy literature, these measures have been defined as: i) anytime, for those individuals belonging to fuzzy set poverty for at least one out four years; ii) continuous, for those belonging to all four years; moreover, we adopt a very recent definition proposed by Verma *et al.* (2017), namely iii) the "fuzzy at-persistent-risk-of-poverty", which refers to those individuals belonging to the fuzzy set in the most recent year, and to two or three years in the last three years: this measure is the fuzzy counterpart of the Eurostat "at-persistent-risk-of-poverty rate", one of the most important Laeken indicators.

From a mathematical point of view, let μ_t be the series of T = 4 membership functions over the four periods, we can define the anytime fuzzy measure as the fuzzy union over the periods, which consists in the maximum of the T values:

$$\mu_{any} = \max(\mu_1, \mu_2, \mu_3, \mu_4) \tag{1}$$

In the same way, the continuous fuzzy measure is defined as the fuzzy intersection over the periods, which consists in the minimum of the T values as:

$$\mu_{cont} = \min(\mu_1, \mu_2, \mu_3, \mu_4) \tag{2}$$

The definition of the fuzzy at-persistent-risk-of-poverty is much more complex, and we suggest to read Verma *et al.* (2017) for a full and detailed description.

3 Model based small area estimation

Sample sizes of surveys like EU-SILC, designed to be representative at national level, are frequently too low to get efficient estimates of indicators at small area level, like NUTS 2. In other words, it means that the measures calculated from such small sub-samples – termed in the related literature as “direct estimators” – have too large variances. Small area estimation theory is concerned with resolving these problems.

This classic EBLUP model proposed by Fay and Herriot (1979) can be extended by considering that the vector of errors follows a Simultaneously Autoregressive Process (SAR) with spatial autoregressive coefficient ρ and proximity matrix $\boldsymbol{W}$ (Cressie, 1993). In this way, the model with spatially correlated random effects is:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{X\beta} + \boldsymbol{Z}(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{u} + \boldsymbol{e}. \tag{3}$$

The estimator is unknown because it depends on some unknown parameters, such as ρ. By substituting them with consistent estimators, a two stage estimator is obtained which can be referred to as a Spatial EBLUP. (see Pratesi and Salvati, 2007, for further details).

In order to estimate the Spatial EBLUP models it is necessary to have the standard errors of the direct estimator $\hat{\theta}$. Since the poverty measures adopted in the present paper are quite complex (such as, for instance, the “fuzzy at-persistent-risk-of-poverty rate”), which are calculated on the basis of a very complex survey such as EU-SILC, we estimate their standard errors by means of the Jackknife Repeated Replication (JRR) in the version of Verma and Betti (2011)[1]..

[1] Verma and Betti (2011) demonstrate how a variant of the JRR method can fit better in case of “complex measures”; moreover, Betti *et al.* (2018) show that the JRR variant of Verma and Betti (2011) is particular adapt for estimating variance of fuzzy poverty measures.

4 Empirical analysis

The reference data for the present work are based on a subset of micro-data from the EU Statistics on Income and Living Conditions (EU-SILC) survey, which is the major source of comparative statistics on income and living conditions in Europe.

Generally, the EU-SILC national surveys are designed with focus on the production of reliable estimates at the national level. In fact, although EU-SILC survey has a very large sample in Spain (13,109 households and 34,756 individuals for 2011), the regional sub-samples are very heterogeneous in size, so that in some NUTS 2 regions estimates are not significant.

From the Spain EU-SILC 2011 Intermediate Quality Report (INE, 2012) we have a detailed description of the sample design that is important both for understanding whether regions form independent sampling domains, and for the construction of the 'computational' PSUs and strata, needed for the estimation of JRR standard errors.

Using such numerical data, here we present analysis of the direct estimates and their relative sampling errors for poverty and deprivation variables. The separate results for each of the 19 regions of Spain allow the input to the Fay and Herriot (FH) and the Spatial EBLUP models. The following two statistics are considered in turn in a longitudinal context: fuzzy monetary poverty rate (FM); and fuzzy supplementary deprivation rate (FS). For each one of these statistics, the longitudinal measures are those described in section 2: any-time poverty, continuous poverty, and at-persistent-of-risk rate.

From the results concerning standard errors of FM fuzzy at-persistent-risk-of-poverty rate, we can appreciate that both FH and SEBLUP are lower than the direct estimates. In general, we have a mean reduction of the standard error for FH of 18%, and for SEBLUP of 26%. The largest reduction standard errors are clearly found in regions with small sub-sample sizes, such as Melilla, Ceuta and Rioja.

However, for the main purpose of this paper, it is quite interesting to observe the larger gain in spatial EBLUP over FH; the geographical information of the w matrix of vicinity clearly supplies an evident added value, being clear that an increase in poverty in a neighbour region can affect the region under investigation as well.

From the results concerning standard errors of FS fuzzy at-persistent-risk-of-poverty rate, the reduction is smaller compared to the corresponding FM measure: 16% for FH and 13% for SEBLUP. In this case, the effect of the geographical information of the matrix of vicinity does not supply an added value.

Analyzing the FM anytime poverty rates, the highest values of are found in Extremadura, Castilla-La Mancha and Andalucia, while the lowest are in Navarra, Aragon and Baleares. The average gains in standard errors for FH and SEBLUB are very similar to the one found for FM at-persistent-risk-of-poverty rate. Again the largest reductions are found in the smallest regions of Melilla and Ceuta.

The measures of the FS anytime poverty rates show that the highest values are found in Galicia, Andalucia, Canarias and Murcia, while the lowest are in Melilla, Navarra, Aragon and Pais Vasco. Again the average gains in standard errors for FH and SEBLUB are very similar to the one found for FS at-persistent-risk-of-poverty rate. The largest reduction is found in the smallest region of Ceuta.

Table 1 reports the FM continuous poverty rates; the highest values of are in Extremadura, Ceuta, Castilla-La Mancha, Andalucia, Murcia and Rioja, the lowest are in Navarra, Pais Vasco, Asturias, Madrid and Melilla. In this case, the average gain in standard error is larger than ones found in all other measures, with mean reduction of about 23% for SEBLUP.

Table 1: FM continuous poverty rates

Region	Direct	se	SEBLUP	se	Gain
Galicia	7.91%	1.14%	8.10%	1.05%	91.59%
Asturias	5.11%	1.03%	5.70%	0.94%	90.98%
Cantabria	7.01%	1.46%	6.49%	1.21%	83.18%
País Vasco	4.43%	1.17%	3.75%	1.06%	91.12%
Navarra	2.13%	0.76%	2.63%	0.73%	96.64%
Rioja	11.24%	3.09%	7.44%	1.58%	51.10%
Aragón	4.77%	1.02%	4.86%	0.93%	91.21%
Madrid	5.59%	0.87%	5.17%	0.88%	101.11%
Castilla y León	8.50%	2.07%	9.04%	1.29%	62.42%
Castilla - La Mancha	13.63%	1.96%	13.31%	1.37%	70.07%
Extremadura	19.49%	1.95%	17.07%	1.55%	79.32%
Cataluña	6.18%	0.92%	6.24%	0.87%	94.36%
Comunitat Valenciana	7.86%	1.17%	8.16%	1.04%	88.80%
Balears	6.04%	1.87%	6.98%	1.59%	85.02%
Andalucía	13.80%	1.58%	13.82%	1.33%	83.91%
Murcia	11.08%	2.96%	11.79%	1.64%	55.29%
Ceuta	14.43%	7.55%	11.83%	2.26%	29.98%
Melilla	5.76%	6.01%	7.68%	2.07%	34.52%
Canarias	9.19%	2.03%	9.02%	1.67%	81.97%
					76.98%

5 Concluding remarks and further considerations

In this paper we propose a series of mathematical procedures to properly estimate longitudinal poverty and deprivation at regional level. First of all, we consider direct estimates of poverty and deprivation measured by means of fuzzy sets theory, and in particular we take into account the new measure “fuzzy at-persistent-risk-of-poverty rate”, recently proposed by Verma *et al.* (2017); then we implant for the first time the procedure of Jacknife Repeated Replications in the version of Verma and Betti (2001) for estimating standard errors of such fuzzy direct estimates.

The primary result obtained is the extension of variance estimation to beyond measures of monetary longitudinal poverty, specifically to fuzzy formulation of those measures and, as a corollary, to multidimensional measures of longitudinal deprivation, which by their very nature are a matter of degree i.e. are fuzzy. To our knowledge, in the literature, no such extension has been published. Moreover, we propose to utilise SAE techniques, such as spatial EBLUP, to further reduce the variability of fuzzy poverty measures at regional level.

Overall, we can conclude that both FH and SEBLUP are able to reduce standard errors by 20-30% in average, with picks of 70% for regions where sample sizes are particular small. Moreover, the larger gain in spatial EBLUP over FH is evident only for FM longitudinal measures, while the gain is practically absent in FS ones, for which the geographical information of the w matrix of vicinity does not supply an added value. So, in conclusion, neighbour region can affect poverty only when we adopt a monetary measure, while it seems to unaffected in the case of a multidimensional or non-monetary measure: further research is necessary to understand reasons of such phenomenon.

References

1. Betti G., Gagliardi F., Verma V. (2018), Simplified Jackknife Variance Estimates for Fuzzy Measures of Multidimensional Poverty, International Statistical Review, doi:10.1111/insr.12219.
2. Cressie N. (1993), Statistics for Spatial Data, New York: Wiley.
3. European Commission (2010), Communication from the Commission. Europe 2020. A strategy for smart, sustainable and inclusive growth. Brussels, 3.3.2010 COM(2010) 2020.
4. Fay R.E., Herriot R.A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data, Journal of the American Statistical Association, 74, pp. 269-277.
5. Foster J.E., Greer J., Thorbecke E. (1984), A class of decomposable poverty measures, Econometrica, 52, pp. 716-766.
6. Instituto Nacional De Estadistica (INE), (2012) Intermediate Quality Report, Survey on Income and Living Conditions Spain (Spanish ECV 2011).
7. Pratesi M., Salvati N. (2007), Small Area Estimation: The EBLUP model based on spatially correlated random effects, Statistical Methods and Applications, 17(1), pp. 113-141.
8. United Nations (2015), Transforming our World: The 2030 Agenda for Sustainable Development, A/RES/70/1, UNITED NATIONS.
9. Verma V., Betti G. (2011), Taylor linearization sampling errors and design effects for poverty measures and other complex statistics, Journal of Applied Statistics, 38(8), pp. 1549-1576.
10. Verma V., Betti G., Gagliardi F. (2017), Fuzzy Measures of Longitudinal Poverty in a Comparative Perspective, Social Indicators Research, 130(2), pp. 435-454.

Weight-based discrimination in the Italian Labor Market: how do ethnicity and gender interact?

L'influenza di genere ed etnia sulla discriminazione basata sul peso corporeo nel mercato del lavoro italiano

Giovanni Busetta, Maria Gabriella Campolo, and Demetrio Panarello[1]

Abstract Access to Italian job market is characterized by a strong degree of discrimination. In this study, we analyze three kinds of discrimination together: gender, race and weight, in order to investigate whether gender and race increase or decrease the impact of weight-based discrimination. To do so, we sent fictitious résumés including photos of either obese or thin applicants in response to online job offers. The results indicate that the strongest kind of discrimination is the racial one, and that discrimination based on weight appears to be stronger for women than men.
Abstract *L'accesso al mercato del lavoro italiano è caratterizzato da un alto grado di discriminazione. In questo studio, analizziamo congiuntamente tre tipologie di discriminazione: di genere, di razza e di peso, al fine di investigare se genere e razza producano un incremento o decremento dell'impatto della discriminazione basata sul peso. Per fare ciò, abbiamo inviato CV fittizi contenenti foto di candidati obesi e normopeso in risposta ad annunci di lavoro online. I risultati indicano che la discriminazione più forte è quella legata alla razza e che la discriminazione basata sul peso sembra essere maggiore per le donne che per gli uomini.*

Key words: Labor market discrimination, Field experiment, Net discrimination

1 Introduction

In the last decades, obesity became a relevant issue for most developed and developing countries [2]. Regarding Europe, OECD showed that obesity rates have doubled in most countries, including UK, Scandinavia and some southern European

[1] Giovanni Busetta, University of Messina; email: gbusetta@unime.it
Maria Gabriella Campolo, University of Messina; email: mgcampolo@unime.it
Demetrio Panarello, Parthenope University of Naples; email: demetrio.panarello@uniparthenope.it

countries [6]. Even if Italy is characterized by one of the slimmest adult population in Europe [7], the country is experiencing a strong increase in the portion of obese people, especially in between children age [5]. Thus, we consider that it will be a relevant and important problem to study for the next years. Moreover, as Italy has always been the country of fashion, a great attention has been devoted to physical appearance. Therefore, we think that weight discrimination must necessarily be a relevant topic, and we want to study discrimination against obese people in Italy.

Furthermore, we compare this kind of labor market discrimination with the other two most common causes of discrimination: against women, and against immigrants. In this respect, according to [8], the most frequent reasons for discrimination in the labor market are indeed age, sex, race and ethnicity, weight and height.

Comparing these three relevant forms of discrimination is then useful to evaluate whether discrimination based on weight has already become worrying. Surprisingly, considering that obesity is a major issue in every developed country, literature has devoted much less attention to the weight bias in the job market, compared to race and gender discrimination [1].

This paper has important policy implications because, as long as discrimination based on weight is rapidly increasing, overweight and obese people would need to be legally protected against it (see [3]).

The rest of the paper is organized as follows. The experimental design and data are presented in Section 2. Section 3 describes the model and shows the main results. Finally, we conclude our study in Section 4.

2 Experimental Design and Data

Our empirical experiment focuses on the impact of obesity in the first step of the hiring process. Therefore, we collected data through a field experiment carried out in Italy in the first semester of 2013. During that period, we sent fictitious résumés in response to 244 real online job postings published in the most important Italian job search websites that do not require registration. Thus, for each offer we prepared and sent 8 ad hoc constructed résumés (in total 1952), one for each fictitious identity that was previously generated accordingly, using fictitious names and addresses and adopting the European format and structure.

To avoid matching problems, we adapted the résumés to the skills required by the firms. For each job offer, we sent 8 equivalent résumés (with equivalent skills and work experience), changing only name, nationality and the photo attached: 4 résumés with different photos of thin applicants (Italian and immigrant woman and man), and 4 with photos of the same candidates modified to appear obese. The photos were manipulated using a software named Fatbooth, available on Apple App Store[2].

3 Methods and Results

Responses were classified as callbacks if the employer requested an applicant to contact them (not just for clarification)[3]. We then analyzed callback rates to see whether some fictitious applicants were discriminated.

We created several indexes measuring separately discrimination against: men, women, locals, immigrants, obese and non-obese candidates. Indexes range from 0 to 1, taking value 0 when no discrimination is observed (this is the case in which the firm replies to all the candidates) and 1 in cases of highest discrimination (none of the categories has been called back). All the values in between are scored 0.25, 0.50 and 0.75, depending on whether the firm decided to call back 3, 2, or 1 categories. In this way, we created several indexes to consider different kinds of discrimination.

Table 1 shows differences in discrimination against men and women, natives and immigrants, normal-weight and obese. They are calculated with respect to the discrimination indexes described above. In all cases, differences are statistically significant, and discrimination operates against women, immigrants and obese with respect to men, natives and non-obese. Results also show that discrimination based on both ethnicity and obesity are quantitatively higher than the one based on gender.

Table 1: Discrimination indexes – men, women, natives, immigrants, normal-weight, obese.

Discrimination against	Mean	Std. Err.	Difference in mean	p-value
Men	0.517	0.017	-0.075	***
Women	0.592	0.017		
Natives	0.419	0.018	-0.271	***
Immigrants	0.690	0.020		
Normal-weight	0.455	0.017	-0.158	***
Obese	0.613	0.018		

Note: *p<0.05; ** p<0.01; *** p<0.001

To make our results comparable with the main literature on the topic, we calculated correspondence tests which are considered a standard methodology [4], in order to use them as robustness check. The correspondence test allows us to calculate the percentage of net discrimination between normal weight individuals and obese in all the considered categories. The preliminary results of our experiment (Table 2, column 11) show that obese are discriminated in all subgroups, and these results are particularly relevant for immigrants. In this case, the concerning call back rate is equal to 0.56 and decreases to 0.46 if the immigrant is a woman. Table 2 also shows

[2] This software, working on the distances between mouth and chin, and enlarging the oval of the face, generates images of obese appearance.

[3] To minimize inconvenience to the employer, invitations were promptly declined since employers who contacted an applicant were contacted themselves by email.

the aggregated results of the correspondence test considering obese as minority group and normal weight as majority one.

Table 2: Correspondence test on weight-based discrimination, by sex and nationality.

	Jobs	No one invited	At least one invited	Equal treatment	Only normal weight invited	Only obese invited	Normal weight	Obese	Net discrimination		Relative call back rate
	1	2	3	4	5	6	7	8	9	10	11
									No.	%	
Woman	244	47	197	104	64	29	168	133	35	17.77	0.79
Man	244	26	218	130	60	28	190	158	32	14.68	0.83
Italian	244	20	224	134	54	36	188	170	18	8.04	0.90
Immigrant	244	93	151	60	75	16	135	76	59	39.07	0.56
Italian W	244	63	181	71	61	49	132	120	12	0.07	0.91
Immigr.W	244	139	105	25	64	16	89	41	48	0.46	0.46
Italian M	244	45	199	93	61	45	154	138	16	0.08	0.90
Immigr.M	244	124	120	40	57	23	97	63	34	0.28	0.65

Note: In the first column, we report the number of the total job offers. We performed a correspondence test on the answers obtained by the firms. In this respect, four are the possible outcomes: no one invited (second column); both invited (third column); only one applicant invited, normal weight or obese (fifth and sixth columns, respectively). In the following two columns, we reported the total number of normal weight and obese invited. Considering these results, we calculated net discrimination as the ratio between the difference in discrimination against obese (column 5) and the one against normal weight (column 6) over the number of jobs for which at least one candidate was invited. Difference in discrimination is reported in column 9, and the previous mentioned ratio is reported in column 10. Finally, we reported the relative callback rate as the ratio between the total of obese and normal weight invited.

The Net Discrimination Index (NDI) calculated shows that weight-based discrimination exists for all the considered subgroups, but it differs depending on the subgroup analyzed. Even if the difference is not high, NDI is higher for women than for men (almost 18% and 15%, respectively). Significant differences are found by comparing Italian (8%) and immigrant candidates (39%). Moreover, if we consider immigrant women, it increases to 46%. From these results, we can easily realize that weight-based discrimination is more relevant for immigrants than Italians. Moreover, in terms of gender gap, it is more pronounced within immigrants than within Italians.

Finally, we estimated a Probit model to analyze whether discrimination based on ethnic origin, weight and sex affect the opportunities of finding a job in the Italian labor market. In our equation model, the dependent variable Y is a dummy variable that takes value 1 if firm replies to the candidate and 0 otherwise (Call back).

$$Y_i = \beta_0 + \beta_1 Woman_i + \beta_2 Obese_i + \beta_3 Immigrant_i + \beta_4 Graduate_i + \beta_5 HighSchool_i + \beta_6 FrontOffice_i + \beta_7 HardWork_i + \beta_8 South_i + \varepsilon_i \qquad (1)$$

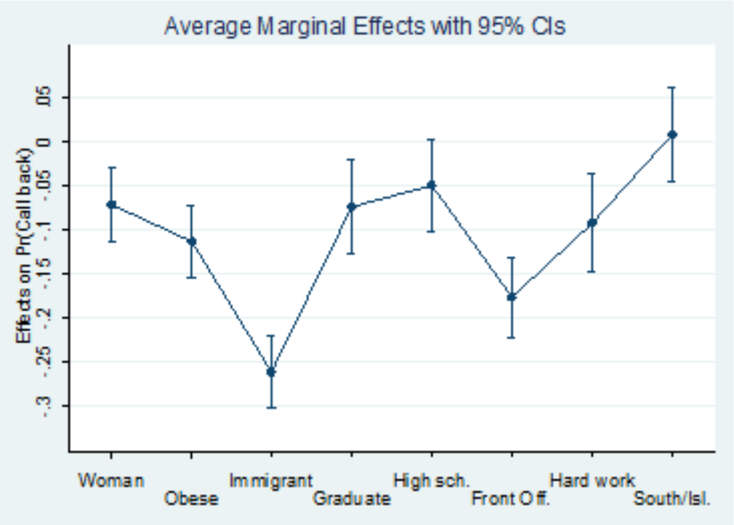
where Y is the latent variable reflecting the probability of the *i-th* subject of receiving a call back. On the right side of equation, with Greek letters we refer to parameters. β_0 is the constant term and ε is the disturbance term. The main explanatory variables are the sex of the candidate (*Woman*: 1=yes; 0=Man); a dummy variable that indicates if the candidate is obese (1=yes; 0=otherwise); a dummy variable that indicates the nationality of the candidate (1=immigrant; 0=Italian); the education level of the candidate, corresponding to the education required by the job opening, divided into three dummy variables (*Graduate*: 1=yes, 0=otherwise; *High School*: 1=yes, 0=otherwise; reference: No title); two dummy variables regarding the job characteristics of the offer (*Front Office:* 1=yes, 0=Back Office, *Hard Work*: 1=yes, 0=Soft work); finally, we consider the geographical area (North-Center=1; South and Islands=0).

Table 3: Probit estimation results.

	Coef.	Std. Err.	p value
Woman	-0.20	0.06	***
Obese	-0.32	0.06	***
Immigrant	-0.71	0.06	***
Graduate	-0.21	0.08	**
High School	-0.14	0.08	
Front Office	-0.49	0.07	***
Hard Work	-0.27	0.08	***
South and Islands	0.02	0.08	
Constant	0.76	0.09	***

Note: *p<0.05; ** p<0.01; *** p<0.001

Figure 1: Probit estimation results – marginal effects at mean



In general, the Probit estimation results (Table 3) show that the minority subgroups (Women, Obese and Immigrant) are more discriminated with respect to the majority ones. The probability of being called back decreases when the education level required by the job increases. Moreover, the probability of being called back is higher for candidates applying for back office and soft work jobs. In Figure 1, we report the marginal effects at mean of our Probit model.

4 Conclusions

In the present analysis, we analyzed weight-based discrimination in the first stage of the hiring process, comparing it with discrimination based on ethnicity and on gender. To do so, we sent fictitious résumés answering to 244 real online job postings published in the most important Italian job search websites that do not require registration. For each offer, we prepared and sent 8 ad hoc constructed résumés, one for each fictitious identity that was previously generated accordingly. In this respect, we compared Italians with immigrant candidates showing that, in the Italian labor market, immigrants are more discriminated than Italians (39% and 8%, respectively). Moreover, results show that weight-based discrimination exists for all the considered subgroups, but it differs depending on the subgroup analyzed. Indeed, it is higher for women than for men (almost 18% and 15%, respectively) and increases even more if women are immigrants (46%).

References

1. Agerström, J., Rooth, D. O.: The role of automatic obesity stereotypes in real hiring discrimination. Journal of Applied Psychology, 96(4), 790 (2011)
2. Finucane, M.M., Stevens, G.A., Cowan, M.J., Danaei, G., Lin, J.K., Paciorek, C. J., ..., Farzadfar, F.: National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9· 1 million participants. The Lancet, 377(9765), 557-567 (2011)
3. Horner, K.: A growing problem: why the federal government needs to shoulder the burden in protecting workers from weight discrimination. Catholic University Law Review, 54, 589 (2005)
4. Jowell, R., Prescott-Clarke, P.: Racial discrimination and white-collar workers in Britain. Race, 11(4), 397-417 (1970)
5. Nardone, P., Spinelli, A., Buoncristiano, M., Lauria, L., Pizzi, E., Andreozzi, S., Galeone, D.: Il Sistema di sorveglianza OKkio alla SALUTE: risultati 2014. Istituto Superiore di Sanità. Roma (2016)
6. OECD: Health at a glance: Europe 2010. OECD Publishing (2010) http://dx.doi.org/10.1787/health_glance-2010-en
7. OECD: Obesity update 2017 (2017) http://www.oecd.org/health/health-systems/Obesity-Update-2017.pdf
8. Roehling, M.V., Roehling, P.V., Pichler, S.: The relationship between body weight and perceived weight-related employment discrimination: The role of sex and race. Journal of Vocational Behavior, 71(2), 300-318 (2007)

The Total Factor Productivity Index as a Ratio of Price Indexes.

L'indice della produttività totale dei fattori come rapporto di indici di prezzo.

Lisa Crosato and Biancamaria Zavanella

Abstract Ever since the seminal work by Robert Solow (1957), total factor productivity (TFP hereafter) has been associated with the movements in the production function caused by technical progress. This has largely influenced the choice of the index numbers involved in the productivity measurement. In fact, the methodology currently in use worldwide is based on the superlative Tornqvist index (Diewert, 1976). In this paper we observe the empirical evidence about TFP in a few countries and propose to re-define the TFP as the ratio of the price index of input to the price index of outputs, using the Sato-Vartia formula.

Abstract *La produttività totale dei fattori viene associata con i movimenti della funzione di produzione causati dal progresso tecnico a partire dal lavoro fondamentale di Robert Solow (1957). Questo ha ampiamente influenzato il modo di misurare la produttività e in particolare la scelta dei numeri indice coinvolti, tanto che la metodologia universalmente usata al momento è basata sull'indice superlativo di Tornqvist (Diewert, 1976). In questo lavoro partiamo dell'evidenza empirica sull'andamento della produttività totale dei fattori in alcuni paesi e ridefiniamo l'indice TFP come rapporto di indici di prezzo, usando la formula di Sato-Vartia.*

Key words: productivity, index numbers, Sato-Vartia, Tornqvist

1 Introduction

Productivity measures the extent to which an economic system transforms the available resources in goods and services. More specifically, we can define productivity as the ratio between the volume of output and of the input utilized to generate it or between the results achieved by the economic system of a region/country and the imputed factors (OECD, 2001). On the empirical side, for any production unit, the

Università di Milano-Bicocca - DEMS - Via Bicocca degli Arcimboldi, 8, 20126 Milano.
e-mail: lisa.crosato@unimib.it; biancamaria.zavanella@unimib.it

total factor productivity index is defined as an output quantity index divided by an input quantity index (Balk, 2010).

The interest in productivity measurement has been strong ever since the late thirties, starting from the introductory papers by Evans and Siegel (1942) and Magdoff (1939). Later on, the seminal work of Solow (1957), briefly followed by the quarrel with Pasinetti (1959) intertwined the total factor productivity (TFP) with technical progress.

The current methodology in use at official statistics offices for measuring Total Factor Productivity (TFP hereafter) derives from the Solow model rooted into a neoclassical framework and ends up in the ratio of two quantity indexes: the Tornqvist quantity input index and an implicit output quantity index derived through the output price index of Paasche (OECD, 2001). This way of constructing the TFP index, however, is not consistent with the axiomatic theory of index numbers and leads to puzzling empirical patterns.

This paper proposes to rethink the TFP index in terms of the ratio of the input price index to the output price index, by means of the Sato-Vartia formula. The underlying idea builds on the empirical evidence on the TFP changes, which for several countries show a fluctuating behaviour hardly consistent with the mechanisms driving technological progress. On the contrary, the relative changes between input and output prices are of a more volatile nature so that our representation could be more consistent with the observed pattern of TFP. Furthermore, the use of Sato-Vartia formulas is in line with the index numbers theory since they satisfy a few axiomatic tests failed by the Tornqvist and the implicit quantity indexes.

Among other factors besides technological progress shaping the pattern of TFP and causing noise, we can surely include the relative input/output cost or price behaviour. So a look at the TFP index on the index price side could maybe give a more consistent interpretation of this puzzling behaviour.

The remaining of the paper is organized as follows: section 2 sketches the definition of TFP index by means of Sato-Vartia input and output price indexes, Section 3 illustrates the empirical evidence for Italy, section 4 concludes.

2 Methodology

Our starting point is the current methodology outlined in the OECD manual, which builds on the Cobb-Douglas production function and then factorizes the parameter representing the tecnhical progress or TFP. Under the usual neoclassical hypotheses, output elasticities with respect to capital and labour match the shares of added value going to the same factors.

Through standard mathematical tools the TFP index can be derived as an output index divided by the Divisia index synthetising the real variation of production factors, weighted by the corresponding shares (Star and Hall, 1976). However, as it is well known, the Divisia index cannot be applied to real data because the index is continuous and the data are discrete, and since Diewert (1976) it has been approximated by the Tornqvist formula, chosen mainly on the basis of its relations to the possible functional forms for the production function. All this results in the definition of the TFP index as the ratio between an implicit quantity index for outputs (given by the nominal variation of value added over the Paasche price output index) and the Tornqvist index for input quantities. In formulas, let n be the number of possible outputs, L and K the labour and capital inputs respectively, V the value, p_{ti} the output prices and y_{ti} the output quantities for $i = 1, \ldots, n$, w_t the cost and L_t the quantity of labour, u_t the cost and K_t the quantity of capital. Then

$$u_t K_t + w_t L_t = \sum_{i=1}^{n} p_{it} y_{it}$$

indicates the aggregate value at time t so that the value index is given by

$$\frac{u_t K_t + w_t L_t}{u_{t-1} K_{t-1} + w_{t-1} L_{t-1}} = \frac{\sum_{i=1}^{n} p_{it} y_{it}}{\sum_{i=1}^{n} p_{it-1} y_{it-1}} =_{t-1} V_t \tag{1}$$

Now the value index can be decomposed into the product of a price Paasche index and a Laspeyres quantity index for output (${}_{t-1}P_t^{Po}$ and ${}_{t-1}Q_t^{Lo}$ respectively)

$$\frac{\sum_{i=1}^{n} p_{it} y_{it}}{\sum_{i=1}^{n} p_{it-1} y_{it-1}} =_{t-1} P_t^{Po} \cdot_{t-1} Q_t^{Lo} =_{t-1} V_t \tag{2}$$

According to the standard methodology of the OECD, the real variation of value added is then measured by the implicit quantity index

$$\sum_{i=1}^{n} p_{it} q_{it} / {}_{t-1}P_t^{Po} \tag{3}$$

and the TFP from $t-1$ to t is defined by

$${}_{t-1}V_t / {}_{t-1}P_t^{Po} =_{t-1} TFP_t \cdot {}_{t-1}Q_t^{TI} \tag{4}$$

with ${}_{t-1}Q_t^{TI}$ being the Tornqvist input quantity index.

However, the Tornqvist index does not satisfy the product test of index numbers and, as Balk (2010) points out, ”the implicit quantity index does not satisfy the Identity Test”. Therefore, the standard TFP index currently used results non consistent into the framework of axiomatic index number theory.

We propose to resort to the Sato-Vartia index instead, substituting it to the Tornqvist formula in (4)

$${}_{t-1}V_t / {}_{t-1}P_t^{Po} =_{t-1} TFP_t \cdot {}_{t-1}Q_t^{SVI} \tag{5}$$

Once again, (5) says that the TFP index is given by the ratio of the output quantity index to the input quantity index with the usual interpretation as a technical progress indicator and shitfting of the production function. Recalling the decomposition of the nominal variation of the value added through the Sato-Vartia index (see for instance Martini (2001))

$$\frac{\sum_{i=1}^{n} p_{it} y_{it}}{\sum_{i=1}^{n} p_{it-1} y_{it-1}} =_{t-1} P_t^{SVo}\,{}_{t-1}Q_t^{SVo} =_{t-1} V_t \tag{6}$$

we can express the TFP as

$$\begin{aligned} {}_{t-1}TFP_t &= {}_{t-1}Q_t^{Lo} / {}_{t-1}Q_t^{SVI} \\ &= \frac{{}_{t-1}V_t}{{}_{t-1}V_t} \frac{{}_{t-1}Q_t^{Lo}}{{}_{t-1}Q_t^{SVI}} \\ &= \frac{{}_{t-1}P_t^{SVI}}{{}_{t-1}P_t^{Po}} \end{aligned} \tag{7}$$

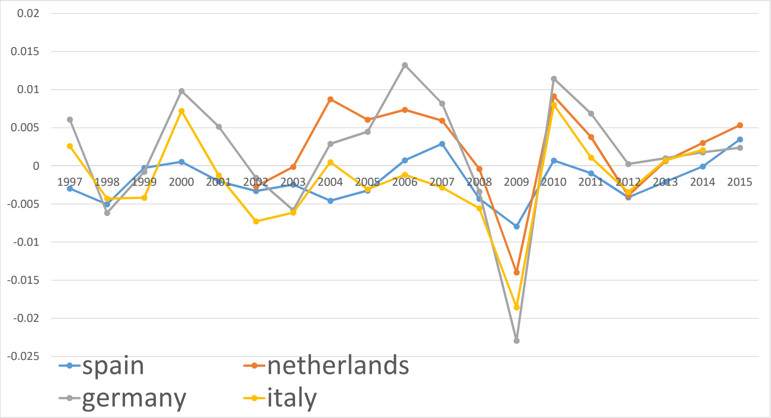
so that the TFP index can be represented, with no loss of information, as the ratio of the Sato-Vartia input cost index and the Paasche output price index. Furthermore, writing the TFP in this way we are explicitating that it is influenced by several different factors that maybe can contribute to justify the ups and downs of TFP detected empirically.

3 Empirical evidence

The empirical evidence we started working on in order to develop the idea underlying this work is based on data from the EUKLEMS database (www.euklems.net). Euklems collects several measures related to economic growth and productivity, at the industry level for all EU members from 1970 onwards and is the result of the homonymous project started in by the European Commission in order to provide quantitative input to evaluate competitiveness, economic growth and policy interventions.

Empirical evidence on log-changes in TFP over the last 18 years reveals a fluctuating behaviour that would indicate ups and downs in the technical progress hardly understandable (see Figure 1 for Italy, Germany, Spain and the Netherlands).

Fig. 1 Log-ratio of ${}_{t-1}TFP_t$. Source: Our elaborations on EU-KLEMS data for the indicated countries.



4 Conclusions

Changes in empirical TFP do not always show the increasing pattern one would associate to the accumulation of technical progress, but for some countries seem rather fluctuating around a null average value. This is not consistent with the technical-progress-ratio-of-quantity-indexes equation exploited worldwide to measure TFP, but could be explained by other factors, for instance by the index used in the deflation process.

Taken the added value deflation technicalities as given and known (due to the widespread use of the chained quantity Laspeyres index), to gain a different point of view we have turned our attention to the input quantity indexes at the denominator, calculated through the Tornqvist index.

In an axiomatic framework, the Tornqvist index has been questioned (Balk, 2010) because it does not satisfy the factor reversal test neither the cofactor identity. In this paper we propose the use of the Sato-Vartia index, which encompasses both problems if the data necessary to construct it are available, as in the EU-KLEMS database.

Although the numerical result of the TFP index is the same with either Tornqvist or Sato-Vartia formula, the choice of the representation in terms of quantity or price indexes is not a pettifogging matter, because the construction of the index influences its intepretation and evaluation in terms of economic policy. For instance TFP indexes, both at the aggregate or sectoral levels are widely used as regressors in empirical economics but, as it comes to the data, shocks in productivity are, for some countries, all but evident. Our contribution provides a possible re-evaluation of this behaviour and suggests TFP index or its changes to be carefully used as explanatory variables.

References

Balk, B. M. (2010). An assumption-free framework for measuring productivity change. *Review of Income and Wealth 56*(s1).

Diewert, W. E. (1976). Exact and superlative index numbers. *Journal of econometrics 4*(2), 115–145.

Evans, W. D. and I. H. Siegel (1942). The meaning of productivity indexes. *Journal of the American Statistical Association 37*(217), 103–111.

Magdoff, H. (1939). The purpose and method of measuring productivity. *Journal of the American Statistical Association 34*(206), 309–318.

Martini, M. (2001). *Numeri indice per il confronto nel tempo e nello spazio*. CUSL.

OECD (2001). Measurement of aggregate and industry level productivity growth. Technical report.

Pasinetti, L. L. (1959). On concepts and measures of changes in productivity. *The Review of Economics and Statistics*, 270–286.

Solow, R. M. (1957). Technical change and the aggregate production function. *The review of Economics and Statistics*, 312–320.

Star, S. and R. E. Hall (1976). An approximate divisia index of total factor productivity. *Econometrica: Journal of the Econometric Society*, 257–263.

Monetary poverty indicators at local level: evaluating the impact of different poverty thresholds

Indicatori monetari di povertà a livello locale: studio dell'effetto di diverse soglie di povertà

Luigi Biggeri, Caterina Giusti and Stefano Marchetti[1]

Abstract The importance of poverty measures at sub-national level is widely attested, both for a detailed planning of policy actions, and for the citizens to evaluate their effects. However, there are relevant issues when computing sub-national poverty indicators that may impact their value: the definition of the poverty line for monetary poverty indicators, and the use of small area estimation methods when the sample size is not enough to obtain accurate estimates at local level. In this work, we estimate the poverty incidence at provincial level in Italy by using small area estimation methods, and we analyze the impact on the poverty incidence of different poverty lines, defined at national, regional and provincial level. The key results underline a strong impact on the poverty incidence when using sub-national poverty lines.

Abstract *Disporre di indicatori di povertà a livello locale è uno strumento indispensabile per i decisori politici, per poter pianificare opportune politiche di intervento, e per i cittadini, per poter valutare l'effetto delle politiche sul proprio territorio. La stima di indicatori di povertà monetaria relativi a livello locale pone due principali problemi metodologici: la definizione della linea di povertà e l'utilizzo di modelli di stima per piccole aree quando la ridotta numerosità campionaria a livello locale non consente di ottenere stime dirette accurate. Obiettivo di questo lavoro è confrontare l'incidenza della povertà nelle province italiane utilizzando linee di povertà alternative, definite a livello nazionale, regionale e provinciale, utilizzando modelli per piccole aree per le stime a livello provinciale.*

[1] Luigi Biggeri, Università degli Studi di Firenze; biggeri@disia.it
Caterina Giusti, Università di Pisa; caterina.giusti@unipi.it
Stefano Marchetti, Università di Pisa; stefano.marchetti@unipi.it

Key words: local poverty indicators; poverty lines; small area estimation.

1 Introduction

The important role played by poverty measures at sub-national and local level in setting policy actions against poverty and social exclusion is widely attested. Local poverty indicators are relevant both for a detailed planning of the policies actions and for the citizens to evaluate their effect.

As it is well known, a common method used to measure the monetary poverty is based on income or consumption levels. Individuals or families are considered as poor if their income or consumption level falls below a minimum level (called poverty line, PL). The PL is usually defined at the national level. However, it is well known that in Italy there is a strong geographical heterogeneity in income and consumption levels (ISTAT 2013a, ISTAT 2013b). Then, we propose an analysis where the geographical heterogeneity is reduced by defining the PL at regional and provincial level. Moreover, it is important to underline that also the different price levels within the country can play a role in the definition of alternative PLs, as observed among others by Ayala et al. (2014) and Giusti et al. (2017). However, we do not treat this last issue in this work.

Here we refer to the household poverty incidence or Head Count Ratio (HCR), the simplest monetary poverty indicator usually elaborated by most of the National Statistical Offices. More in detail, we use consumption expenditures data from the Italian Household Budget Survey (HBS) 2012 to estimate the HCR for the 20 regions and the 110 provinces in Italy.

We first estimate the regional HCRs using two alternative PLs: the PL defined at national level and the PLs defined at regional level. The national PL for households of two components is defined by ISTAT as the mean per-capita consumption expenditure at national level. Then, this PL is adjusted for households with a different number of components by using the Carbonaro equivalence scale, which takes the following values: 0.6 for households with one component, 1.33 with three, 1.63 with four, 1.90 with five, 2.16 with six and 2.40 for households with seven or more components. We define the PLs at regional level in the same way but computing the mean per-capita consumption expenditure separately for each region.

Since we observe a high impact of the regional PL definition on the regional HCRs, we then extend the analysis at the provincial level. When computing the HCR at provincial level, the PL can be defined not only at national or regional level, but also at provincial level. The 2012 HBS sample size at provincial level, varying from zero to 1037, with a median value of 146, is for most of the provinces too small to obtain reliable estimates both of the HCRs and of the PLs at provincial level.

Therefore, we use a small area model to obtain more accurate estimates, as better explain in the next sections.

2 Estimating regional HCRs with different poverty lines

The PL used in the computation of the HCR with expenditures data depends on the level of the mean per-capita consumption expenditures that in Italy varies strongly among regions, with a percentage difference that reaches the 50% comparing Northern with Southern Italian Regions. Therefore, it is important to evaluate the impact of the use of sub-national poverty lines in measuring the poverty incidence.

Fig. 1 shows the estimates of the household HCRs for the 20 Italian regions by using the national PL and the regional PLs. The results clearly show that the variability of the spatial distribution of the HCRs is quite smaller when using the regional PLs rather than the national one. Moreover, the HCRs of north-east and north-west regions increase when using regional PLs, the HCRs of the central regions remain more or less the same, while the HCRs of the Southern regions strongly decrease. Thus, the use of different PLs has strong geographical implications in the evaluation of Italian households' poverty. On the other hand, the choice of the poverty definition and of the PL depends on the level of analysis and the kind of the policy to be implemented (Kangas and Ritakallio, 2007). However, for comparing relative monetary poverty at regional (local) level, it seems justified the use of region-specific PL (Mogstad et al., 2007).

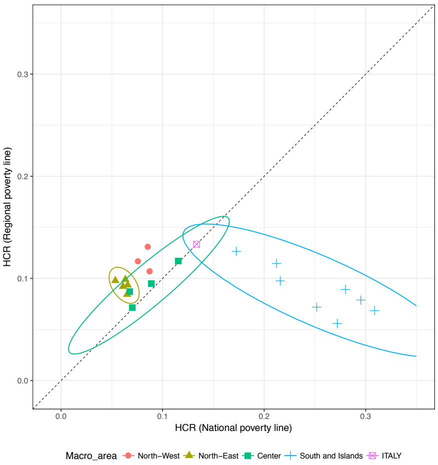


Figure 1: Estimates of the Head Count Ratio (HCR) for the 20 Italian regions with national versus regional poverty lines

3 Estimating the provincial HCR with different poverty lines

In this section we analyse the impact of three different thresholds on HCR estimates at provincial level in Italy. The HCR estimates at the province level are estimated using an area-level Fay-Herriot model (Fay and Herriot, 1979). This method uses aggregated auxiliary data to model direct estimates of the HCR to reduce their variability. As auxiliary variables at the province level we use the per-capita taxable income (information available from the “Agenzia delle entrate” database 2012) and the share of households who own their house (from the Population Census 2011). To save space we do not report here the model parameters, the model diagnostics and other details related to the estimation procedure. As a general result, the average decrease in variability of the HCR estimates is about 23.7% compared to the direct estimates.

We also estimate the provincial PLs by using a small area model, equal to the one used for provincial HCR estimates. Fig. 2 shows two plots: on the left the HCRs for the 110 Italian provinces based on the national PL versus the HCRs based on regional PLs, on the right the HCRs based on the national PL versus the HCRs based on provincial PLs. The two plots are similar: they show that when the PLs are estimated at more detailed geographical level we observe a strong decrease of the HCRs in the southern provinces, an increase for most of the north-east provinces, while the HCRs in central and north-west provinces increase in some provinces and decrease in others. Switching from regional to provincial PLs affect the HCRs more or less in the same way.

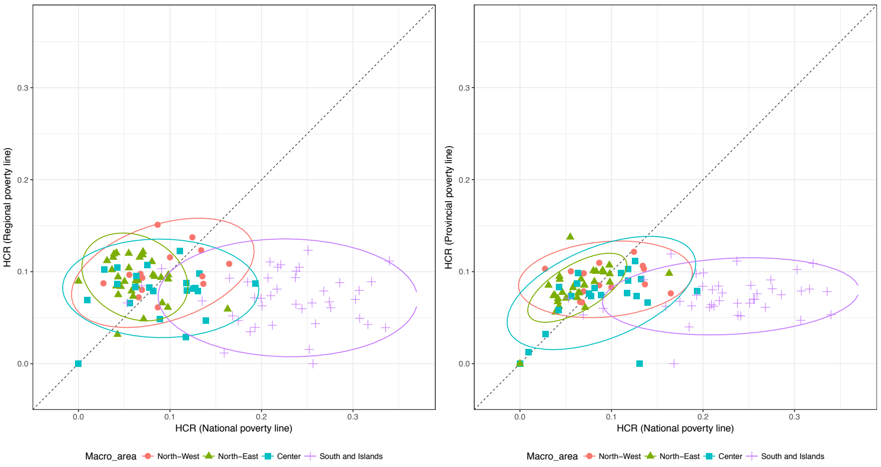


Figure 2: Estimates of the Head Count Ratio (HCR) for the 110 Italian provinces with national versus regional poverty lines (left) and with national versus provincial poverty lines (right).

Finally, Fig. 3 shows the HCRs based on regional vs. HCR based on provincial PLs. In this case it seems that the HCRs increase or decrease independently on the geographical level of estimation of the PLs. Moreover, the change in the values of the HCRs is very small compared to that observed when comparing results obtained using the national PL (Fig. 2).

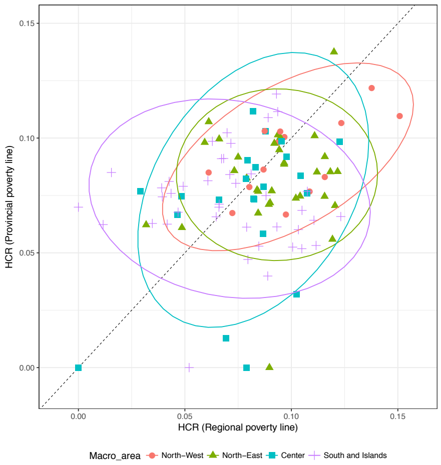


Figure 3: Estimates of the Head Count Ratio (HCR) for the 110 Italian provinces with regional versus provincial poverty lines

The main results presented in this section suggest that measuring the monetary poverty incidence at provincial level using national or local (regional or provincial) thresholds strongly change the picture of the phenomena. On the contrary, using provincial rather than regional PLs seems to not affect HCR estimates at the province level.

4 Concluding remarks

In this work we have presented alternative estimates of the HCR for Italian regions and provinces by using data on households' consumption expenditure. The aim was to evaluate the impact of subnational PLs on the HCRs. To estimate the HCRs and the PLs at provincial level we suggested the use of a small area model defined at the area level. Our results show that the choice of the PL is very relevant when the aim is to compare local relative poverty indicators.

The results can be extended in several directions, for example by also taking into consideration the different level of the prices in the regions and provinces, since also this aspect can highly impact the value of the HCRs. Moreover, to get a more complete picture of the poverty and living condition in the areas of interest, it would be important to complement the estimation of the HCRs with other monetary and non-monetary poverty indicators.

References

1. Ayala, L., Jurado, A., Perez-Mayo, J.: Drawing the Poverty Line: Do Regional Thresholds and Prices Make a Difference? Appl. Econ. Persp. Pol. (2014) doi:10.1093/aepp/ppt053
2. ISTAT (2013a) La povertà in Italia, Anno 2012. Statistiche report.
3. ISTAT (2013b) Reddito e condizioni di vita, Anno 2012. Statistiche report.
4. Giusti, C., Masserini L., Pratesi, M.: Local Comparisons of Small Area Estimates of Poverty: An Application Within the Tuscany Region in Italy. Social Indicators Research (2017), 131 (1): 235-254.
5. Kangas, O., Ritakallio, V.: Relative to What? Cross National Pictures of European Poverty Measured by Regional, National and European Standards. European Societies (2007), 9 (2): 119-145.
6. Mogstad, M., Langørgen, A., Aaberge, R.: Region-specific versus Country-specific Poverty Lines in Analysis of Poverty. Journal of Economic Inequality (2007), 5 (1): 115-122.

A gender inequality assessment by means of the Gini index decomposition

Una valutazione della disuguaglianza di genere attraverso la scomposizione dell'indice di Gini

Michele Costa

Abstract This paper proposes to measure and to evaluate gender gaps and gender inequalities by means of the decomposition of an inequality measure. A three-terms decomposition of the Gini index is applied, thus allowing to take into account also the role of overlapping between female and male subpopulations. An analysis of the income distribution of the Italian households shows how gender gaps represent a major source of inequality, without particular improvements over the last 20 years.
Abstract *In questo lavoro si propone di analizzare e di valutare i differenziali e la disuguaglianza di genere grazie alla scomposizione di una misura di disuguaglianza. Il ricorso ad una scomposizione dell'indice di Gini articolata su tre componenti permette di tenere conto anche della sovrapposizione tra le distribuzioni delle sottopopolazioni femminile e maschile. L'analisi della distribuzione del reddito familiare in Italia mostra che i differenziali di genere rappresentano un importante fattore di disuguaglianza, sostanzialmente stabile durante gli ultimi 20 anni.*

Key words: Gender gap, Gender income inequality, Inequality decomposition, Gini index

1 Introduction

Gender inequalities and gender gaps are a worldwide concern and represent the core of uncountable actions and policies developed by either governments and institutions. Gender inequalities are firstly a primary and fundamental issue of justice. Consequences of gender inequalities are frequently overlooked or underestimated, while it exists an interesting literature which analyzes the relation between gender inequality and welfare, pointing out gender gaps as a constraint for economic growth.

Michele Costa
Department of Economics, University of Bologna, e-mail: michele.costa@unibo.it

We assess the role of gender in income inequality by decomposing the Gini inequality ratio following the approach introduced by Dagum in 1997. First we evaluate the inequality within male and female subgroups, second we analyse the contribution to total inequality attributable to the differences between female and male subpopulations. The Dagum's Gini index decomposition also allows to evaluate the effect on overall inequality of the overlapping between female and male subpopulations, which represents a relevant element in gender inequality studies.

2 The Dagum's Gini index decomposition

The Gini index is one of the most important measure of inequality and, during its over 100 years of file, has experienced many different interpretations, expressions and formulas. For the case of a population disaggregated into k subgroups of size n_j, with $\sum_{j=0}^{k} n_j = n$, the Gini index G can be expressed as follows

$$G = \frac{1}{n\bar{y}^2} \sum_{j=1}^{k} \sum_{h=1}^{k} \sum_{i=1}^{n_j} \sum_{r=1}^{n_h} |y_{ji} - y_{hr}| \tag{1}$$

where $\bar{y}$ is the arithmetic mean of y in the overall population, y_{ji} is the value of y in the i-th unit of the j-th subgroup and, accordingly, y_{hr} is the value of y in the r-th unit of the h-th subgroup. For a detailed discussion of the Gini index see, e.g., [3],[7].

Among the many methods which allow to decompose the Gini index (see, e.g., [4],[8],[9]), we use the decomposition proposed by Dagum [5], where the differences $|y_{ji} - y_{hr}|$ in (1) are assigned to G_w, the component of inequality within subgroups, when $j = h$, to G_b, the component of inequality between subgroups, when $j \neq h$, $\bar{y}_j \geq \bar{y}_h$, $y_{ji} \geq y_{hr}$, and to G_t, the component of overlapping, when $j \neq h$, $\bar{y}_j \geq \bar{y}_h$, $y_{ji} < y_{hr}$.

The component of inequality within can be obtained quite esaily from the relation $G_w = \sum_{j=1}^{k} G_j p_j s_j$, where G_j is the Gini index of the j-th subgroup, while $p_j = n_j/n$ and $s_j = (n_j\bar{y}_j)/(n\bar{y})$ are the population share and the income share of the j-th subgroup, respectively.

For the other two components, G_b and G_t, which in the original version require some substantial computational effort, are available [1] simplified expressions, which are $G_b = G_b^* + 0.5(G - G_w - G_b^*)$ and $G_t = 0.5(G - G_w - G_b^*)$, where $G_b^* = \sum_{j=1}^{k-1} \sum_{h=1,j=k}^{k} \frac{p_{hj}^* - s_{hj}^*}{p_{hj}^* s_{jh}^* + p_{jh}^* s_{hj}^*} (p_j s_h + p_h s_j)$, $p_{hj}^* = p_h/(p_h + p_j)$ and $s_{hj}^* = s_h/(s_h + s_j)$.

In order to achieve a better understanding of the inequality structure, it is also possible to compare the decomposition obtained by using all n observations, that is $G_y = G_{wy} + G_{by} + G_{ty}$, to the decompositions obtained by referring only to subsamples of observations. In particular, it is useful to analyze the decompositions for the lower values of y, $G_{y|ymin} = G_{wy|ymin} + G_{by|ymin} + G_{ty|ymin}$ as well as for the higher values of y, $G_{y|ymax} = G_{wy|ymax} + G_{by|ymax} + G_{ty|ymax}$. When the structure of the decomposed indices $G_{y|ymin}$ and $G_{y|ymax}$ is similar, we get that the underlying inequality factor operates uniformly on y, while different structures indicate that particular regions of y are more affected by the inequality factor.

Following a similar approach, we can also evaluate the influence of a further inequality factor x by ranking y on the values of x and by comparing the decompositions of $G_{y|xmin}$ and $G_{y|xmax}$. Similar decompositions suggest that the x and the y are independent, while different decompositions indicate a relation between the two inequality factors.

A final element of interest refers to the evaluation of the inequality between, which is usually performed on the basis of the ratio G_w/G, where G acts as the maximum of G_b. The scenario $G_b = G$ implies $G_w = G_t = 0$: while $G_t = 0$, that is the absence of overlapping, doesn't present particular difficulties, the hypothesis $G_w = 0$, that is the equidistribution of y within each subgroup, represents a relevant departure from real situations. In order to achieve an evaluation of G_b more coherent with the observed data ([6], [2]), it is possible to keep $G_t = 0$ but to replace $G_w = 0$ with $G_w = G_{wmin}$, which is the minimum inequality within compatible with the observed data. In this case G_b is evaluated as $G_b/(G - G_{wmin})$.

3 The gender income inequality among Italian households

The Dagum's decomposition of the Gini index presented in Section 2 is extremely useful to analyze the relevance of gender in income inequality. The component G_w allows to evaluate how the income variability existing within the female and male subpopulations influence total inequality, while the contribution attributable to the differences between the female and male subpopulations is given by G_b and G_t. The meaning of G_b is straightforward, but as far as G_t it is useful to point out that high levels of overlapping indicate a small contribution of gender to income inequality, while low levels of overlapping suggest a stronger contribution.

The data used in this study are from the Survey on Households Income and Wealth of the Bank of Italy; the results illustrated in the following refer to the equivalent income obtained by means of the OCSE equivalence scale.Table 1 shows the p_i, s_i and G_i for the Italian households by gender of the head of the household: it is possible to observe some well known stylized facts of income inequality in Italy, that is the differences $(p_i - s_i)$. When $p_f = s_f$, the gender gap is equal to 0, while $p_f > s_f$ indicates the existence of a gender gap. The aggregate data of Table 1 suggest the presence of a gender gap, but also its reduction over time, since $(p_f - s_f)$ decreases from 3.8% in 1993, to 3.2% in 2004 and to 3% in 2014.

Moving from the aggregate and gross evaluation provided by $(p_f - s_f)$ to the more detailed and accurate information contained on the decomposed Gini index (Table 2), we obtain a different picture on gender income inequality. First, the importance of G_w on total inequality strongly decreases (from 62% in 1993 to 50% in 2014), thus indicating a weaker variability within the female and male subpopulations. Second, the overlapping between the female and male subpopulations in-

Table 1 Population share, income share and Gini index for the Italian households by gender of the head of the household

		1993			2004			2014	
	female	male	tot	female	male	tot	female	male	tot
p	0.275	0.725	1.000	0.388	0.612	1.000	0.471	0.529	1.00
s	0.237	0.763	1.000	0.356	0.644	1.000	0.441	0.559	1.00
G	0.319	0.334	0.333	0.307	0.336	0.327	0.311	0.324	0.320

creases: the importance of G_t rises from 13.5% in 1993 to 20.3% in 2014. A greater overlapping represents a positive signal for the reduction of the gender gap, since it suggests that the distributions of the subpopulations share a larger area. Third, the inequality between increases: the importance of G_b rises from 24.9% in 1993 to 29.7% in 2014. The relevance of the inequality between is fully understandable by comparing G_b to its maximum compatible with the observed data (last column of Table 2): in this case G_b represents 37.9% of total inequality in 1993, rising to 42.9% in 2014. Overall, the decrease of G_w is balanced by the increase of both G_b and G_t. While a greater G_t alleviates the role of gender as inequality factor, an increase of G_b leads to a stronger gender inequality from 1993 to 2014.

Table 2 Income inequality decomposition by gender of the head of the household

	Gw	Gb	Gt	Gw/G	Gb/G	Gt/G	Gb/(G-Gwmin)
1993	0.205	0.083	0.045	0.616	0.249	0.135	0.379
2004	0.175	0.092	0.060	0.535	0.281	0.183	0.407
2014	0.160	0.095	0.065	0.500	0.297	0.203	0.429

In order to better understand the results of Table 2, we focus on the tails of the distribution, taking into account the bottom and the top 20% of the income. Table 3 reports the p_i, s_i and G_i for the female and male subpopulations for the two cases and it is possible to observe some relevant differences.

By comparing the decomposed Gini indexes for the bottom and the top incomes (Table 4), we note that the two decompositions, initially quite different, are more or less similar in 2014. The importance of G_b shows a relevant increase, especially for the top incomes.

A further analysis of the gender income inequality refers to the study of specific population characteristics, such as educational level and geographical area of residence, chosen among the main inequality factors acknowledged by the literature. The Gini index decomposition is applied not to all n observations, but only to the subsample of households with the particular characteristic which we are analyzing. More specifically, we compare the female/male decompositions obtained on two subgroups related to two different values of the character under examination. When the two decompositions are substantially similar, the underlying factor is not rele-

Table 3 Population share, income share and Gini index for the Italian households by gender of the head of the household

		1993			2004			2014	
	female	male	tot	female	male	tot	female	male	tot
20% bottom income									
p	0.362	0.638	1.000	0.448	0.552	1.000	0.482	0.518	1.000
s	0.374	0.626	1.000	0.450	0.550	1.000	0.476	0.524	1.000
G	0.173	0.203	0.193	0.154	0.164	0.159	0.232	0.223	0.227
20% top income									
p	0.186	0.814	1.000	0.310	0.690	1.000	0.389	0.611	1.000
s	0.176	0.824	1.000	0.290	0.710	1.000	0.378	0.622	1.000
G	0.160	0.189	0.185	0.187	0.231	0.219	0.161	0.185	0.176
up to elementary school									
p	0.388	0.612	1.000	0.503	0.497	1.000	0.595	0.405	1.000
s	0.357	0.643	1.000	0.490	0.510	1.000	0.582	0.418	1.000
G	0.275	0.293	0.289	0.267	0.274	0.271	0.251	0.280	0.263
with university degree									
p	0.172	0.828	1.000	0.331	0.669	1.000	0.481	0.519	1.000
s	0.164	0.836	1.000	0.289	0.711	1.000	0.434	0.566	1.000
G	0.253	0.304	0.297	0.259	0.345	0.324	0.309	0.306	0.312
south islands									
p	0.247	0.753	1.000	0.407	0.593	1.000	0.475	0.525	1.000
s	0.221	0.779	1.000	0.382	0.618	1.000	0.440	0.560	1.000
G	0.312	0.351	0.344	0.292	0.320	0.310	0.312	0.341	0.330
north									
p	0.298	0.702	1.000	0.368	0.632	1.000	0.456	0.544	1.000
s	0.247	0.753	1.000	0.338	0.662	1.000	0.422	0.578	1.000
G	0.298	0.295	0.302	0.278	0.310	0.301	0.267	0.293	0.284

vant for the interpretation of the gender inequality, while, on the contrary, different decompositions indicate an influence on gender inequality. Table 3 illustrates the p_i, s_i and G_i for two subgroups: for the educational level we compare the up-to-elementary-school group to the group with a university degree, for the geographical area the group living in the north to the group living in the south or islands.

The related decompositions of the Gini index for the analysis of the gender gap are shown in Table 4. The comparison between the decompositions suggests that the educational level influences the gender income inequality more than the geographical area. We also confirm the decrease of the importance of G_w, together with an increase of the relevance of G_t and G_b, especially for the more affluent subgroups.

4 Conclusions

The decomposition of an inequality index can be extremely useful into the study of the gender income inequality, where the decomposition refers to the female and male subpopulations. The analysis of the income distribution of the Italian house-

Table 4 Income inequality decomposition by gender of the head of the household

	Gw/G	Gb/G	Gt/G	Gw/G	Gb/G	Gt/G
	20% bottom income			20% top income		
1993	0.544	0.259	0.197	0.717	0.168	0.114
2004	0.509	0.252	0.239	0.591	0.250	0.159
2014	0.500	0.263	0.237	0.534	0.267	0.199
	up to elementary school			with university degree		
1993	0.531	0.288	0.181	0.734	0.145	0.121
2004	0.498	0.273	0.229	0.583	0.272	0.145
2014	0.510	0.270	0.221	0.495	0.328	0.177
	south islands			north		
1993	0.648	0.212	0.140	0.734	0.145	0.121
2004	0.526	0.277	0.197	0.547	0.277	0.177
2014	0.500	0.303	0.197	0.505	0.307	0.187

holds shows how gender gap explained 24.9% of total inequality in 1993, rising to 29.7% in 2014. The scenario is even worse when evaluating inequality attributable to the differences between female and male subpopulations without the traditional assumption of null inequality within: in this case gender gaps are accountable for 37.9% of total inequality in 1993, rising to 42.9% in 2014. Inequality decomposition also allows to evaluate the relation between gender and other inequality factors: educational level of the head of the household and geographical area of residence are taken into account, with the former showing a greater influence on gender income inequality.

References

1. Costa, M.: Transvariation and inequality between subpopulations in the Dagum's Gini index decomposition. Metron, **67**, 134–120 (2009)
2. Costa, M.: The evaluation of the inequality between population subgroups. In Verde, R., Racioppi, F., Petrucci, A.: Statistics and Data Science: new challenges, new generations. Firenze University Press, Florence, 313-318 (2017)
3. Dagum, C.: Gini ratio. In: The New Palgrave Dictionary of Economics. Mac Millian Press, London (1987)
4. Dagum, C., Zenga M.: Income and Wealth Distribution, Inequality and Poverty. Springer, Berlin (1990)
5. Dagum, C.: A new decomposition of the Gini income inequality ratio. Empirical Economics, **22**, 515–531 (1997)
6. Elbers, C., Lanjouw, P., Mistiaen J.A., Ozler, B.: Reinterpreting Between-Group Inequality. Journal of Economic Inequality, **6**, 231–245 (2008)
7. Giorgi, G.M.: Gini's scientific work: an evergreen. Metron, **63**, 299–315 (2005)
8. Giorgi, G.M.: The Gini inequality index decomposition. An evolutionary study. In Deutsch, J., Silber, J.: The measurement of individual well-being and group inequalities. Routledge, London (2011)
9. Yitzhaki, S., Lerman, R.: Income stratification and income inequality. Review of Income and Wealth, **37**, 313–329 (1991)

Socio-Economic Statistics

The NEETs during the economic crisis in Italy, Young NEETs in Italy, Spain and Greece during the economic crisis

I giovani NEET negli anni della crisi economica in Italia, Spagna e Grecia

Giovanni De Luca, Paolo Mazzocchi, Claudio Quintano, Antonella Rocca

Abstract The recent economic crisis has further increased the socio-economic disparities across European countries. In particular, the levels of NEETS (young people Not in Employment, Education and Training) has resulted greater than before in European Mediterranean countries, more severely hit by the crisis. The aim of this paper is to analyze the peculiarities of NEETs in Italy, Spain and Greece in the years from 2007 to 2016 and verify if the introduction of the Youth Guarantee Schemes has contributed to decrease these high shares. To this aim, some econometric models have been applied to the Labour Force Survey data and Youth Guarantee Schemes effects have been tested using the Chow test.

Abstract *La recente crisi economica ha ulteriormente inasprito le differenze socio-economiche tra i paesi europei. In particolare, nei paesi del Sud Europa, maggiormente colpiti dalla crisi, si è assistito ad una crescita del livello dei giovani NEET (giovani che non lavorano, né studiano, né seguono corsi di formazione). Scopo di questo lavoro è analizzare le peculiarità di tale fenomeno in Italia, Spagna e Grecia negli anni dal 2007 al 2016 e verificare se l'introduzione del Fondo Garanzia Giovani ha prodotto una riduzione significativa dei livelli. A tale scopo, alcuni modelli econometrici sono stati applicati ai dati della Rilevazione continua sulle Forze Lavoro, mentre gli effetti del Fondo Garanzia Giovani sono stati verificati attraverso l'applicazione del test di Chow.*

Key words: NEETs, Youth unemployment rate, Early school leavers

[1] Giovanni De Luca, giovanni.deluca@uniparthenope.it; Paolo Mazzocchi, paolo.mazzocchi@uniparthenope.it; Claudio Quintano, claudio.quintano@uniparthenope.it; Antonella Rocca, rocca@uniparthenope.it; Department of Management and Quantitative Studies, University of Naples "Parthenope", Naples, Italy.

1 Introduction

The late 2000s crisis and the increased complexity of the labour markets **have** exacerbated the social and economic inequalities almost *in* all European countries and *among* European countries. Countries whose economies already suffered of structural weakness were hit more hardly by the economic crisis than the most prosperous ones. In particular, the South of Europe is one of the regions in the world where the consequences of the crisis have become most salient. Furthermore, within each country the most vulnerable segments of population, such as young people and migrants, resulted more penalized in terms of job losses and worsening in their condition on the labour market.

In this paper the authors have analyzed and compared the changes occurred in the levels of NEETs (young people Not in Employment, Education and Training, aged 15-29 years) in the years of economic crisis, from 2007 to 2016, in Italy, Spain and Greece in a gender perspective, in order to highlight their main peculiarities. Furthermore, the authors have investigated if the overcoming of the deepest phase of economic crisis has led to significant reductions in the share of NEETs in light also of the recent labour market reforms. The reference is to the introduction of the Youth Guarantee Found, a recent European reform finalized to create opportunities to exit from the NEET status, ensuring that all young people receive a good-quality offer of job, apprenticeship, traineeship, or continued education (European Commission, 2017).

The structure of the paper is as follows: Section 2 introduces the framework of analysis; Section 3 shows the methodology while in Section 4 the main results and some conclusive considerations are reported.

2 The framework of analysis

During the economic crisis of the last decade, some European countries, such as Italy, Spain and Greece experienced one of the great recessions of the last century; contrariwise, other economies in Europe only slowed their growth (Germany and Poland, for example). Indeed, comparing the data in 2007 and 2016 – before and after that the economic crisis produced its effect (Eurostat on-line database, http://ec.europa.eu/eurostat/web/lfs/data/database) – while at EU-28 country level total unemployment (age class 25-74) registered only a slight increase, moving from 6.1% to 7.5%, in Italy, Spain and Greece, it suffered an increase of 5.1, 10.9 and 15 percentages points, respectively. The worsening was still greater and the gap higher for the youth unemployment rate, that in these years at EU-28 level increased only of 2.1% while for Italy, Spain and Greece, the increase was of 17.4, 26.3 and 24.6 percentage points respectively (passing from 20.4% to 37.8% in Italy, from 18.1% to 44.4% in Spain and from 22.7% to 47.3% in Greece). In the same period, in these countries also the share of NEETs remarkably increased. Indeed, while at European level it resulted almost stationary, passing from 11% in 2007 to 11.5% in 2016, in

Italy, Spain, and Greece it increased of 3.7, 4.5, and 2.6 percentage points respectively. Indeed, it moved from 16.1% to 19.8% in Italy, from 12.0% to 14.6% in Spain and from 11.3% to 15.8% in Greece. Therefore, while Spain, but especially Greece, resulted more penalized in terms of youth unemployment rates, Italy suffered the highest levels of NEETs.

Young people represent one of the most vulnerable segments of population, because of their less work experience, their weaker work contracts and of a labour market that primarily tends to protect older workers. Besides unemployment, other causes hide behind the high levels of NEETs and concern inactivity. While unemployment is mainly correlated with the economic cycle, inactivity represents an alarming social phenomenon, as it is linked to the exclusion of young people from education and training (Bell and Blanchflower, 2011) and from the labour market because these young people do not take any action, remaining outside the labour force. The permanence in the NEET status for a prolonged period produces serious consequences on the future life perspectives because it predisposes to long term unemployment and social exclusion (OECD, 2014; Balan, 2015). For these reasons, the European Commission, solicited by the European Parliament, implemented and financed the Youth Guarantee Fund, successfully already rolled out in Scandinavian countries (Escudero and Lopez Mourelo, 2015). It consists in a number of plans to target youth unemployment, including apprenticeships and traineeships programmes and support schemes for young business starters. Beneficiaries are all European regions with a youth unemployment rate higher than 25%. All Greek and Spanish regions were eligible for the Youth Employment initiative while in Italy the regions of Veneto and Trentino-Alto Adige should have been excluded; anyway, it was chosen to extend the scope to the Belluno, Rovigo and Venice provinces, because their youth unemployment rates were higher than 25% (www.garanziagiovani.gov.it). Even if the Youth Guarantee implementation plans were submitted in December 2013, they started to produce the first effects in the Spring of 2014. The introduction of these plans and the contextual overcoming of the deepest recession phase produced a reduction in both the unemployment and NEET rates everywhere. Therefore, in this paper, besides studying the peculiarities and determinants of NEETs across Italy, Greece and Spain, the authors aim at verifying if in the Spring of 2014 a structural break in the NEET trend occurred or not.

Data come from the Labour Force Survey (LFS), currently the main European reference source for comparable and multidimensional socio-economic statistics on employees and working conditions, waves from 2007 to 2016. The need to get an adequate number of observations and to detect changes occurred in short time intervals **have** motivated the construction (through ad hoc elaborations) of four-weekly time intervals. The LFS is in fact a continuous survey carried out during every week of the year. In order to account for the Youth Guarantee attendance, the analysis involves only young people from 15 to 24 years, the only implicated in these initiatives (in Spain the eligibility is extended to disables under-30s).

3 The statistical methodology

The analysis of the dynamics of the share of NEETs is based on a Seasonal Autoregressive Distributed Lag (SARDL) model, in which the dependent variable – the share of NEETs on the 15-24 years population – is studied according to some own lags (in order to consider the persistence of the NEET phenomenon over time and its seasonal component) and also to some independent variables (Almon, 1965). The model is given by

$$y_t = \mu + \sum_{k=1}^{p} \phi_k y_{t-k} + \sum_{K=1}^{P} \Phi_K y_{t-Ks} + \sum_{s=0}^{q} \beta_s x_{t-s} + \boldsymbol{\gamma}' \mathbf{z}_t + u_t$$

where y_t is the share of NEETs at time t, y_{t-k} are the lagged values; the second sum includes seasonal lags, x_{t-s} is the independent variable for which also some lags are considered and z_t is the r x 1 vector of additional not-delayed variables. Finally, ϕ_k, Φ_K, β_s and $\boldsymbol{\gamma}$ the corresponding parameters. These models are particularly easy to implement because the OLS estimates are appropriate.

In a second step, in order to study if the introduction of the Youth Guarantee Fund has produced a significant variation in the share of NEETs, we **have** define the dummy variable

$$g_t = \begin{cases} 0 & t \in T_b \\ 1 & t \in T_a \end{cases}$$

where T_b is the set of times before the introduction of the Youth Guarantee Fund and T_a is the set of times after. The Chow test has been used to compare the results of the model applied to all the data – defined as restricted model – with the results of the models separately run for the two periods (Chow, 1960). The unrestricted model can be then written as a two regime SARDL model,

$$y_t = \begin{cases} \mu_1 + \sum_{k=1}^{p} \phi_{1k} y_{t-k} + \sum_{K=1}^{P} \Phi_{1K} y_{t-Ks} + \sum_{s=0}^{q} \beta_{1s} x_{t-s} + \boldsymbol{\gamma}_1{}' \mathbf{z}_t + u_t & \text{if} \quad g_t = 0 \\ \mu_2 + \sum_{k=1}^{p} \phi_{2k}\, y_{t-k} + \sum_{K=1}^{P} \Phi_{2K} y_{t-Ks} + \sum_{s=0}^{q} \beta^{\bullet}_{2s} x_{t-s} + \boldsymbol{\gamma}_2{}' \mathbf{z}_t + u_t & \text{if} \quad g_t = 1 \end{cases} \}$$

The null hypothesis of the Chow test is the stability of the parameters across the two regimes, that is: H_0: $\theta_1 = \theta_2$ where θ_1 and θ_2 are the parameter vectors in the first and second regime, respectively. The Chow statistics is given by

$$F = \frac{(RRSS - URSS)/j}{URSS/(N_1 + N_2 - k)} \sim F_{j, N_1 + N_2 - k}$$

where RRSS is the restricted residuals sum of squares obtained estimating a unique model for all the period, while URSS denotes the unrestricted residuals sum of squares obtained as sum of the RSS of two models, the former considering the N_1 $\mathbf{N_1}$periods before the introduction of the Youth Guarantee Fund, the latter considering the N_2 $\mathbf{N_2}$periods after. Finally, j is the number of restrictions and k is the number of parameters of the unrestricted model.

Under the null hypothesis, the F-ratio follows the F distribution with j and (N_1+N_2-k) degrees of freedom in the numerator and denominator, respectively. However, the Chow test does not inform us whether the difference in the two models is due to the differences in the constant term, in a single coefficient, in all the coefficients of a variable regardless the lags, or finally in a specific subset of coefficients.

In particular, we are interested in checking if the Youth Guarantee Fund has involved a reduction in the constant term without any impact on the dependence structure with the other variable, which imply a decrease in the share of NEETs. In this case, the unrestricted model is

$$y_t = \mu_1(1 - g_t) + \mu_2 g_t + \sum_{k=1}^{p} \phi_k y_{t-k} + \sum_{K=1}^{P} \Phi_K y_{t-Ks} + \sum_{s=0}^{q} \beta_s x_{t-s} + \gamma' z_t + u_t$$

where μ_1 is the constant in the first regime, observed before the introduction of the Youth Guarantee Fund, and μ_2 is the constant of the second regime, characterized by its presence, and the hypothesis to test is H_0: $\mu_1 = \mu_2$.

When the Chow test is computed on a subset of coefficients, then the null hypothesis is: H_0: $\theta_{1V} = \theta_{2V}$, where θ_{1V} and θ_{2V} are the parameter vectors of the variable V in the first and second regime, respectively.

4 Results

Among both males and females, the incidence of NEETs has been higher in Italy, where, also in 2016, it was around the 20%[1]. Contrariwise, the unemployment rates, both with reference to the whole population and only to young people, were higher in Greece and Spain than in Italy. However, in Italy young people resulted more penalized on the labour market in comparison to the older people. The ratio between the youth and the total unemployment rates is indeed higher for Italy. While in Greece and Spain the high levels of NEETs result mainly linked to unemployment, Italy shows the higher share of inactive NEETs. Therefore, in Italy the causes of the NEET phenomenon are harder to find. According to education, Spain highlights the highest rates of early school leavers, but also of high-educated young people[2].

The SARDL models applied to the subsamples of young people identified in each country according to gender highlight the high persistence of the phenomenon over time and the strictly dependence from the unemployment rate (with the exception of Italian men), that for Spain and Italian women includes also some delays (Tab.1). According to education, the share of early school leavers is always significant, with the exception of Greek men. In Spain and for Italian and Greek women the annual moving average of early school leavers results significant too.

Finally, according to the Chow test, a significant change in the dynamics of NEET levels can be observed in Italy, where it is mainly due to the increase in the educational levels, and for Greek men, where, instead, it depends also on a structural break in the NEETs, unemployment and educational levels.

Therefore, since the introduction of the Youth Guarantee Fund some effects are evident, especially for young Italian people and for Greek men. Furthermore, even if

[1] The descriptive and graphical analyses are here not reported for sake of brevity, but they are available on request by authors.

[2] Anyway, results are affected by the differences in the education system of the 3 countries compared (education is compulsory until 15 years of age in Greece while in Spain and Italy until 16 years; tertiary education should be attained at 22 years of age in Spain and at 24 years of age in Greece and Italy).

in Italy the major contribution to this result comes from the reduction on the share of early school leavers, the lack of statistical significance for the test on the unemployment covariates should suggest that this Scheme has also produced a positive effect on the inactive NEET component. Therefore, in the Italian context, results suggest to deepen the analysis on the extent to which the reduction of early school leavers is linked to the Youth Guarantee Scheme or, otherwise, to the general increase in the levels of education.

Table 1: *Seasonal Autoregressive Distributed Lag Models for NEETs by gender (2007- 2016).*

Neets	***Italy***		***Spain***		***Greece***	
	Men	***Women***	***Men***	***Women***	***Men***	***Women***
Neets$_{(t-1)}$	0.343^{***}	0.218^{***}	0.142^{**}	0.160^{**}	0.142	0.194^{**}
Neets$_{(t-2)}$	0.186^{**}	0.056	0.121^{*}	0.015	0.120	$-.0159^{*}$
Neets$_{(t-3)}$	0.165^{**}	0.167^{**}	-0.032	0.051	0.274^{***}	0.197^{**}
Neets$_{(t-13)}$	0.384^{***}	0.370^{***}	0.135^{**}	0.242^{***}	-	0.414^{***}
Unempl$_{(t)}$	-	0.133	0.414^{***}	0.039	0.301^{***}	0.206^{***}
Unempl$_{(t-1)}$	-	-0.298	-0.293^{*}	-0.270^{*}	-	-
Unempl$_{(t-2)}$	-	0.390^{***}	0.320^{**}	-0.136	-	-
Unempl$_{(t-3)}$	-	-	-	0.506^{***}	-	-
Early$_{(t)}$	0.531^{***}	0.766^{***}	0.668^{***}	0.650^{***}	-0.022	0.394^{***}
Early$_{(m12)}$	-0.220	-0.408^{**}	-0.261^{***}	-0.274^{*}	0.270	0.730^{***}
Constant	-6.193^{*}	-2.417	-7.272^{***}	-0.213	-1.394	-5.124^{***}
F-test	94.68^{***}	36.18^{***}	24.65^{***}	16.25^{***}	122.72^{***}	39.118^{***}
R^2 Adj.	0.829	0.732	0.647	0.568	0.862	0.697
H_0: $\theta_1 = \theta_2$	1.690^{*}	2.130^{**}	1.028	1.566	2.788^{***}	0.969
H_0: $\mu_1 = \mu_2$	1.964	3.529^{*}	3.502^{*}	0.123	7.672^{***}	0.589
H_0: $\theta_{1N} = \theta_{2N}$	1.162	1.103	1.620	0.869	2.225^{*}	1.119
H_0: $\theta_{1u} = \theta_{2u}$	-	1.551	1.713	1.158	5.670^{***}	0.292
H_0: $\theta_{1E} = \theta_{2E}$	2.748^{**}	3.714^{***}	1.549	0.717	4.322^{***}	0.443

*** Significant at 1%; ** Significant at 5%; * Significant at 10%.

References

1. Almon, S.: The distributed lag between capital appropriations and expenditures. Econometrica 33, pp.178–196 (1965).
2. Balan, M.: Methods to estimate the structure and size of the "neet" youth. Procedia Econ. and Financ., 32, pp. 119 – 124 (2015).
3. Bell, D.N.F.: Blanchflower, D.G.. Young people and the Great Recession. Oxford Rev. of Econ. Policy, 27(2), pp. 241-267 (2011).
4. Chow, G.C.: Test of equality between sets of coefficients in two linear regressions. Econometrica, 28(3), pp. 591-605. http://dx.doi.org/10.2307/1910133} (1960).
5. Escudero, V., Lopez M., E.: The Youth Guarantee programme in Europe: Features, implementation and challenges. International Labour Office, Working Paper, 4, (2015).
6. European Commission: Data collection for monitoring of Youth Guarantee schemes: 2015. Employment, Social Affairs & Inclusion. January (2017).
7. OECD: Education at a Glance 2014. OECD Indicators (2014).

Camel or dromedary? A study of the equilibrium distribution of income in the EU countries.

Cammello o dromedario? Un'analisi della distribuzione di equilibrio del reddito nei paesi dell'UE.

Crosato L., Ferretti C., Ganugi P.

Abstract We face here the problem of analysing the presence of bimodality of the equilibrium distribution of incomes in the EU countries, using EU-SILC data about 2012-2015. As a first step we visually inspect the kernel distribution and calculate the Sarle's bimodality coefficient. We evaluate also the relationship between bimodality and inequality. As a second step we propose to use some suitable stochastic models to analyse the shape (camel/dromedary) of the estimated the long-run income distribution. The chosen models are the classical Markov Chain and the Mover Stayer model.

Abstract *Questo lavoro affronta il problema di valutare la presenza o meno della bimodalità nella distribuzione di equilibrio dei redditi dei Paesi europei, per mezzo dei dati EU-SILC relativi agli anni 2012-2015. Come primo passo, analizziamo le stime kernel delle distribuzioni dei redditi e calcoliamo il coefficiente di bimodalità di Sarle. Come secondo passo, proponiamo di utilizzare alcuni processi stocastici per analizzare la forma (cammello/dromedario) della distribuzione dei redditi stimata sul lungo periodo. I modelli scelti sono la Catena di Markov classica e il modello Mover-Stayer.*

Key words: long-run income distribution, equilibrium, bimodality, bipolarization, inequality.

Lisa Crosato, Dip. di Economia, Metodi quantitative e Strategie di impresa, Univ. degli Studi Milano-Bicocca; email: lisa.crosato@unimib.it

Camilla Ferretti, Dip. di Scienze Economiche e Sociali, Univ. Cattolica del Sacro Cuore di Piacenza; email: camilla.ferretti@unicatt.it

Piero Ganugi, Dip. Di Ingegneria e Architettura, Univ. degli Studi di Parma; email: piero.ganugi@unipr.it

1 Introduction: Why is Bimodality relevant?

Investigating the presence of bimodality in the equilibrium distribution of income (i.e. the long-run distribution such that frequencies among income classes have achieved stability w.r.o. time) in EU countries is relevant both for political and theoretical reasons.

The political relevance of the topic depends directly from the fact that the achievement of the Welfare Systems after the Second World War in different decades within the European Countries has always found the decisive support in the Middle-Income Population. Given the crucial role of this part of Population in the formation of consensus, its eventual structural reduction in favor of Low and High Incomes introduces the premises of a breakdown in the consensus toward the same System and, in turn, toward the political parties which defend it. The intricacy of the problem is empowered by the fact that an eventual structural bimodality is entirely a different fact from a return to the XIX century Income Distribution with respectively predominant and scanty masses in the Low Incomes and in the tail of the distribution. On the contrary, bimodality involves “polarization” i.e. not only large mass for the working poors but also, even if far more curbed, for the rich employees.

The theoretical reason to study the eventual structural (long period) character of bimodality sources by the fact that just today in some European countries Income Distribution seems to be featured by a noticeable second mode (cfr. Fig. 1). It is then natural to work to ascertain if the same second peak is the result of temporary and now seemingly fading recession and so fated to be reabsorbed by a new period of (equalizing) growth.

Literature on bimodality of Incomes can be divided in two distinctive branches depending on the nature of the data. One of these two branches on Incomes has used the pro capita GDP at World or European level and has aimed to ascertain the eventual polarization of pro capita GDP in the World (Bourguignon and Morrisson, 2002; Pittau, 2005). The second branch of literature on Incomes uses Personal Income micro-data and aims to ascertain bimodality within the single country. Within this second field of research, the problem of measuring bimodality, or bipolarization, has been deepened among others by Chakravarty et al. (2007), Chakravarty and D'Ambrosio (2010), Lasso de la Vega et al. (201), Deutsch et al. (2013).

Our paper contributes to the second strand of the literature, building on the empirical evidence on the bimodality of income distribution in a few European Countries detected using the EU-SILC database. Our goal is to verify whether Income distributions display a “camel” or a “dromedary” shape and if they preserve such shape with respect of time and in the long-run period.

The paper is structured as follows: section 2 illustrates the data source and the bimodality of income distribution, section 3 applies the kernel density estimation and Sarle’s coefficient to evaluate the presence of bimodality and to have a look on its relationship with the Gini index, section 4 illustrates the expected results and section 5 concludes.

2 Data Description and evidence of bimodality in EU countries.

Our analysis is based on the data obtained from the European Union Statistics on Income and Living Conditions (EU-SILC) longitudinal study (see Krell et al., 2017, for a full description of the dataset and an analysis of the consistency of the data). EU-SILC has become the EU reference source for statistics on income distribution and social exclusion at European level and supplies, among other variables, the Net employee cash or near cash income (variable PY010N). According to the EU-SILC description, employee income is “the total remuneration, in cash or in kind, payable by an employer to an employee in return for work done by the latter during the income reference period”. The net income component is then given by the gross income component but for the tax at source, the social insurance contributions, or both, which are deducted (see the methodological guidelines and description of EU-SILC target variables).

We start by analyzing the net employee income for the last available year, i.e. 2015 income extracted from the longitudinal component of the database, because further on we will need to build transition matrices for the calculation of the equilibrium distribution. Figure 1 reports kernel densities estimates for the countries that, in 2012 and 2015 have at least 1,000 non-NAs and non-null observations for the variable at hand.

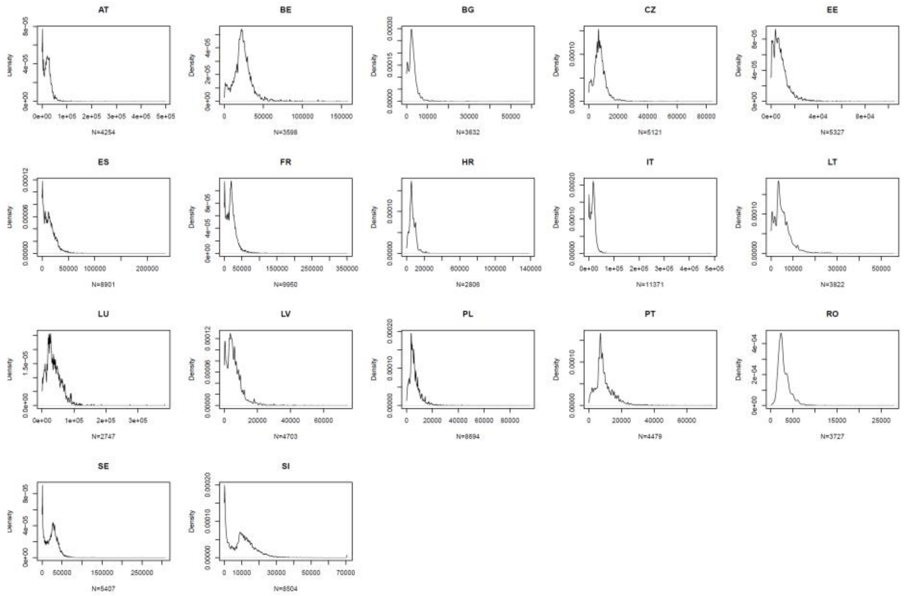


Figure 1: Kernel density estimates for 17 EU countries. Source: Author’s elaboration on EU-SILC LONGITUDINAL UDB 2015 – version 1 of March 2017. Variable PY10N, Net employee cash or near cash income greater than zero, countries with at least 1,000 non-null values only.

As can be seen, bimodality can be clearly detected in several countries. The presence of bimodality can be also evaluated through the Sarle’s bimodality coefficient (see

Ellison, 1987), whose corrected version for finite samples is given by the following formula:

$$b = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

where *n* is the sample dimension, *g* is the sample skewness and *k* is the sample excess kurtosis. In this case however the index neatly confirms bimodality only for Belgium, which is not consistent with the kernel densities of figure 1, suggesting the case for a further assessment of the extent to which bimodality takes place in the different countries.

The coefficient *b* seems positively related with the Gini index when considering the former communist countries (see figure 2), and, somewhat surprisingly, Belgium. The remaining countries instead show a not so clear relation.

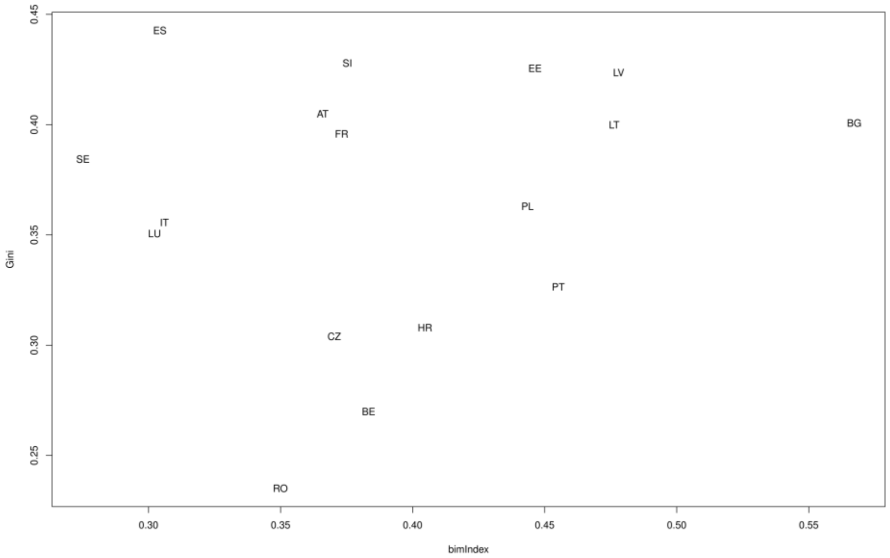


Figure 2: Scatterplot of the Gini concentration ratio (y-axis) versus the Sarle's bimodality index (x-axis) for 27 EU countries. Source: Author's elaboration on EU-SILC LONGITUDINAL UDB 2015 – version 1 of March 2017. Variable PY10N, Net employee cash or near cash income greater than zero, countries with at least 1,000 non-null values only.

3 Our proposal for evaluating the long-run two peaks distribution

Having ascertained the presence of bimodality in a relevant group of EU countries in the years 2012-2015, we tackle the problem to evaluate the long-run behavior of EU net incomes. We propose to estimate it using two stochastic models, Markov Chain

(MC) and Mover Stayer (MS) (see Anderson and Goodman, 1957, and Goodman, 1961). Both are based on the empirical yearly transition matrices, which summarize the probability to move among a set of suitably defined income classes. Such classes should be able to mirror the division among Low-, Middle- and High-Income individuals, as for example in Bourguignon (2002) or more recently in Xuehui and Shang-Jin (2017). The main difference is that the MS model supposes the existence of a bulk of individuals never moving from their starting state. In detail, let $S = diag\{s_i\}$ be the diagonal matrix where s_i denotes the probability that a EU citizen with a given income in 2012 is a Stayer and that consequently he/she will never leave from the corresponding income class. The MS global one-step transition matrix is given by the formula:

$$P = S + (I - S) * M,$$

where I is the identity matrix, and M is the transition matrix for the not-Stayers (Movers), which are supposed to move following a classical Markov Chain ruled exactly by M. Given the starting distribution p_0 (in this case coinciding with the percentage of citizens that in 2012 belongs to each income class), the long-run distribution is given by the equilibrium distribution π evaluated as:

$$\pi = p_0 * \left(\lim_{t\to+\infty} P^{(t)}\right),$$

where $P^{(t)}$ is the global t-steps transition matrix given by $P^{(t)} = S + (I - S) * M^{(t)},$ When $s_i = 0$ for every income class, the previous formula coincides with the classical MC equilibrium distribution. Parameters of both MC and MS can be estimated using the techniques proposed in Anderson and Goodman (1957) and Frydman et al. (1985).

4 Expected results

The long-run distribution gives a glance on what happen to the income distributions if the 2012-2015 economic conditions remain stable also in the future. We can expect only two possible results: 1) a camel or dromedary distribution maintains its starting shape, or 2) a camel/dromedary distribution tends respectively to lose/gain one “humpback”. If also the equilibrium distribution remains or becomes a two-peaks one, we can claim that bipolarization has become chronic in some EEC countries, and the income is neatly divided between Low- and High-Income individuals.

Bipolarization and Gini indices have to be calculated also on the equilibrium distributions and their values have to be compared to the values obtained on the initial income distributions p_0. It is in fact intriguing to ascertain if bipolarization involves also a rise in inequality. With the aim to obtain a more robust analysis we will calculated different bipolarization indices as proposed in Chakravarty et al. (2007), Chakravarty and D'Ambrosio (2010), Lasso de la Vega et al. (2010) and Deutsch et al. (2013).

5 Further research

Further research will regard mainly two aspects:

1) the estimation of the long-run distribution through more complex models, as for example a mixture of Markov Chains in which individuals are characterized by different speeds, such as in Frydman et al. (2002);

2) the modelization of the bimodal density distribution through an analytical bimodal density distribution as proposed in Ferretti et al. (2017).

References

1. Anderson, T. and Goodman, L. (1957). Statistical Inference about Markov Chains. *Ann. Math. Statist.*, **28**(1), 89–110.
2. Bourguignon F, Morrisson C (2002). Inequality among World Citizens. *Am. Econ. Re*v., **92**(4):727–744.
3. Chakravarty, S. R., D'Ambrosio C. (2010). Polarization Orderings of Income Distributions, Review of Income and Wealth **56**(1): 47-64.
4. Chakravarty, S., Majumder A., Roy S. (2007). A Treatment of Absolute Indices of Polarization, *The Japanese Economic Review* **58**(2): 273-293.
5. Deutsch J, Fusco A., Silber J. (2013). The BIP Trilogy (Bipolarization, Inequality and Polarization): One Saga but Three Different Stories. *Economics: The Open-Access, Open-Assessment E-Journal*, **7**(2013-22): 1–33.
6. Ellison, A.M. (1987). Effect of seed dimorphism on the density-dependent dynamics of experimental population of *Atriplex triangularis* (Chenopodiaceae). *Am. J. Bot.*, **74**(8), 1280-1288.
7. Ferretti C., Ganugi P., Zammori F. (2017), Change of Variables theorem to fit Bimodal Distributions, in: *SIS 2017 Statistical and Data Sciences: new challenges, new generations*, Conference Proceeding, 417-422.
8. Frydman, H., & Kadam, A. (2002). Estimation in the continuous time Mover Stayer model with an application to bond rating migration. *App. Stoch. Mod. in Business and Industry,* **20**, 155–170.
9. Frydman, H., Kallberg, J., and Kao, D. (1985). Testing the adequacy of Markov Chain and Mover-Stayer Models as representation of Credit Behavior. *Oper. Res.*, **33**(6), 1203–1214.
10. Goodman, L. (1961). Statistical Methods for the Mover-Stayer Model. *J. Amer. Statist. Assoc.*, **56**, 841–868.
11. Krell, K., Frick, J. R., & Grabka, M. M. (2017). Measuring the Consistency of Cross-Sectional and Longitudinal Income Information in EU-SILC. *Rev. of Income and Wealth*, **63**(1), 30-52.
12. Lasso de la Vega, C., A. Urrutia and H. Diez (2010). Unit Consistency and Bipolarization of Income Distributions, Review of Income and Wealth **56**(1): 65-83.
13. Pittau M.G. (2005), Fitting Regional Income Distributions in the European Union, *Oxford Bull. Econ. Stat.*, **67**(2), 135-161.
14. Quah, D. T. (1996). Twin Peaks: Growth and Convergence in Models of Distribution Dynamics, Center for Economic Performance, Discussion paper No. 280. London School of Economics and Political Science.
15. Xuehui H., Shang-Jin W. (2017), Re-examining the middle-income trap hypothesis (MITH): What to reject and what to revive?, *J. Intern. Money and Finance*, **73A,** 41-61

Small Area Estimation of Inequality Measures

La stima per piccole aree di indicatori di disuguaglianza

Maria Rosaria Ferrante and Silvia Pacei

Abstract In order to estimate inequality measures at local level, small area estimation methods may be used to improve the reliability of estimates when the sample size is low. Small area models specified at area level, incorporate the design based estimates (direct estimates) as inputs, that are typically unbiased even though unreliable for small samples. Nevertheless, in the case of inequality measures, design based estimates are instead known to be biased for small sample sizes. In this work we focus on the search for a correction that can produce approximately unbiased direct estimators, taking into account the complexity of the survey design. We use data taken from the EU-SILC sample survey for Italy in 2013. Those modified estimators can then be used in small areas models.

Abstract *Allo scopo di stimare indicatori di disuguaglianza a livello locale, si possono impiegare metodi di stima per piccole aree per migliorare l'affidabilità delle stime quando la dimensione campionaria è piccola. I modelli per piccole aree specificati a livello di area, si basano su stimatori basati sul disegno (diretti), tipicamente corretti ma non affidabili per piccoli campioni. Gli stimatori degli indicatori di disuguaglianza sono invece distorti per piccoli campioni. L'obiettivo di questo lavoro è proporre una correzione che possa portare a stimatori diretti approssimativamente corretti, tenendo conto della complessità del disegno campionario. A questo scopo si usano i dati ottenuti per l'Italia dall'indagine* EU-SILC *del* 2013.

Key words: mean log deviation, complex sample survey, Fay-Herriot model.

[1] Maria Rosaria Ferrante; University of Bologna; email: maria.ferrante@unibo.it.

Silvia Pacei; University of Bologna; email: silvia.pacei@unibo.it.

1 Introduction

The increased interest for reliable information for restricted domain with reference to inequality measures is due to different reasons. One of the most relevant is to better plan policies to reduce inequality at local level. In this regards, an increasing gap in inequality and social exclusion among regions within the different EU member States has been observed in recent years. This issue is particularly relevant for Italy whose economic system is characterized by a strong territorial disparity.

Using data taken from the EU-SILC sample survey for Italy in 2013, we consider the estimation of inequality measures for the Italian provinces. Nevertheless the number of units sampled from many provinces is too low to provide reliable estimates using a “direct” estimator, that is an estimator calculated simply using the sample weights. This problem happens because EU-SILC survey is planned to provide reliable estimates for areas that are larger than those we are interested in.

To solve that problem we may resort to a small area estimation strategy. We consider area level models that incorporate the direct estimates as inputs. These estimates are typically obtained through unbiased estimators even though unreliable for small samples. Nevertheless, in the case of inequality measures, design based estimators are instead known to be biased for small sample sizes. The reason is that inequality measures can be written as ratios of random variables, both of which are estimated from the sample. They are thus biased in small sample, because the expected value of a ratio of random variables is not generally equal to the ratio of the expected values. The bias of the sample measure is $O\left(\frac{1}{n}\right)$, where n is the sample size.

In this work we focus on the search for a correction that can produce approximately unbiased direct estimators, taking into account the complexity of the survey design. Those modified estimators can then be used in small areas models.

We consider the class of generalized entropy (GE) measures, having the merit of satisfying the decomposability axiom, that allows to decompose the total inequality into the part due to inequality within areas and the part due to differences between areas. GE measures can be expressed as:

$$GE(\alpha) = \frac{1}{\alpha(\alpha-1)}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_j}{\bar{y}}\right)^{\alpha} - 1\right]; \quad j = 1, \dots, n; \quad \alpha \in [0, \infty) \qquad [1]$$

where $\bar{y}$ denotes the sample mean.

Specific special members of this family include Theil’s mean log deviation ($\alpha = 0$), Theil’s Index ($\alpha = 1$) and half the squared coefficient of variation ($\alpha = 2$). We start considering the mean log deviation for different reasons: *i*) it is used to study the “Inequality of opportunity” (Checchi and Peragine, 2010) with the purpose of assessing to what extent circumstances and efforts determine advantages; *ii*) it is particularly sensitive to changes in the tails of distribution, that are particular interesting in the case of income data; *iii*) it is found to be the less biased among the three indices mentioned (Breunig and Hutchinson, 2008).

2 Estimating the Theil's mean log deviation

We are interested in estimating the mean log deviation of the individual equivalized income, Y, for small areas indexed by i$= 1, \dots, m$. In the case of complex sample surveys, the direct estimator may be calculated using sample weights as follows:

$$ge(0)_i = \frac{1}{\hat{N}_i}\left[\sum_{j=1}^{n_i} w_j log\left(\frac{\bar{y}_i}{y_j}\right)\right]; \quad i = 1, \dots, m \qquad [2]$$

where $\bar{y}_i$ denotes the small domain sample mean calculated using sample weight, $\bar{y}_i = \frac{\sum_{j=1}^{n_i} w_j y_j}{\sum_{j=1}^{n_i} w_j}$, and $\widehat{N}_i = \sum_{j=1}^{n_i} w_j$.

In the literature a few papers consider the small sample bias issue for inequality measures (see Breunig, 2001, 2002; Giles, 2005; Breunig and Hutchinson, 2008) and propose a correction, but only in the simple random sample context. Breunig and Hutchinson (2008), for example, write the GE measures as functions of the population mean, μ, and some other population functions and then derive corrections for the GE measures, based on a second-order Taylor's series expansion of the sample estimates around the population values.

Regarding the mean log deviation, they obtain the following result for the approximate bias:

$$ABias\big(ge(0)\big) = -\frac{1}{2}\mu^{-2}Var(\hat{\mu}) \qquad [3]$$

They suggest to estimate [3] from sample data and then subtract it from $ge(0)$ to obtain a bias approximately corrected inequality value.

They also warn about the fact that the correction tends to increase the variability of the estimator, and that the overall reliability of estimates have to be considered.

Extension of this bias correction to the weighted estimator in equation [2] is not trivial. We consider an heuristic solution by substituting μ with the weighted sample mean and $Var(\hat{\mu})$ with the estimate obtained using the standard procedure used by Eurostat for a two-stage stratified sample (Eurostat, 2013). In particular, in EU-SILC survey carried out in Italy a stratified sample of municipalities is selected in the first stage and, in the second stage, a sample of households is randomly selected from the municipalities included in the first stage. The largest municipalities are always included in the sample (therefore they are called auto-representative or AR), while the other ones are selected according on a stratified sample where strata are defined by the administrative regions and the number of inhabitants (non auto-representative municipalities or NAR). The procedure used for estimating $Var(\hat{\mu})$ involves two different methods for AR and NAR municipalities. In our case, both estimates of μ and $Var(\hat{\mu})$ are calculated at small area level.

3 Simulation study

We carry out a simulation study to assess both the magnitude of the bias of the non-corrected estimator and the effectiveness of the correction adopted to reduce that small sample bias. To this purpose we consider the EU-SILC sample as target population and then repeatedly select random samples from it. We prefer to base our study on the EU-SILC dataset, rather than use data generated under some distribution model, to have a more realistic view of the small area estimation problem considered.

We consider as small areas the administrative regions and repeatedly select 1,000 two-stage stratified samples, mimicking the sample strategy adopted in the EU-SILC itself: in the first stage, AR municipalities are always included in the sample, while a stratified sample of NAR municipalities are selected; in the second stage, a simple random sample of households is selected from each municipality included in the first stage. We consider two overall sampling rates, 1.5 and 3%, to better understand the extent of the problem and the effectiveness of the solution with reference to different sample sizes. In our simulation setting the small area sample size ranges from a minimum of 6 to a maximum of 28 for the 1.5% sample, and almost twice for the 3% sample. $ge(0)$ and its bias corrected version, from now on $geCorr(0)$, are calculated considering the individuals, as usual. Individual equivalized income is, by definition, the same for all members of the same household.

$ge(0)$ and $geCorr(0)$ are compared in terms of bias and accuracy using the average absolute relative bias (AARB) and the average absolute relative error (AARE):

$$AARB = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{1000}\sum_{r=1}^{1000}(est_{ri}/GE(0)_i - 1)\right| \qquad [4.a]$$

$$AARE = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{1000}\sum_{r=1}^{1000}|est_{ri}/GE(0)_i - 1| \qquad [4.b]$$

where est_{ri} denotes the value of an estimator (alternatively $ge(0)$ or $geCorr(0)$) obtained for the r.th simulated sample and i.th small area, and $GE(0)_i$ is the true small area mean log deviation.

Percentage values of indicators in [4.a] and [4.b] are reported in Table 1. Results show that the correction considered greatly reduces the bias of the non-corrected estimator, although the corrected estimates remain a little biased on average. On the other hand, with respect to the concern about the reduction of the overall reliability of the estimates due to the correction, we find instead a negligible increase in the accuracy indicator.

Table 1: Percentage performance measures based on the simulation study

	1.5% sample		***3% sample***	
	$ge(0)$	$geCorr(0)$	$ge(0)$	$geCorr(0)$
AARB%	15.9	4.0	7.9	2.6
AARE%	51.8	52.3	37.8	38.2

References

1. Breunig R. (2001), An almost unbiased estimator of the coefficient of variation, Economics Letters, 70, 15-19
2. Breunig R. (2002), Bias correction for inequality measures: an application to China and Kenia, Applied Economics Letters, 9, 783-786
3. Breunig R., Hutchinson D.L.A. (2008), Small sample bias corrections for inequality indices, in "New Econometric Modeling Research", William N. Toggins ed., Nova Science Publishers: New York
4. Checchi D., Peragine V. (2010), Inequality of Opportunity in Italy, Journal of Economic Inequality, 8(4), 429-450.
5. Giles D.E. (2005), The bias of inequality measures in very small samples: some analytic results, Econometric Working Paper EWP0514, ISSN 1485-6441, University of Victoria, Department of Economics
6. EUROSTAT (2013), Standard error estimation for the EU-SILC indicators of poverty and social exclusion, EUROSTAT methodologies and working papers

Testing the Learning-by-Exporting at Micro-Level in light of influence of “Statistical Issues” and Macroeconomic Factors

Verifica dell’ipotesi Learning-by-Exporting a livello micro alla luce dell’influenza delle “questioni statistiche” e dei fattori macroeconomici

Maria Rosaria Ferrante and Marzia Freo

Abstract This study aims at testing the Learning-by-Exporting (LBE) on the TFP from the perspective of the evaluation literature. The focus is posed on the distribution of the outcome, the pre-entry selection bias is addressed, and both “statistical issues” and the influence of macroeconomic factors are accounted for. Basing upon a panel of Italian manufacturing firms, we design an experiment by aligning and pooling cohorts of starter, incumbent exporter and domestic firms and we further address the panel drop out. Main findings are that internationalisation has an impact on firms’ TFP, which is larger for best performing firms. Then it is shown that estimates of LBE impact are biased when i) the heterogeneous influence of macro-factors across groups and cohorts, and ii) the drop out of some firms from the panel are not accounted for.

Abstract *Il presente contributo indaga l’effetto del Learning-by-Exporting (LBE) nella prospettiva degli studi di valutazione. L’attenzione è posta alla distribuzione dell’outcome, e si tengono sotto controllo i meccanismi di auto-selezione, le questioni statistiche e l’influenza dei fattori macro-economici. Sulla base dell’analisi di un panel di imprese manifatturiere italiane (1998-2007), attraverso un elaborato disegno di ricerca, si verifica che l’internazionalizzazione ha un impatto sulla PTF delle imprese, specialmente sulle migliori. Inoltre viene mostrato*

[1] Maria Rosaria Ferrante, Department of Statistics, University of Bologna, Bologna; email: maria.ferrante@unibo.it

Marzia Freo, Department of Statistics, University of Bologna, Bologna; email: marzia.freo@unibo.it

Summary of the submitted paper “Detecting Learning by Exporting Effects on Firms’ Productivity Distribution in Presence of Alternate Phases of Export Demand” by same authors.

che le stime dell'impatto del LBE sono distorte se non si tiene conto i) *dell'influenza eterogenea dei fattori macro e della caduta e uscita di alcune imprese dal panel.*

Key words: Learning-by-Exporting, TFP, Panel Attrition, Macroeconomic Factors.

1 Introduction

During last years, many empirical studies have investigated the hypothesis that firms experience an increase in productivity during the period following their entrance into international markets (the Learning-By-Exporting or LBE hypothesis) and have displayed evidences for different countries with "mixed" conclusions ([4]). To investigate the mechanisms through which LBE may be explained and to identify a number of stylised facts for policy conclusions two meta-analyses have been run ([3], [5]) which highlight the roles played by issues related to sampling and methodological heterogeneity, on one side, and to different country-level macroeconomic environments, on the other side.

This study is a further contribution which aims to test the presence of the LBE effect on productivity growth, taking into account for meta-analysis's findings and going back to the firm-level. With respect to the most of previous micro-level studies, we design an ad-hoc research process which, besides micro-level factors, allows us to explicitly recognize macro-effect and the commonly called "sampling issues".

Differences between exporters and non-exporters are analysed from the perspective of the evaluation literature, by using counterfactual methods, and enlarging the focus to cover the whole probability distribution of the LBE effects looking at differences on various quantiles.

As far as the "sampling issues" are concerned, we test the LBE effect using a ten year panel data of Italian manufacturing firms observed during the period 1998-2007, and examine six cohorts of firms which started to export from 2000 to 2005. This data allows us to investigate firms behaviour during the pre- and post-entry periods. It is confirmed that the number of firms that may be observed to enter and remain in the international market for some years is quite low, and that the panel data suffer from attrition. Both these sampling issues have frequently emerged within the applied literature ([2]) but, at the best of our knowledge, they have never been jointly afforded.

Coming to the macroeconomic factors, their heterogeneity across domestic and international markets and over time is explicitly recognised. This is made possible by the length of the available panel that allows one to track the performance of firms over time and permits the observation and the control of cyclical movements. At the best of our knowledge in the literature on the evaluation of the LBE effect, the influence of the cycle is not considered at the micro-level.

In summary, in this study, we design a manifold research experiment in order to address these further sources of heterogeneity, concerning sampling and macroeconomic factors, and to obtain reliable statistical results on the LBE effect.

2 The Research Design

The detection of the LBE is managed as an empirical evaluation problem by comparing the TFP evolution of the treated group of firms starting to export to two untreated control groups which are respectively formed by domestic firms and by incumbent exporter firms. Exploiting the ten-year panel data, we define the treated "starters exporter" firms all firms observed to export for at least three years after having not exported for the two previous years. Performance of the treated groups is compared to two alternative control groups: the incumbent exporter firms which are always observed to export and the domestic firms which are observed selling their products only to domestic market, during all but one year. In order to take into account for the attrition bias due to drop-out, the study furtherly singles out groups of incumbent and domestic firms which from one year onwards exit from the panel (named "exiter") according this pattern: exiter firms are continuously present in the panel during the first five year and in some year after drop out. This allows us to run the analysis on the data both when they are balanced and when they are unbalanced. It is worth noting that many further firms in the panel have been considered not eligible for this study because they have not a well-defined pattern concerning the export status.

Since the number of export starters size for each cohort is small, to obtain a group of starters large enough to allow a reliable statistical analysis, we decide to align the sequences of firm-observations at the year in which each firm begins to export, which we fix as the time of treatment $j=0$ and to pool firms that belong to the six cohorts. In this way we obtain a group of 478 export starters and we consider the time to treatment variable in terms of the advance or delay to the treatment time (from $j=-2$ to $j=2$, where $j=0$ represents the treatment time). Note that the definition of the starters group leads us to observe a two-year-long period before the treatment ($j=-2,-1$) and a three-year-long period after the treatment ($j=0,1,2$). Moreover after the alignment, observations which share the same value of time-to-treatment variable are generated in different years and therefore they are affected by different level of macroeconomic variables.

Insofar, to detect the LBE effect on TFP, the six singled-out cohorts of starters are compared to the corresponding cohorts of domestic or incumbent firms. Then , the alignment and pooling of the cohorts of starters also requires that five-year-long sequences nested within longer sequence of incumbents and domestics are aligned and pooled. This action involves some caution to avoid deteriorating the comparisons in terms of the relative weights of the observations and in terms of the membership to specific target groups and years. In fact, the comparison is based on five-year-long sequences, and while starters have always five-year-long sequences, domestics and incumbents, due to their definition, exhibit longer sequences. In order to avoid domestic or incumbent firms being over-represented with respect to starter firms, we apply the following procedure for each comparison of the starter exporter to domestic (or incumbent) firms:

1. for each firm sequence among domestic (or incumbent), we extract all the five-year long sub-sequences; we align the sub-sequences at time to treatment j=0 and append to each sub-sequence the inclusion probability defined as the inverse of the number of sub-sequences which may be extracted by each firm sequence;
2. we sample five-year long sub-sequences according to their inclusion probabilities;
3. we pool the sampled sub-sequences to form the control group and we run analysis over the treated and the control groups.

This three-steps aligning-sampling-pooling procedure is repeated 100 times with a bootstrap.

3 The Evaluation Model

Based on the previously reported considerations, to properly measure the premium actually due to entry into the international market, we shall account for: 1) the pre-entry selection bias on TFP levels, 2) the bias due to drop out, and 3) the differential effect of the macroeconomic factors.

To estimate the LBE effect on the whole TFP distribution, we adopt the quantile decomposition methodology (QD, hereinafter) ([1]). According to QD, in the first step two conditional quantile models are used to establish the relationships between the productivity and covariates of the firms in two different groups along the entire distribution. In the second step, the conditional distributions implied by the estimated quantile model for firms in a certain group are applied to the observed covariate distribution of firms in the other state such that a counterfactual unconditional TFP distribution is estimated. Finally, the observed differences among the TFP distributions across the groups are decomposed into a component explained by the differences in the composition of covariates and a component explained by different returns to covariates (coefficients). In this way, it becomes possible to compute the impact of each of the components on the overall outcome distribution. The differences among distributions are evaluated at different quantiles. The component due to the covariates can be interpreted as the effect induced by the heterogeneity in characteristics, that is, by the self-selection mechanism on observables. The component due to the coefficients can be interpreted as the net export productivity premium.

To detect the post-entry TFP premium, we will estimate the net premium of starter firms, by comparing the dynamics of the productivity of the starter firms after entrance into international markets to the dynamics of the productivity of domestic and incumbents firms. These results are attainable by running QD over the bootstrapped samples drawn first over the balanced dataset and then over unbalanced dataset. These two cases may represent two bounds of values which premiums may attain. In fact, on one side, the measured TFP premium computed over always-present firms is likely to be underestimated because of the observed positive selection of domestic and incumbent always present firms. On the other side, the TFP premium that refers to Unbalanced dataset, which is enlarged to cover drop out observations, is likely overestimated because the premiums of the exiter firms are deteriorated by the crisis pattern.

With the aim of adjusting the TFP of exiters for drop out selection, we modify the QD approach by resorting to the Heckman selection model and to the two-stages Heckman estimator.

We adopt as the outcome variable the TFP change in years after the entrance. Concerning the specifications of conditional models adopted for estimating the net TFP premium, the specified covariates represent the observable characteristics for which outcome is controlled. This specification is chosen to further reduce, at least partially, the bias induced by the selection process on unobservable characteristics, while the QD mainly controls for the selection on the observables characteristics.

The general specification used for the quantile conditional model regresses the yearly rate of growth of TFP at time-to-treatment j=-1,0,1 over the set of structural firm characteristics that are supposed to explain the self-selection mechanism (industrial groupings dummies, macro-area dummies, TFP level, size in term logarithm of number of employees, and per capita wage at fixed time to treatment j=-2). Including the TFP pre-entry level allows us to control for unobserved pre-entry heterogeneity. To control for cyclical effects, the specification of the equation includes a vector d of five-year dummies This allows controlling for heterogeneity across cohorts and groups differently affected by the diminishing pull of export demand, as the TFP export premium is partially affected by macro-effects, which act differently across groups and cohorts. In the conditional model over unbalanced data as further term is added the inverse Mills' ratio which adjusts for drop out selection.

4 Main Results and Conclusions

LBE effects have been estimated as differentials in TFP rates of growth in favour of starter firms against either domestic or incumbents, respectively over balanced and unbalanced data and by adjusting unbalanced data for the drop out.

Estimates over balanced data show differential growth of TFP in favour of starter firms which are suddenly positive for the best-performing firms while they become positive two years after the entry for the slower-performing firms.

When the TFP premium growth rates are estimated over unbalanced data similar findings emerge, with the peculiarity that premiums are always higher with respect to those computed over balanced data and they are remarkably higher when considered versus domestics. They also are in general statistically significant.

The acceleration of premium growth during the third year is still higher when it is adjusted for drop out selection, even if not significant, because it discounts the higher standard errors of two-step estimators influencing the test results. Also in estimates adjusted for drop out it appears that the initial sunk costs of internationalization could produce lower growth for starter firms during the year of entrance, but during the post-entry periods, these premiums evolve faster. Also in this case, the premiums become positive earlier for the best-performing starter firms, which accelerate versus domestics just after entrance and versus incumbents one year after entrance.

In summary, according to expectations, if post-entry effects had not intervened, the net premiums of starters would have experienced a growth of the same strength of incumbents' or domestics' premiums, resulting in a null differential growth. Thus, starter exporting firms, after an initial deceleration, increase their TFPs more than incumbents and domestic firms. During the period after the entrance, the best-performing starters accelerate compared with domestics and the following period compared with incumbents. Two periods later, even the lesser- and medium-performing starter firms grow faster than the incumbents and even more than the domestics. Moreover, premiums become higher when the comparisons are enlarged to always present and exiter firms and are still higher when they are adjusted for drop-out. These results are evidence of LBE effects that intervene sometime after firms begin to export. Insofar, we are able to show, that neglecting these statistical and economical aspects could lead to hidden the LBE effect or to obtain biased estimation of the effect itself.

Concerning previous research questions, we find support to the LBE effect during the post-entry period. In particular, firms starting to export are in general observed to increase their TFP faster than domestic and incumbent firms two years after their entrance into international markets, and also before when they are best performing firms, that is they are at the top section of TFP's growth distribution. We also find that LBE effects would have been lower or even absent if we had not accounted for the macroeconomic cycle and the drop-out selection.

References

1. Chernozhukov, V., Fernández-Val, I., Melly, B.: Inference on counterfactual distributions. ECONOMETRICA. (2013) doi: 10.3982/ECTA10582.
2. ISGEP - International Study Group on Exports and Productivity: Understanding Cross-Country Differences in Exporter Premia: Comparable Evidence for 14 Countries. Rev. World Econ. (2008) doi: 10.1007/s10290-008-0163-y.
3. Martins, P.S., Yang, Y. The impact of exporting on firm productivity: a meta-analysis of the learning-by-exporting hypothesis. Rev. World Econ. (2009) doi: 10.1007/s10290-009-0021-6.
4. Wagner, J. Exports and productivity: A survey of the evidence from firm-level data. World Econ. (2007) doi: 10.1111/j.1467-9701.2007.00872.x
5. Yang, Y., Mallick S. Explaining cross-country differences in exporting performance: The role of country-level macroeconomic environment. Int. Bus. Rev. (2014) doi: 10.1016/j.ibusrev.2013.04.004

The mobility and the job success of the Sicilian graduates

La mobilità e il successo lavorativo dei laureati magistrali siciliani

Ornella Giambalvo and Antonella Plaia and Sara Binassi

Abstract The paper focuses on the job success of Sicilians Graduates comparing people who remain in Sicily for work reasons (Stayers) and who moves from Sicily to another Italian region or abroad (Movers) during 2016. Data referred to the *Almalaurea Consortium* annual survey collecting data of first and second level graduates interviewed at one, three or five years from graduation. We model the probability to be a successful graduate by means of a logistic regression model considering job and curriculum characteristics. The results show that many factors influence the job success, with some differences between graduates intervíwed 1 or 5 years after graduation, but it is clear that to be a Mover increases significantly the probability to be successful graduate.

Abstract *Il lavoro vuole analizzare il successo lavorativo dei giovani laureati siciliani che dopo un anno o dopo cinque anni dalla laurea lavorano. Nello specifico si analizzano le differenze fra le caratteristiche dei laureati Movers (quelli che lasciano la Sicilia) e Stayers (quelli che rimangono nella regione di studio) attraverso l'adozione di un modello logit che prende in considerazione variabili legate alle caratteristiche del lavoro e al curriculum universitario. I risultati mostrano che molti fattori influenzano il successo nel lavoro, con differenze tra gli intervistati a 1 e a 5 anni, ma è chiaro che essere un Mover aumenta significativamente la probabilità di essere un laureato di successo.*

Key words: Mobility, Survey Graduates' employment condition

1 Introduction

The complex reform process regarding issues of governance of the Italian University system (such as the progressive financial autonomy of the University and the reshape

Ornella Giambalvo, Antonella Plaia
Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Palermo, Viale delle Scienze, Edificio 19, 90128 Palermo, Italy, e-mail: ornella.giambalvo/antonella.plaia@unipa.it

Sara Binassi
Consorzio Interuniversitario AlmaLaurea, Viale Masini 6, 40126, Bologna, Italy, e-mail: sara.binassi@almalaurea.it

of the academic curricula) also caused the increasing spirit of competition among the universities. Every year, before the student's enrolment, Censis Institute, other research Institutions and other organizations (including Media), produce university rankings. Each of these takes into account several aspects referred to the university's organization, reputation and student benefits or referred to the student's actual characteristics. Some of them consider also the graduates employment rate one or three years after graduation. The Lisbon Strategy stressed the importance of the quality of employment regarding especially young people and young graduates. The strategic context for European cooperation in the education field estimates that the value of the graduates employment rate, for young people (20-34 years old) would be 82% in 2020. In 2016, the EU-28 employment rate of recent graduates (based on graduates – aged 20–34 – who had completed their education or training between one and three years prior to the survey) was 78.2%, ranging from highs of 90.2% in Germany and 96.6% in Malta to lows of 49.2% in Greece (52.9% in Italy). During the crisis period this employment rate followed a decreasing trend (-14% from 2008 to 2013) while in EU28 the decreasing trend was -6%. In Italy the economic crisis continues to produce its negative effects in employment and unemployment rates, particularly referring to young people (15-29 years old) and graduates: in 2016 the first rate reaches 49.5% whereas the second is 22.4%. Furthermore, the graduates employment rate varies according to the territorial differences: in 2016 it reaches the 43.4% in the Southern regions *vs.* the 65.9% in the Northern ones. Moreover, in the last fifteen years, students' enrolment has decreased significantly, while students' migration from the South to the Northern or the Central regions has increased [2]. This phenomenon creates further inequalities in within the country and a cultural and socio-economic loss for the South.

Based on the above results the paper focuses on the analysis of the job success of Sicilians Graduates between who remains in Sicily for work reasons (Stayers) or otherwise moves from Sicily to another region in Italy or abroad (Movers). Data referred to the *Almalaurea Consortium* annual survey collecting data of first and second level graduates interviewed at one, three or five years from graduation. We focus our attention to the second-level graduates interviewed after one and five years. In particular we consider the differences in job success according to some job characteristics and graduates' curriculum between Stayers and Movers. Who are the graduates that migrate and who are the students that stay? Is the migration of Sicilian graduates a consequence of personal and family dynamics? How do students that migrate from the South perform compared to students that stay? Who find a better job and when? This paper attempts to gives some responses to these important questions.

2 Data and variables

The 2016 AlmaLaurea survey on graduates' employment conditions [1] involved more than 16,000 first- and second-level graduates in 2015 enrolled in the universities of Sicily interviewed one year after graduation and 7,288 second-level graduates in 2011, interviewed five years after graduation. Since a high share of first-level graduates who comes from the universities of Sicily, continues their education with a master's degree (58% versus 56% of the overall population of first-level graduates), here we analyse the employment performance of second-level degree, focusing on one and five years after graduation. One year after graduation, the share of Sicilians graduates of 2015 employed who remain in Sicily for work reasons is 71%. Most of all women, with a degree in medicine, architecture, humanities and economic-statistics. The 29% decides to move, for work reasons, outside the region, most of all women, mostly from engineering and economic-statistics. Movers seem to be younger at graduation than Stayers and they are more regular in complying with the expected duration of the degree program. The graduation mark is slightly better for Movers than Stayers. In evaluating the characteristics of the work carried out, such as, for example, the type of work activity, remuneration and the correspondence between university studies and work carried out, it should be noted that, among the Stayers employed, a part of them entered the labour market already during the Master studies, or even before (33 against 13% of Movers). A further 9% changed their jobs after the conclusion of their studies, a value slightly lower than that for the Movers, equal to 14%. Finally, 58% entered the labour market only at the end of the Master degree (73% among Movers employed). One year after obtaining the title, self-employment concerns 14% of Stayers employed (the share is decidedly higher than the percentage of Movers, 6%). Almost in line between Stayers and Movers is the share of employed on permanent contracts (including the contract with increasing protection): 29% on Stayers compared to 32% on Movers. Stayers employed claim to receive a lower than Movers average salary.However, regardless of the type of work performed, part-time activities concern the 48% of those employed in Sicily against 18% of those who moves from the region. A year after the title, Stayers employed is effective or very effective (an indicator which combines a formal request title for the exercise of one's own work and use, in the work carried out, of the skills learned during the university) for more than half of the graduates workers (58% against 56% of Movers). Five years after graduation, the percentage of Sicilians graduates of 2011 employed who remain in Sicily for work reasons is 65%. Self-employment reaches 37% of Stayers employed (the share is decidedly higher than the percentage of Movers, 12%). The share of employed on permanent contracts rises among Movers: 53% compared to 40% on Stayers. Large part of the Stayers declare to carry out their activities in the private sector (73% against 63% of the Movers). More in detail they work within the services (88%; 5 percentage points more than the Movers). The industry sector, on the other hand, absorbs 9% of the Stayers and 16% of the Movers. Stayers employed claim to receive, also at five years after graduation, a lower average salary than Movers: the net monthly salary in fact, it is equal to 1,126 euros for Stayers, compared to 1,479 euros for the Movers (part-time activities concern the 26% of Stayers against 13% of Movers). Stayers employed is effective or very effective for more than half of the graduates workers (68% against 63% of Movers). Finally, in terms of job satisfaction, the gap is still positive for Movers.

3 First Results and some comments

As described in the previous sections, in this paper we want to compare the *job success* of what we called *Movers* vs *Stayers*. We define a Successful Graduate a person with a monthly neat earnings over the average (conditioning on his degree subject grouping and the type of degree) who uses greatly the skills acquired through the degree course. Conversely, an unsuccessful graduates, is a person with a monthly net earnings below the average (conditioning on his degree subject grouping and the type of degree) who uses limitedly or in no way the skills acquired through the degree course. We model the probability to be a successful graduate by means of a logit model [3] with the following covariates (chosen according to previous literature evidence [1]): the subject grouping of the degree, the graduation mark, the gender, the duration of studies, the age at graduation, the type of degree (single cycle vs second-level degree), the job sector of activity, part-time vs full-time job, satisfaction for the current job, together with some socio-economic characteristics of the graduate's family, such as parents' education level. Moreover, the interactions between Parents' education level and Migration, Gender and Migration, Parents' education level and gender have been considered. Tables 1 (graduates in 2015 interviewed after 1 year) and 2 (graduates in 2011 interviewed after 5 year) show the results with all the significant covariates and interactions.

According to Table 1, all the degree subjects show a probability of being a successful graduate lower than a graduate in Medicine; Males show a higher probability than female; a graduation delay below the average corresponds to a higher probability, even if these results are not coherent with the significance of the Age at graduation (we need to investigate further this aspect); the probability increases if at least one parent is graduated; Movers show a higher probability than Stayers; the probability increases if, during studies, graduates were resident in a province different from the University one; graduates have a higher probability to be a successful one if he/she did not work at graduation.

It's interesting to notice that, if we limit our analysis to only those students that did not work at graduation (table not shown), the *Graduation mark* and both the *Graduation delay* and the *Age at graduation* are no more significant (and this in someway explains the apparently discordant effect of these two variables shown in Table 1), nor they are all the previous significant interactions.

Considering 2011 graduates interviewed after 5 years (Table 2), coming from the universities of Catania and Palermo increases (with respect to Messina) the probability to be a successful graduate; only some of the degree subjects show a probability of being a successful graduate significantly lower than a graduate in medicine; Males show a higher probability than female; a graduation delay below the average corresponds to a higher probability, even if, again, this results is not coherent with the significance of the Age at graduation; Movers show a higher probability than Stayers; the probability decreases if the graduate works in a private or not-for-profit company or is part-time, while it increases with the satisfaction for the job.

Finally, focusing again on Movers and Stayers, it is evident that, even if we cannot say that "Being a Movers" is the first factor to be a successful graduate, the

Table 1 2015 Graduates at a Sicilian universities (interviewed after 1 year) participating 2016 Almalaurea's Graduates' employment conditions Survey: logit model for the probability to be a successful graduates. Baseline:Female Graduates in Medicine, with a graduation delay above the average, Stayers with no graduated parents, not employed during the studies and at graduation, resident in the same province of the University.

	B	S.E.	Sign.	Exp(B)
Graduation mark	0,043	0,015	0,005	1,044
Graduation delay (above average = 0)				
Below the average, conditioning on degree subject and type of degree	0,484	0,207	0,020	1,622
Age at graduation	0,143	0,022	0,000	1,154
Degree subject (Medicine = 0)				
Maths, Physics and Natural sciences	-4,316	0,695	0,000	0,013
Chemistry and Pharmacy	-2,810	0,564	0,000	0,060
Geology, Biology and Geography	-5,106	0,597	0,000	0,006
Healthcare sciences	-5,015	0,645	0,000	0,007
Engineering	-5,206	0,545	0,000	0,005
Architecture	-3,800	0,549	0,000	0,022
Agriculture and Veterinary	-3,641	0,744	0,000	0,026
Economics and Statistics	-4,084	0,522	0,000	0,017
Politics and Social Sciences	-5,036	0,620	0,000	0,007
Law	-4,748	0,659	0,000	0,009
Humanities	-4,142	0,550	0,000	0,016
Foreign languages	-3,843	0,599	0,000	0,021
Education	-3,880	0,601	0,000	0,021
Psychology	-4,214	0,597	0,000	0,015
Physical education	-4,490	0,721	0,000	0,011
Gender (Female = 0)				
Male	0,908	0,258	0,000	2,479
Parents' education level (none with a degree = 0)				
At least one is graduated	0,978	0,279	0,000	2,659
Movers to get the current job (Stayers = 0)				
Movers from Sicily	2,024	0,291	0,000	7,569
Parents' education level and Movers to get the current job (Stayers-none with a degree = 0)				
Movers - At least one is graduated***	-0,923	0,377	0,014	0,397
Employment condition during studies (no = 0)				
Lavoratore-Studente	1,484	0,348	0,000	4,412
Studente-Lavoratore*	0,162	0,197	0,411	1,176
Residence and University location (same province = 0)				
residence in a differente province	0,535	0,180	0,003	1,708
Residence in a different Region	-0,597	0,339	0,078	0,550
Employment condition at graduation (Working = 0)				
Not working	0,443	0,216	0,040	1,557
Intercept	-7,337	1,992	0,000	0,001

odd ratio of Movers relative to Stayers is 7.6 for 2015 graduates and 2.5 for 2011 graduates.

Table 2 2011 Graduates at a Sicilian universities (interviewed after 5 years) participating 2016 Almalaurea's Graduates' employment conditions Survey: logit model for the probability to be a successful graduates. Baseline:Female Graduates in Medicine, with a graduation delay above the average, Stayers with no graduated parents, not employed during the studies and at graduation, resident in the same province of the University.

	B	S.E.	Sign.	Exp(B)
University (Messina = 0)				
Catania	0,646	0,219	0,003	1,908
Palermo	0,618	0,225	0,006	1,855
Degree subject (Medicine = 0)				
Maths, Physics and Natural sciences	-3,290	0,748	0,000	0,037
Geology, Biology and Geography	-1,068	0,618	0,084	0,344
Healthcare sciences	-3,005	0,688	0,000	0,050
Engineering	-2,014	0,564	0,000	0,133
Architecture	-0,965	0,474	0,042	0,381
Agriculture and Veterinary	-1,516	0,654	0,020	0,220
Economics and Statistics	-2,120	0,576	0,000	0,120
Politics and Social Sciences	-1,487	0,587	0,011	0,226
Foreign languages	-1,140	0,609	0,061	0,320
Education	-2,091	0,642	0,001	0,124
Gender (Female = 0)				
Male	0,684	0,232	0,003	1,982
Movers to get the current job (Stayers = 0)				
Movers from Sicily	0,933	0,234	0,000	2,541
Sector of activity (pubblic = 0)				
Privat	-1,087	0,197	0,000	0,337
Not-for-profit	-1,237	0,479	0,010	0,290
Fyll-time/part-time (Full-time = 0)				
Part-time	-1,891	0,240	0,000	0,151
Satisfaction with the current job	0,492	0,050	0,000	1,635
Age at graduation	0,035	0,015	0,019	1,036
Graduation delay (above average = 0)				
Below the average, conditioning on degree subject and type of degree	0,458	0,161	0,004	1,580
Intercept	-4,663	1,871	0,013	0,009

References

1. AlmaLaurea. XIX Survey on Graduates' employment condition. Report 2017. Available on http://www.almalaurea.it/en/universita/occupazione/occupazione15
2. Enea, M. From South to North? Mobility of Southern Italian students at the transition from the first to the second level university degree. In Proocedings of 48th Scientific Meeting of the Italian Statistical Society. Università degli Studi di Salerno (2016)
3. D. Collett Modelling Binary Data, Second Edition, Chapman and Hall/CRC (1991).

Statistical Analysis of Energy Markets

Forecasting Value-at-Risk for Model Risk Analysis in Energy Markets

Angelica Gianfreda and Giacomo Scandolo

Abstract We consider the assessment of mis-specification risk when forecasting Value-at-Risk on a daily horizon. In particular, we focus on Energy Markets (electricity, oil, gas), where the impact of model risk may be relevant. Within an AR-GARCH framework to capture known features of volatility, we consider nine competing distributions for the standardized innovations and we apply a recently proposed measure of model risk to quantify the amount of model uncertainty in the procedure. Our approach is made more robust by discarding, on a daily basis, the worst performing models by using a set of weights built upon the Bayesian Information Criterion. The analysis covers the period 2001-2015, allowing for an in-depth assessment of the dynamics of model risk.

Key words: Building Weights, Risk Management, Electricity, Natural Gas, Brent Crude Oil

1 Introduction

In the financial literature, it has been recognized that the choice of the underlying probabilistic model for the risk factors can have a significant impact on a risk forecast. The hazard of producing a poor risk assessment due to the choice of an unsuited model is usually termed *model risk*. A distinction is usually made between two aspects of model risk: *estimation risk* and *mis-specification risk*. The former one refers to the uncertainty arising from parameters estimation, once a parametric family of distributions has been chosen. Instead, the latter one refers to the choice of

Angelica Gianfreda
Free University of Bozen-Bolzano, Universitätsplatz 1, Bozen e-mail: angelica.gianfreda@unibz.it

Giacomo Scandolo
University of Firenze, Via delle Pandette 9, Firenze e-mail: giacomo.scandolo@unifi.it

the parametric family itself. Given that quantifying and managing mis-specification risk is more difficult and it has been investigated to a lesser extent[1], we aim at focussing on this issue considering energy markets over almost fifteen years of data (from 2001 to 2015), so that we are able to assess model risk on a long-run basis and depict its historical evolution.

2 Methodology

Given a financial portfolio, the Value-at-Risk $\mathrm{VaR}_{\alpha,t+1}$, from day t to day $t+1$ at level α, is implicitly defined through the equality $P(L_{t+1} > \mathrm{VaR}_{\alpha,t+1}) = \alpha$, where $L_{t+1} = V_t - V_{t+1}$ is the loss from day t to day $t+1$ (here, V_t and V_{t+1} are the portfolio market values at days t and $t+1$). We consider the values $\alpha = 1\%$ and 5%, typical for market risk.

The assessment of VaR naturally depends on a probabilistic model. Henceforth, at any given date, competing models will produce competing VaR forecasts and model risk arises when these forecasts are dispersed. Several measures of model risk have been proposed in the literature. Here, we consider the Relative Measure of Model Risk (henceforth, RMMR) as defined in Barrieu and Scandolo (2015). Let VaR_i be the forecast of $\mathrm{VaR}_{\alpha,t+1}$ under model i, and VaR^* the forecast under a *reference* model; then, the RMMR is defined as the number

$$\mathrm{RMMR} = \frac{\max_i \mathrm{VaR}_i - \mathrm{VaR}^*}{\max_i \mathrm{VaR}_i - \min_i \mathrm{VaR}_i} \tag{1}$$

It easily turns out that this number lies in the interval $[0,1]$ (provided the reference model is among the competing models), and it is insensitive to the amount invested in the portfolio. We observe that the closer is RMMR to 1, the lower is VaR^* with respect to the other competing forecasts and therefore the higher is the amount of model risk involved.

In this work, we consider three portfolios, each of them investing in one of the following energy-related assets: Brent crude oil (*Oil*), ICE UK natural gas (*Gas*), and day–ahead auction prices for electricity observed on the European Energy Exchange (EEX) for delivery in the German/Austrian zones (*Electricity*). The oil and gas series always displayed positive prices, while several occurrences of negative prices were observed for electricity. Therefore, for the former two portfolios we set $L_t = -(e^{X_t} - 1)$, where X_t is the daily log-return observed at day t; for the electricity portfolio, we simply set $L_t = -X_t$, where X_t is the daily price change. In each case, we model the series (X_t) through an AR(5)–GARCH(1,1) process. Specifically, we have $X_t = \mu_t + \sigma_t Z_t$, where $\mu_t = \overline{\mu} + \sum_{i=1}^{5} \phi_i X_{t-i}$ is the conditional mean following an AR(5) process, and $\sigma_t^2 = \omega + \alpha(X_{t-1} - \mu_{t-1})^2 + \beta\sigma_{t-1}^2$ is the conditional variance following a GARCH(1,1) model. The IID innovation series (Z_t) can follow any of 9 competing standard distributions, which are: *normal* (NORM) and *skew*

[1] See for instance Cont (2006) and Daníelsson et al. (2016); among others.

normal (SNORM); *Student-t* (STD) and *skew Student-t* (SSTD); *Generalized Error Distribution* (GED) and *skew GED* (SGED); *Johnson's S_U family* (JSU); *Normal Inverse Gaussian* (NIG); *Generalized Hyperbolic family* (GHYP). Notice that some of these models are nested[2] and some of them allow for extra-parameters that give control on asymmetry and/or tail behaviour.

For any fixed distribution for the innovations, the parameters of the AR-GARCH model are estimated day-by-day by ML using a rolling window of 256 past daily data. This allows us to obtain a forecast VaR_i $(i = 1, \ldots, 9)$ of the daily Value-at-Risk under all competing models. Once a reference model is fixed throughout (NORM, for instance), the final output is a daily series of model risk measures.

As a result of the rolling estimation process, we obtain the maximized log-likelihood $\widehat{\ell_i}$ $(i = 1, \ldots, 9)$ for each day and competing model, and then the series of Bayesian Information Criterion, defined as $BIC_i = -2\widehat{\ell_i} + \log(n)p_i$, where n is the length of the dataset ($n = 256$ in our analysis) and p_i is the number of parameters in model i. [3] We observe that BIC is a measure of fitting ability (the lower is BIC, the better is the fitting) which penalizes for over-parametrization, and we use these values in two ways. First, on a daily basis we can rank the 9 models according to their fitting ability. In particular, at each day we single out the *daily best* model, as the one with the lowest BIC value; we then employ this model as the reference one in the computation of RMMR. Second, we can attach to each model a *percentage weight*, defined as

$$w_i = \frac{a_i^2}{\sum_{j=1}^{K} a_j^2}, \tag{2}$$

where

$$a_i = \frac{\max_j BIC_j - BIC_i}{\max_j BIC_j - \min_j BIC_j}.$$

Notice that w_i is decreasing in BIC_i: higher fitting ability is therefore associated with higher weights. We use these weights to discard, on a daily basis, the worst fitting models; specifically, after ranking the models with increasing weights, we retain models until the cumulative weight 0.95 is reached. We think this step is crucial in obtaining a measure of model risk that does not strictly depend on the initial choice of the competing models. As a consequence, RMMR do not necessarily lie in the interval $[0,1]$ if the reference model is among the discarded ones. Finally, we use the weights to obtain an average forecast, naturally defined as $\text{VaR}_{\text{avg}} = \sum_i w_i \text{VaR}_i$, which can be used as a possible reference model throughout (here, discarded models are left out of the average and weights are therefore properly normalized).

[2] For instance, NORM is a particular case of SNORM.

[3] For instance, $p_i = 10$ for the STD model: 6 parameters for the AR process, 3 for the GARCH process and 1 for the STD distribution.

3 Empirical results

As anticipated, for each day we obtain various measures of model risk, by considering as reference model either one specific distribution, or the daily best, or the average forecast $\mathrm{VaR}_{\mathrm{avg}}$. We always discard the worst fitting models, as explained above (on average, 2 or 3 models are discarded each day). We repeat the entire procedure for VaR at $\alpha = 1\%$ and $\alpha = 5\%$ and for all three portfolios of individual assets (oil, gas and electricity). All time series have been collected from Datastream, from 01/01/2001 to 31/12/2015, and are quoted on a basis of 5 days per week, for a total of 3914 observations. We used the R-package `rugarch`. Some of the RMMR series and related empirical findings are shown next.

The ”overall best” model for each of the three assets is identified by looking at the majority of days in which it is found to be the best fitting model. Overall best models (and in parenthesis the ”overall worst” models, being the best daily models in the fewest number of days) are: STD (GHYP) for oil, GED (GHYP) for gas, and STD (SNORM) for electricity. We observe that the Normal model is not the overall worst model for any of the three assets because the use of BIC penalizes models with many parameters.

In Table 1 we show some descriptive statistics of the RMMR (for $\alpha = 1\%$) for some choices of the reference model. We emphasize that while the overall best/worst models are fixed throughout for any given asset, the daily best model may change on a daily base. As explained before, weights are computed for each model: they are used for discarding the worst fitting models and to compute the average forecast $\mathrm{VaR}_{\mathrm{avg}}$.

Reference model		Oil	Gas	Ele
Normal	mean	1.04	0.54	0.64
	std. dev.	0.52	0.74	0.91
	max	5.95	4.75	5.04
Overall best	mean	0.52	0.64	0.77
	std. dev.	0.32	0.33	0.44
	max	2.28	3.44	8.25
Overall worst	mean	1.59	1.31	0.73
	std. dev.	0.46	0.44	0.50
	max	5.99	5.55	9.41
Daily best	mean	0.31	0.20	0.29
	std. dev.	0.30	0.28	0.34
	max	1.00	1.00	1.00
Average forecast	mean	0.48	0.49	0.44
	std. dev.	0.11	0.13	0.14
	max	0.87	0.88	0.95

Table 1 Some descriptive statistics for RMMR ($\alpha = 1\%$).

Looking at the mean levels, we see that the daily best models provide on average the lowest amount of model risk. The overall best gives less model risk than the overall worst for Oil and Gas; a bit surprisingly, this is not the case for Electricity. The model risk associated to the average VaR is stable around 0.5. Looking at the maximum levels we see that the RMMR sometime reach levels around 10, which signal a huge amount of model risk.

Figure 1 shows the daily dynamics of RMMR for Gas during 2002 (with $\alpha = 1\%$), when using the overall best (GED) and overall worst (GHYP) as reference models. We can see that, even though RMMR with respect to GHYP is consistently higher than for GED, the two series are highly volatile (as confirmed by the standard deviations in Table 1). Therefore, to ease the presentation, we used rolling means of the RMMR computed on a 256-day basis, and we show the dynamics of the RMMR ($\alpha = 1\%$) for two portfolios in Figure 2 (gas and oil shows similar RMMR dynamics). Finally, we also compare the RMMRs for both $\alpha = 1\%$ and $\alpha = 5\%$, showing[4] that the amount of model risk depends on α

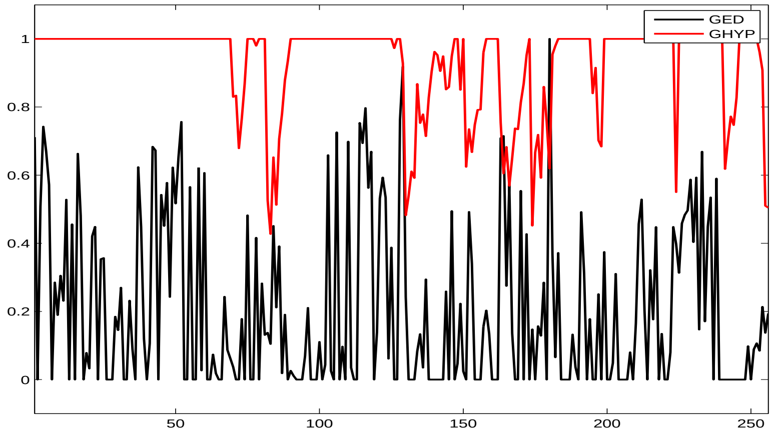


Fig. 1 Dynamics of RMMR for $\text{VaR}_{1\%}$ for Gas in 2002, using the overall best (GED) and the overall worst (GHYP) as reference models.

4 Conclusions

Quantifying and managing mis-specification risk has been less investigated so far, hence we have provided for the first time the empirical assessment of mis-specification risk when studying energy assets. Relaxing the assumption of normality and using a wide range of alternative distributions, we have quantified model risk under the well-established setting of GARCH models. Our empirical results emphasize that the distributional assumptions made in price modelling can produce a relevant discrepancy in risk figures and then trigger substantial model risk. In general, we find that better models tend to produce less model risk. Although not

[4] This figure is omitted for lack of space.

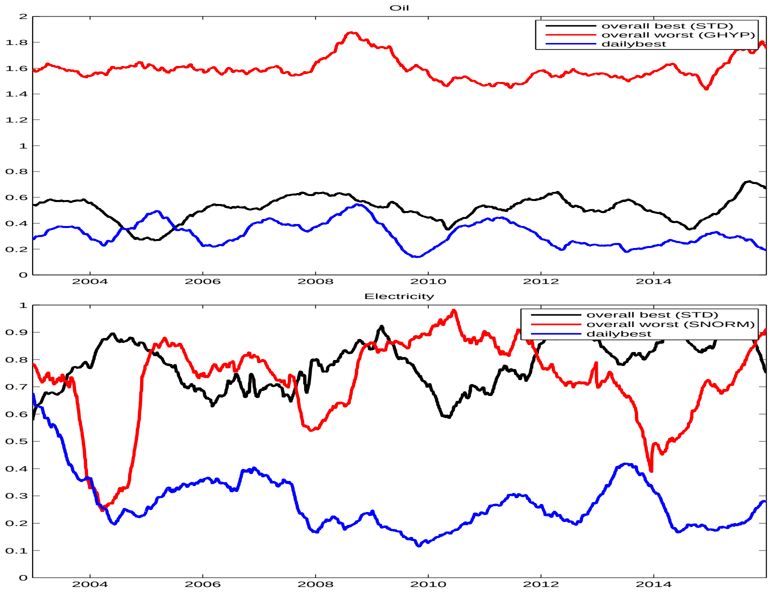


Fig. 2 Dynamics of rolling mean RMMR for $\alpha = 1\%$ using the overall best, the overall worst, and the daily best as reference models for Oil (top panel) and Electricity (bottom panel).

completely surprising, this pattern is quite evident across different assets and levels of VaR.

It is worth highlighting that our analysis intentionally addresses the choice of the innovations distribution, which is one among a number of possible sources of model risk affecting the final VaR figures. Future research may indeed address the misspecification risk due to the choice of the number of lags in the ARMA-GARCH structure or even due to the choice among different types of specification for the conditional mean/volatility.

Acknowledgements The authors kindly acknowledge the Free University of Bozen-Bolzano, for the internal research funding of the project *FOrecasting and MOnitoring electricity Prices, volumes and market Mechanisms*, “FoMoPM”.

References

1. Barrieu, P. and Scandolo, G.: Assessing financial model risk. European Journal of Operational Research. **242**, 546-556 (2015)
2. Cont, R.: Model uncertainty and its impact on the pricing of derivative instruments. Mathematical Finance. **16**, 519–547 (2006)
3. Daníelsson, J., James, K.R., Valenzuela, M., Zer, I.: Model risk of risk models. Journal of Financial Stability. **23**, 79–91 (2016)

Prediction interval of electricity prices by robust nonlinear models

Lisa Crosato, Luigi Grossi and Fany Nan.

Abstract It is well known that volatility of electricity prices estimated through GARCH-type models can be strongly affected by the presence of extreme observations. Although the presence of spikes is a well-known stylized effect observed on electricity markets, their presence has been often neglected and robust estimators have been rarely applied. In this paper we try to fill this gap introducing a robust procedure to the study of the dynamics of electricity prices. The conditional mean of de-trended and seasonally adjusted prices is modeled though a robust estimator of SETAR processes based on a polynomial weighting function (Grossi and Nan, 2015), while a robust GARCH is used for the conditional variance. The robust GARCH estimator relies on the extension of the forward search by Crosato and Grossi (2017). The robust SETAR-GARCH model is applied to the Italian electricity markets using data in the period spanning from 2013 to 2015. The purpose of this application is therefore twice: first, it is possible to enhance the prediction from point to intervals with associated probability levels, second, we set up a procedure to detect possible extreme prices which are commonly observed in electricity markets.

1 Introduction

In this paper we introduce a doubly robust approach to modelling the volatility of electricity spot prices, minimizing the misleading effects of the extreme jumps that characterize this particular kind of data on the predictions. The purpose of this paper is twice: first, it is possible to enhance the prediction from point to intervals with associated probability levels, second, we can detect possible extreme prices which are commonly observed in electricity markets. Although many papers have applied quite sophisticated time series models to prices and demand time series of electricity and gas very few have considered the strong influence of jumps on estimates and the need to move to robust estimators (Nowotarski et al.; 2013). Thus, the big differences between our paper and the previous literature related to jumps

Lisa Crosato
University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano. e-mail: lisa.crosato@unimib.it.

Luigi Grossi
University of Verona, Via Cantarane, 24, 37129 Verona. e-mail: luigi.grossi@univr.it.

Fany Nan
Joint Research Center of the European Commission, Ispra. e-mail: fany.nan@ec.europa.eu.

in electricity prices are twofold: 1) we don't detect extreme observations and we don't focus on the prediction of jumps; 2) We apply robust estimators which are not strongly influenced by the jumps. In this way we improve the prediction of normal prices which represent the majority of data. Our approach could then be integrated with other methods dealing with jumps forecasting.

Our methodology is applied separately to the hourly time series of Italian electricity price data from January 1st, 2013 to December 31th, 2015. In the first step, we apply a threshold autoregressive model (SETARX) to the time series of logarithmic prices. In the second step the weighted forward search estimator (WFSE) for GARCH(1,1) models is applied to the residuals extracted from the first step in order to estimate and forecast volatility. The weighted forward search is a modification of the Forward Search intending to correct the effects of extreme observations on the estimates of GARCH(1,1) models. Differently from the original Forward Search, at each step of the search estimation involves all observations which are weighted according to their degree of outlyingness.

2 A robust SETAR model for electricity prices

It is well-known that the series of electricity prices have long-run behaviour and annual dynamics, which change according with the load period. A common characteristic of price time series is the weekly periodic component (of period 7), suggested by the spectra that show three peaks at the frequencies $1/7$, $2/7$ and $3/7$, and a very persistent autocorrelation function.

We assume that the dynamics of log prices can be represented by a nonstationary level component L_{tj}, accounting for level changes and/or long-term behaviour, and a residual stationary component p_{tj}, formally:

$$\log P_{tj} = L_{tj} + p_{tj}. \tag{1}$$

To estimate L_{tj} we use the wavelets approach (Lisi and Nan, 2014). We consider the Daubechies least asymmetric wavelet family, LA(8), and the coefficients were estimated *via* the maximal overlap discrete wavelet transform (MODWT) method (for details, see Percival and Walden, 2000).

The detrended prices p_t, will then be modeled by a two-regime Self-Exciting Threshold AutoRegressive model SETAR(7,1) which is specified as

$$p_t = \{ \mathbf{x}_t \beta_1 + \varepsilon_t, \quad \text{if} \quad p_{t-1} \leq \gamma \mathbf{x}_t \beta_2 + \varepsilon_t \quad \text{if} \quad y_{t-1} > \gamma \tag{2}$$

for $t = 1,...,N$, where p_{t-1} is the threshold variable and γ is the threshold value. The relation between p_{t-1} and γ states if p_t is observed in regime 1 or 2. $\beta_{\mathbf{j}}$ is the parameter vector for regime $j = 1,2$ containing 7 coefficients to account for weekly periodicity. $\mathbf{x_t}$ is the t-th row of the $(N \times 7)$ matrix $\mathbf{X}$ comprising 7 lagged variables of p_t (and eventually a constant). Errors ε_t are assumed to follow an $\mathrm{iid}(0, \sigma_\varepsilon)$ distribution.

Parameters can be estimated by sequential conditional least squares. In the case of robust two-regime SETAR model, for a fixed threshold γ the GM estimate of the autoregressive parameters can be obtained by applying the iterative weighted least squares:

$$\hat{\beta}_j^{(n+1)} = \left(\mathbf{X}_j' \mathbf{W}^{(n)} \mathbf{X}_j\right)^{-1} \mathbf{X}_j' \mathbf{W}^{(n)} \mathbf{p}_j \tag{3}$$

where $\hat{\beta}_j^{(n+1)}$ is the GM estimate for the parameter vector in regime $j = 1,2$ after the n-th iteration from an initial estimate $\hat{\beta}_j^{(0)}$, and $\mathbf{W}^{(n)}$ is a weight diagonal matrix, those elements depend on a weighting function $w(\hat{\beta}_j^{(n)}, \hat{\sigma}_{\varepsilon,j}^{(n)})$ bounded between 0 and 1. The threshold γ can be estimate by minimizing the objective function $\rho(r_t)$ over the set Γ of allowable threshold values. Weights are calculated as

$$w(\hat{\beta}_j, \hat{\sigma}_{\varepsilon,j}) = \psi\left(\frac{y_t - m_{y,j}}{C_y \hat{\sigma}_{y,j}}\right) \psi\left(\frac{y_t - \mathbf{x}_t \hat{\beta}_j}{C_\varepsilon \hat{\sigma}_{\varepsilon,j}}\right)$$

where ψ is the polynomial weight function (Maronna et al., 2006) and $m_{y,j}$ is a robust estimate of the location parameter (sample median) in the j-th regime. $\hat{\sigma}_{y,j}$ and $\hat{\sigma}_{\varepsilon,j}$ are robust estimates of the scale parameters σ_y and σ_ε respectively, obtained by the median absolute deviation multiplied by 1.483. C_y and C_ε are tuning constants fixed at 6.0 and 3.9 respectively.

3 Robust volatility estimation through the forward search

Now we apply the weighted forward search (WFS) estimator to derive robust prediction intervals for the volatility of electricity prices, starting from the residuals of the SETAR model estimated in section 1. Let ε_t denote an observed time series of heteroscedastic residuals. For electricity prices $\varepsilon_t = p_t - \hat{p}_t$ where $\hat{p}_t$ is the price fitted by the robust SETAR model. We proceed now by estimating a standard GARCH(1,1) model on residuals ε_t, so that $\varepsilon_t | F_{t-1} \sim N\left(0, \sigma_t^2\right)$ and

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{4}$$

with $\alpha_0 > 0, \alpha_1 \geq 0, \beta \geq 0, \alpha_1 + \beta < 1$.

The WFS (see Crosato and Grossi, 2017, for details) starts with the selection of a small subset of observations showing the minimum median of squared residuals. Then, at each step of the search, MLE estimation is carried out on all observations weighted to account for their degree of outlyingness so that the estimation pattern along the search is not influenced by the presence of spikes. All units, a part of the first b, are ranked according to their squared standardized residuals with respect to the model estimated at the previous step of the search. Proceeding along the search, an increasing number of units are given weight 1 while the remaining are downweigthed by the corresponding value of the complementary cumulative distribution function of the squared standardized residuals defined on the whole sample. The weights range from approximately 0 to 1 so that outliers will be heavily downweighted until the end of the search while the closer the weight to 1 the stronger is the degree of agreement of the observation with the estimated GARCH. This way the temporal structure of the time series is respected, filling the time gaps created by the forward search ordering but on the same time observations are ordered consistently with the GARCH model estimated until the last steps when all observations enter with their original value.

A plus of the WFS approach is that outliers are identified automatically through a test reducing the arbitrariness in declaring a given observation as outlier. Once outliers have been identified the forward plots of coefficient estimates are cut automatically and WFS GARCH estimates are obtained.

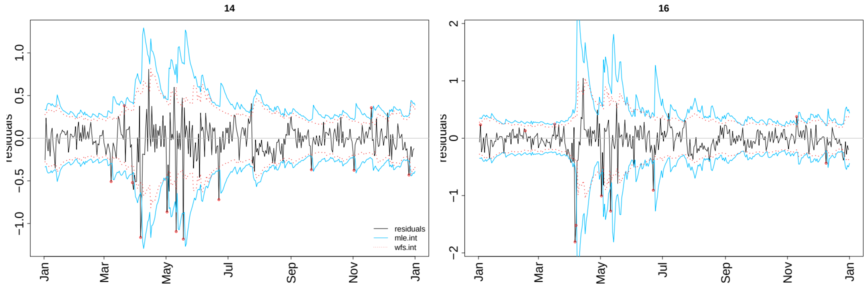


Fig. 1 Forecasted conditional confidence intervals of volatility for hourly electricity prices in Italy (hour 14 and 16, January to December 2015) obtained by MLE (light blue) and WFS (dotted red).

The outliers identified through the WFS test vary from a minimum of 29 (hour 21) to a maximum of 59 (hours 8 and 9). A few outliers characterize many hours of the same day, as is the case of the 21st and 27th of April, 2013 for 14 hours as well as mot days of June 2013. Spikes in 2014 characterize mainly the months of February, March and April. Outliers for 2015, hours 14 and 16, are highlighted by red circles in figure 1.

Using the MLE and WFS estimated coefficients and applying equation (4) for the conditional variance recursively, we then obtain the forecasted price volatility and the corresponding 95 percent confidence intervals (see figure 1). As can be seen, the WFS intervals (dotted red) are tighter around the realized volatility than the MLE ones (light blue). The gain in prediction is evident in the aftermath of a spike or a drop, for instance in April and May. Note that the correction provided by the WFS works also for smallest jumps, as in July and December.

References

1. Crosato L., Grossi L.: Correcting outliers in GARCH models: a weighted forward approach. Statistical Papers (2017) doi:10.1007/s00362-017-0903-y.
2. Grossi L., F. Nan: Robust estimation of regime switching models. In: Morlini I., Minerva T., Palumbo F. (eds.) Advances in Statistical Models for Data Analysis, pp. 125–135. Springer International Publishing, Switzerland (2015).

3. Lisi, F., Nan, F.: Component estimation for electricity prices: Procedures and comparisons. Energy Economics **44**, 143–159 (2014).
4. Nowotarski J., Tomczyk J., Weron R.: Robust estimation and forecasting of the long-term seasonal component of electricity spot prices. Energy Economics **39**, 13–27 (2013).
5. Percival, D., Walden, A.: Wavelet Methods for Time Series Analysis. Cambridge University Press. (2000).

Bias Reduction in a Matching Estimation of Treatment Effect

Sulla Riduzione della Distorsione nella Stima Matching dell'Effetto del Trattamento

Maria Gabriella Campolo, Antonino Di Pino and Edoardo Otranto

Abstract The traditional matching methods for the estimation of the treatment parameters are often affected by selectivity bias due to the endogenous joint influence of latent factors on the assignment to treatment and on the outcome, especially in a cross-sectional framework. In this study, we show that the influence of unobserved factors involves a cross-correlation between the endogenous components of the propensity scores and causal effects. A correction for the effects of this correlation on matching results leads to a reduction of bias. A Monte Carlo experiment supports this finding.

Abstract *I tradizionali stimatori matching dei parametri del trattamento spesso producono stime affette da selettività dovuta all'influenza endogena di fattori latenti, specialmente nelle analisi cross-section. In questo studio, mostriamo che l'influenza di fattori non osservabili mette in correlazione la propensione a sottoporsi al trattamento e gli effetti causali prodotti da quest'ultimo. La correzione delle stime matching per gli effetti di questa correlazione consente di ridurre la distorsione dovuta alla selettività. Questo risultato è supportato dalle evidenze di una serie di esperimenti Monte Carlo.*

Key words: propensity score matching, endogenous treatment, measurement and transition equation.

[1] Maria Gabriella Campolo, University of Messina; email: mgcampolo@unime.it
Antonino Di Pino, University of Messina; email: dipino@unime.it
Edoardo Otranto, University of Messina; email: eotranto@unime.it

1 Introduction

A typical assumption of models for treatment effects is based on the hypothesis that the decision of a subject to receive a certain treatment depends on the difference in the outcomes potentially gained by the subject under the two alternative regimes of treatment and control, respectively (see, e.g., Winship and Morgan, 1999). Starting from this assumption, the decision of a subject to undergo the treatment is endogenous with respect to the potential outcome. The non-random selection of the units into the treatment regime, due to the endogeneity of treatment, involves that important unobserved covariates influence jointly the propensity of a subject to undergo the treatment and the outcomes. As a consequence, matching estimation of the treatment effect, based on the comparison of treated and untreated units with the same propensity score, is biased (e.g. Austin, 2011).

In this context, a natural solution, as the detection of new statistically significant covariates in the treatment choice equation, could not reduce the bias; in fact, Heckman and Navarro-Lozano (2004) show that this is the case when these variables are not exogenous with respect to the outcome.

In this study, we try to circumvent the problem of misspecification of the selection equation in matching methods based on propensity score, assuming that the potentially omitted endogenous factors can be represented by a stochastic component correlated with the causal effects of the treatment. This implies that the causal effect of each subject is correlated with the causal effect of another subject with similar propensity score; moreover, the stochastic component is autocorrelated, as causal effects relative to similar propensity scores will be more similar. In order to assess this endogenous relationship, we model the causal effects adopting a sort of state-space model (see, for example, Harvey, 1990), where a common latent factor is detected in correspondence of the endogenous stochastic component of the propensity score sorted in an ascending (or descending) order. State-space models are generally adopted for time series; the extension to this framework is simple, substituting the ordering of the observations in terms of dating with the order in terms of increasing propensity score. The predictions of these components are used as correction terms in the matching procedure. The estimation method proposed, called State-Space Corrected Matching (*SSCM*), is based on the Kalman filter (see Harvey, 1990) and possesses the nice characteristic of not imposing conditions of identification of the probability to undergo the treatment as in the randomized experiments.

We verify the performance of this method comparing its bias with respect to the bias occurring with traditional propensity score matching (cf., among others, Rosenbaum and Rubin, 1983) by Monte Carlo experiments. In the Monte Carlo experiment we generate data in a cross-sectional context, adopting a two-regime model whose data generation process (*DGP*) is affected by endogeneity. Applying our correction method, we obtain a marked reduction of bias in the estimated average treatment effect for the treated (*ATT*) in comparison with the traditional Propensity Score Matching estimator (*PSME*, Rosenbaum and Rubin, 1983).

In next Section we describe this new procedure, whereas in Section 3 we show the results of the Monte Carlo experiments comparing the performance of *SSCM* estimation method and traditional *PSME* in terms of prediction of the *ATT* parameter.

2 The Model

In this paper we insert several novelties with respect to the present literature. The most relevant is the individuation of an autoregressive process that characterizes, jointly, individual propensity scores and causal effects. As a consequence, another important novelty is given by the correction term based on the estimation of a State-Space model in which the endogenous component common to causal effects and propensity scores is specified by a "measurement" equation and a "transition" equation, respectively. In addition, in our model it is not necessary to reproduce conditions of identification of the probability to undergo the treatment such as in a randomized experiment.

In this analysis, we start to consider the potential outcome gained by choosing one of the two treatment status as a relevant (endogenous) determinant of the decision undergoing the treatment. In particular, we specify the model assuming that the difference between the potential (expected) outcomes, y_{1i} and y_{0i}, obtainable, respectively, under the regimes T_i =1 (if the subject belongs to the treatment group) and T_i=0 (if the subject belongs to the comparison group), determines, at least in part, the choice of the regime.

We specify a Probit (or Logit) model, where the (latent) propensity to undergo the treatment of the *i-th* subject, T^*_i, depends linearly on the covariates in **Z**:

$$T*_i = \mathbf{z}'_i \boldsymbol{\beta} + v_i \tag{1}$$

where $\mathbf{z}'_i$ is the *i*-th row of the matrix **Z**, $\boldsymbol{\beta}$ is a vector of unknown coefficients and v_i is a zero-mean random disturbance with unit variance. If $T^*_i > 0$, $T_i = 1$ (the subject is undergone to treatment), otherwise $T_i = 0$ (the subject is not undergone to treatment).

Assuming that the assignment to treatment is endogenous, T^*_i will depend on the causal effect $\Delta_i = y_{1i} - y_{0i}$.

Formally, we can explain autocorrelation and endogeneity specifying our model similarly to a generalized Roy model (e.g., Carneiro et al., 2003). In doing this, we add to the above selection equation (Eq. 1) two equations that specify the outcome of treated and untreated subjects, as follows:

$$y_{1i} = \mu_{1i} + u_{1i} \qquad \text{if } T_i = 1; \tag{2a}$$

$$y_{0i} = \mu_{0i} + u_{0i} \qquad \text{if } T_i = 0; \tag{2b}$$

In Eqs. (2a) and (2b) μ_{1i} and μ_{0i} are the expected outcomes, respectively, of treated and untreated subjects, depending on the decision to undergo the treatment (T = 1) or not (T = 0). The error terms u_{1i} and u_{0i} are normally distributed with zero mean and variances equal to σ^2_1 and σ^2_0 respectively. The covariances σ_{1v} and σ_{0v} of the disturbances of both outcome equations, u_{1i} and u_{0i}, with the disturbances of the selection equation (1), v_i, can be different from zero as a consequence of endogeneity. The covariances σ_{1v} and σ_{0v} are measurements of the endogeneity of the propensity to undergo the treatment, T^*_I with respect to the outcome gained under T =1 and T = 0.

Correlation between outcomes and propensity scores, as well as the autocorrelation of the causal effects, may be specified starting from the definition of causal effects, Δ_i. Hence, subtracting Eq. (2b) from Eq. (2a), we obtain:

$$\Delta_i = y_{1i} - y_{0i} = \mu_{1i} - \mu_{0i} + (u_{1i} - u_{0i}) \tag{3}$$

Imposing a linear relationships between the error terms of the outcome equations and the selection equation, we have:

$$u_{1i} = \sigma_{1v} v_i + \varepsilon_{1i} \tag{4a}$$

$$u_{0i} = \sigma_{0v} v_i + \varepsilon_{0i} \tag{4b}$$

where ε_{1i} and ε_{0i} are i.i.d. disturbance with zero mean. By substituting (4a) and (4b) into Eq. (3):

$$\Delta_i = y_{1i} - y_0 = \mu_{1i} - \mu_0 + (\sigma_{1v} - \sigma_{0\text{v}})v_i + (\varepsilon_{1i} - \varepsilon_{0i}) \tag{5}$$

Putting $\mu_{1i} - \mu_{0\text{i}} = \mu_i$; $(\sigma_{1v} - \sigma_{0\text{v}})v_i = \sigma\, v_i$ and $(\varepsilon_{1i} - \varepsilon_{0i}) = \varepsilon_i$, Eq. (5) can be written as a *measurement equation* of a state-space model, as follows (cf., among others, Harvey, 1990):

$$\Delta_\text{i} - \mu_i = \sigma\, v_i + \varepsilon_\text{i} \tag{6}$$

In Eq. (6), ε_i is a vector of n×1 disturbance terms uncorrelated across i. The variable v_i can be considered as the state variable whose elements are not observable, but are assumed to be generated by a first-order Markov process, such as the following transition "equation":

$$v_i = \rho v_{i-1} + \eta_\text{i} \tag{7}$$

The dependent variable of Eq. (6), $\Delta_\text{i} - \mu_i$, may be considered as the stochastic component of the causal effect Δ_i, endogenous with respect to the decision to undergo to treatment. Starting from this result, the selectivity effect due to the endogeneity of the decision to undergo the treatment may be corrected by estimating σv_i in Eq. (6), and using the predicted values, $\sigma\hat{v}_i$, as a correction term in the matching estimation of the causal effects. In doing this, a preliminary estimation of causal effects Δ_i is obtained at a first stage by applying a propensity score matching procedure. Then, at a second stage, matching is replicated using the corrected outcomes y_{1i} - $\sigma\hat{v}_i$ so as to obtain the corrected causal effects Δ_i -$\sigma\hat{v}_i$ $= \hat{\mu}_i$. We call this estimator the State-Space Corrected Matching (*SSCM*) estimator.

3 Monte Carlo Experiment

We propose a Monte Carlo experiment to compare the performance of the *SSCM* procedure with that of the *PSME* in terms of bias reduction under both the conditions of heterogeneous and homogeneous covariates between regimes. For this purpose, we generate 500 data sets of 2,000 units from the Two-Regime model above in Eqs. (1), (2a) and (2b). The exogenous covariates Z are generated in order to reproduce the very frequent condition of heterogeneity in observed covariates between treatment and comparison group, and the condition of homogeneity in the observed covariates between regimes. We consider two different *DGPs*, with and without endogeneity, so as to fix two distinct set of population parameters under the condition of endogeneity and exogeneity, respectively.

Table 1 summarizes the estimated *ATT* values obtained by embedding different endogeneity conditions into the *DGP*. Computing the bias with respect to the population *ATT* value (set equal to 5), the *SSCM* estimator performs better than the *PSME* procedure. The bias resulting from the application of *SSCM* is markedly smaller than the one resulting from *PSME*.

Table 1: Simulation Results. Estimated ATT parameters. Population *ATT* value = 5. Generated sample size: n = 2000. No of reps. 500. Simulated endogeneity by setting σ_{1v} and σ_{0v}.

		SSCM			*PSME*	
Endogeneity	***ATT***	*95% CI*		***ATT***	*95% CI*	
σ_{1v} 5.4; σ_{0v} 2.4	4.974	*4.930*	*5.018*	7.996	*7.970*	*8.022*
σ_{1v} 5.4; σ_{0v} -2.4	4.320	*4.279*	*4.362*	6.814	*6.782*	*6.846*
σ_{1v} 5.4; σ_{0v} 0.8	4.983	*4.939*	*5.027*	7.571	*7.534*	*7.607*
σ_{1v} 5.4; σ_{0v} -0.8	4.729	*4.688*	*4.770*	7.572	*7.536*	*7.608*
	% BIAS*	*St.Dev.*	*t***	***% BIAS****	*St.Dev.*	*t***
σ_{1v} 5.4; σ_{0v} 2.4	-0.51%	*0.022*	222.160	59.92%	*0.013*	*607.170*
σ_{1v} 5.4; σ_{0v} -2.4	-13.59%	*0.021*	205.720	36.28%	*0.016*	*414.280*
σ_{1v} 5.4; σ_{0v} 0.8	-0.35%	*0.022*	222.480	51.41%	*0.018*	*410.660*
σ_{1v} 5.4; σ_{0v} -0.8	-5.42%	*0.021*	225.710	51.45%	*0.018*	*412.000*

Note: * % of Bias [(Est. ATT-5)/5]%; ** t-ratio: ATT/St.Dev.)

We can observe, in particular, that, if we reproduce the “more common” endogeneity conditions (characterized by covariances, σ_{1v} and σ_{0v}, with the same sign) in the *DGP*, the confidence intervals obtained by the *SSCM* estimates include the population *AT*T value. In the less frequent case, in which the propensity to undergo the treatment is endogenously affected in the two regimes with opposite sign, confidence intervals of the *SSCM* estimates do not include the population parameter. However, the percentage of bias of *SSCM* estimation does not exceed 15% in absolute value.

4 Conclusion

The aim of this study is to improve the propensity-score matching approach so that estimation results do not overly suffer from the influence of the endogeneity of treatment. We show that, applying a state-space model, we can estimate the endogenous component of the causal effects, so as to use the result of this estimate as a correction term. In particular, the results of the Monte Carlo experiments here reported confirm that, simulating endogeneity of the selection into treatment in a Two-Regime model, the predicted components of causal effects can be successfully used, at a second stage of the estimation procedure, to correct the matches outcomes.

As the results of our Monte Carlo analysis show, this method allows us to reduce the selectivity bias in matching without imposing, to the data or the model, any restriction usually adopted to reproduce a condition comparable to the randomization. At this stage of our research, we have deepened the characteristics of the SSCM estimator only through Monte Carlo experiments. However the inferential properties must still be investigated. This will be the next aim of this research.

References

1. Austin, P. C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioural Research, 46, 399-424 (2011)
2. Carneiro, P., Hansen, K. T., Heckman, J. J.: Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. International Economic Review 44(2), 361-422 (2003)
3. Harvey, A. C.: Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge: Cambridge University Press (1990)
4. Heckman, J. J., Navarro-Lozano, S.: Using matching, instrumental variables, and control functions to estimate economic choice models. The Review of Economics and Statistics, 86(1), 30-57 (2004)
5. Rosenbaum, P. R, Rubin, D. B.: The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55 (1983)
6. Winship, C., Morgan S. L.: The estimation of causal effects from observational data. Annual Review of Sociology, 25, 659-706 (1999)

Statistical Inference and Testing Procedures

Comparison of exact and approximate simultaneous confidence regions in nonlinear regression models

Confronto tra la regione di confidenza esatta ed approssimata nei modelli di regressione nonlineari

Claudia Furlan and Cinzia Mortarino

Abstract Accuracy measures for parameter estimates represent a tricky issue in nonlinear models. Practitioners often use the separate marginal confidence intervals for each parameter. However, these can be extremely misleading due to the curvature of the parameter space of the nonlinear model. For low parameter dimensions, routines for evaluating approximate simultaneous confidence regions are available in the most common software programs, but the degree of accuracy also depends on the intrinsic nonlinearity of the model. In this paper, the accuracy of the marginal confidence intervals, Hartley's exact simultaneous confidence region (sCR), and the most widespread approximate sCR are compared via both real data and simulations, for discrete time diffusion models in the class of nonlinear regression models.
Abstract *Nei modelli non lineari non è scontato riuscire ad ottenere misure di accuratezza delle stime. Nella pratica spesso si usano gli intervalli di confidenza marginali per ogni parametro, ma questa procedura può portare a risultati inaffidabili a causa della curvatura dello spazio parametrico tipico dei modelli non lineari. Nei più comuni software si può trovare implementato il calcolo della regione di confidenza simultanea approssimata per un numero ridotto di parametri, ma il livello di copertura esatto dipende dal grado di non linearità intrinseca del modello. In questo lavoro, nell'ambito dei modelli di regressione non lineari e in particolare per i modelli di diffusione a tempo discreto, si confrontano fra loro i livelli di copertura degli intervalli di confidenza marginali, della regione di confidenza simultanea (sCR) esatta di Hartley e della sCR approssimata più utilizzata.*

Key words: nonlinear models, simultaneous confidence region, Hartley's simultaneous confidence region, Bass model

1 Introduction

Nonlinear models are the natural modelling framework for many real-world phenomena. Unlike linear models, accuracy measures for parameter estimates, such as confidence intervals or confidence regions, may represent a difficult task due to the intrinsic curvature of the parameter space. A common mistake is relying on

Claudia Furlan e-mail: furlan@stat.unipd.it · Cinzia Mortarino e-mail: mortarino@stat.unipd.it
Department of Statistical Sciences, University of Padova, Italy.

marginal confidence intervals, whose use can be misleading. Simultaneous confidence regions are usually available in the most commonly used software programs, only in the approximate form at least for low parameter dimensions.

The problem of constructing exact confidence regions for the parameters of nonlinear models has received little attention in the past (Lee et al, 2002), since this is computationally intensive. Given the complexity of obtaining an exact simultaneous confidence region (sCR), a few approximations have been proposed (Seber and Wild, 1989) under the normality assumption of homoscedastic errors. For instance, the so-called ‘approximate’ sCR is derived by approximating the nonlinear model via a linear Taylor expansion, thereby taking advantage of the asymptotic normality of the estimator. Thus, the approximate confidence levels of the ‘approximate’ sCR are valid asymptotically. This approximation is computationally more attractive, since it corresponds to hyperellipsoids.

Under the normality assumption of homoscedastic errors, Hartley (1964) proposed an exact sCR based on inverting an exact test. However, the power of the exact test, and thus the coverage probability of the corresponding sCR, depends on the choice of the idempotent projection matrix. More recently, Demidenko (2017) studied the exact statistical properties in small samples.

Among the nonlinear regression models, in this paper, we focus on two of the most widespread discrete time diffusion models of products and technologies; these are the Bass model (BM) and the Generalized Bass model (GBM) which have three and six parameters, respectively. We analyze two case studies based on real data, namely Algerian natural gas production and Austrian solar thermal capacity. In this paper we derive and compare Hartley (1964)’s exact sCR with Guseo (1983)’s projection matrix with the ‘approximate’ sCR, in terms of accuracy, via simulation studies. The simulation studies are performed to explore the effect of increasing the parameter dimension with different model structures. Specifically, the BM is considered in a constrained version and in its full form (two and three parameters respectively), while the GBM is considered with two different intervention functions (six parameters in both versions). The simulation studies are performed for lifecycles with the same diffusion characteristics as those of the two case studies.

2 Diffusion models

Let us denote with n the number of observations used to fit the model and with y the n-dimensional vector obtained by stacking the observed y_i values, $i = 1,2,\dots,n$. Similarly, $f(\vartheta;t)$ will denote the vector $f(\vartheta;t) = \{f(\vartheta;t_1), f(\vartheta;t_2),\dots, f(\vartheta;t_n)\}'$.

For a general nonlinear regression model,

$$y_i = f(\vartheta;t_i) + \varepsilon_i, \quad \vartheta \in \mathbb{R}^k, \quad i = 1,2,\dots,n, \tag{1}$$

where $\mathfrak{C}(\Theta) \subset \mathbb{R}^k$ is the exact sCR with confidence level $1-\alpha$.

In the application of diffusion models, the starting point is given by an observed time series reporting sales data or consumption/production of a technological innovation. In this work, we consider the BM and GBM (Bass et al, 1994). Let $z(t)$ be the cumulative data, at time t, and $w(t)$ an intervention function. The GBM is:

$$z(t) = m\frac{1 - e^{-(p+q)\int_0^t w(\tau)d\tau}}{1 + \frac{q}{p}e^{-(p+q)\int_0^t w(\tau)d\tau}}, \tag{2}$$

where m is the market potential, p is the innovation coefficient, q is the imitation coefficient, and $w(t)$ can be any integrable function. Below, we examine the model arising when $w(t)$ is specified by the so-called *exponential* shock (Guseo and Dalla Valle, 2005),

$$w(t) = 1 + c_1 e^{b_1(t-a_1)} I_{t \geq a_1}, \tag{3}$$

which allows us to describe the diffusion of a product for which, at time a_1, we observe a rapid and reversible ($b_1 < 0$) shock with intensity c_1; and when $w(t)$ is specified by the so-called *rectangular* shock,

$$w(t) = 1 + c_1 I_{a_1 \leq t \leq b_1}, \tag{4}$$

which allows us to describe the diffusion of a product for which we observe a constant shock with intensity c_1, in the time interval $[a_1, b_1]$. We will denote the GBM of Eq. (2) with $w(t)$ as in (3) by GBM$_{exp}$, and with $w(t)$ as in (4) by GBM$_{rect}$. These structures are a special case of model (1), where $f(\vartheta;t)$ is represented by $z(t)$ in Eq. (2), with $w(t)$ specified as in (3) or (4). The cumulative time series data (y) can easily be used to jointly estimate all the parameters (m, p, q, a_1, b_1, c_1) of the model using nonlinear least squares. Finally, the BM is the special case of the GBM, when $w(t) = 1$, $t \in \mathbb{R}^+$.

3 Exact and approximate inference

Hartley (1964) proposed a method for evaluating an exact sCR. This method gives the possibility of verifying whether any point in the parameter space Θ belongs to the exact $(1-\alpha)$ level sCR. Hartley (1964)'s exact sCR is

$$\mathfrak{C}(\Theta) = \left\{ \vartheta : \frac{[y - f(\vartheta;t)]' P [y - f(\vartheta;t)]}{[y - f(\vartheta;t)]'[I_n - P][y - f(\vartheta;t)]} \leq \frac{k}{n-k} F_{1-\alpha}(k, n-k) \right\}, \tag{5}$$

where $F_{1-\alpha}(k, n-k)$ is the $1-\alpha$ percentile of a Snedecor's F distribution with k and $n-k$ degrees of freedom. The confidence level is exact if ε's components can be assumed to be independent and distributed according to a Gaussian distribution. In this work, we assume error homoscedasticity with variance σ^2, and we use the projection matrix proposed by Guseo (1983), $P = F(F'F)^{-1}F'$, where F denotes the $n \times k$ matrix obtained by deriving the vector $f(\vartheta;t)$ with respect to the k-dimensional vector ϑ, $F = \frac{\partial f(\vartheta;t)}{\partial \vartheta}$. This assumption about P leads to the exact sCR that is used in this work:

$$\mathfrak{C}(\Theta) = \left\{ \vartheta : \frac{[y - f(\vartheta;t)]' F(F'F)^{-1}F' [y - f(\vartheta;t)]}{[y - f(\vartheta;t)]'[I_n - F(F'F)^{-1}F'][y - f(\vartheta;t)]} \leq \frac{k}{n-k} F_{1-\alpha}(k, n-k) \right\}. \tag{6}$$

The so-called ‘approximate’ sCR, denoted here by $\mathfrak{I}(\Theta)$, is derived by approximating the nonlinear model by a linear Taylor expansion, taking advantage of the asymptotic normality of the estimator. The ‘approximate’ sCR, $\mathfrak{I}(\Theta) \subset \mathbb{R}^k$, is

$$\mathfrak{I}(\Theta) = \left\{ \vartheta : (\vartheta - \hat{\vartheta})' \hat{F}' \hat{F} (\vartheta - \hat{\vartheta}) \leq k s^2 F_{1-\alpha}(k, n-k) \right\}, \tag{7}$$

where $\hat{F} = F(\hat{\vartheta})$, and s^2 is the sample variance. As the linear approximation is valid asymptotically, $\mathfrak{I}(\Theta)$ will have the correct confidence level of $1-\alpha$, asymptotically.

To evaluate $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$, we derived the expression of the components of F for the GBM$_{exp}$, GBM$_{rect}$, and BM, but we omit them here for brevity.

4 Real data analysis

One field that is currently under the public eye is the diffusion of renewable and nonrenewable energy systems. One energy system of each type was chosen in this study, namely Algerian natural gas production, in billion cubic metres (BCM), with annual data from 1970 to 2004 ($n = 35$, source: www.bp.com), and Austrian thermal solar capacity, in MW$_{th}$, with annual data from 1982 to 2008 ($n = 27$, source: www.estif.org). The data are shown in Figure 1.

We compare the $\mathfrak{C}(\Theta)$ of Eq. (6) and $\mathfrak{I}(\Theta)$ of Eq. (7) in both time series, for increasing parameter dimensions. To accomplish this, we have selected three nested models: the constrained BM (with m fixed, thus $k = 2$), the BM ($k = 3$), and GBM ($k = 6$). In this paper, we decide to show $\mathfrak{I}(\Theta)$ and $\mathfrak{C}(\Theta)$ only for the BM $k = 3$. To do that, we used a grid of 12,190,801 points. Each point in the grid has been subsequently tested to assess inclusion in $\mathfrak{C}(\Theta)$ or $\mathfrak{I}(\Theta)$, via conditions (6) and (7), respectively, with $1-\alpha = 0.95$. For the natural gas production, the proportion of common points with respect to $\mathfrak{C}(\Theta)$ is 0.706, while it is 0.7 for $\mathfrak{I}(\Theta)$. For the solar thermal capacity, the proportion of common points with respect to $\mathfrak{C}(\Theta)$ is 0.442, while it is 0.512 for $\mathfrak{I}(\Theta)$. The degree of overlap is smaller in both energy systems compared with what we found with $k = 2$. The representation of these points, for both time series, is shown in Figure 2, together with the representation of a grid covering the parallelepiped generated by combining the marginal confidence intervals of level 0.95, evaluated separately with the Bonferroni method for the three parameters. For both time series, the difference between $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$ is larger than that observed for the case with $k = 2$. In particular, for the solar thermal capacity, the discrepancy between $\mathfrak{I}(\Theta)$ and $\mathfrak{C}(\Theta)$ is much bigger, and the shape of $\mathfrak{C}(\Theta)$ is far from being ellipsoidal.

Moving to the case with $k = 6$, we fitted the GBM$_{exp}$ to the natural gas production and the GBM$_{rect}$ to the solar thermal capacity. Only the fitted values when $k = 6$ are plotted in Figure 1, since the GBMs were found to be the best models, according to the F-test for nested models. For the natural gas production, the proportion of common points with respect to $\mathfrak{C}(\Theta)$ is 0.060, while it is 0.103 for $\mathfrak{I}(\Theta)$. For the solar thermal capacity, the proportion of common points with respect to $\mathfrak{C}(\Theta)$ is 0.333, while it is 0.984 for $\mathfrak{I}(\Theta)$. The degree of overlap is drastically low, denoting a high curvature of the space $f(\vartheta;t)$. Especially in this final case, $\mathfrak{I}(\Theta)$ appears to be extremely small with respect to $\mathfrak{C}(\Theta)$, thereby excluding many values that belong to $\mathfrak{C}(\Theta)$.

5 Simulation study

In this section, $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$ are compared in terms of coverage probability, for the models used in Section 4: the constrained BM, BM, and GBM. For each model, we generated N=1,000 simulated time series, using estimates as true values. Moreover, the length of simulated time series corresponds to the number of real data ($n = 35$ for the natural gas production and $n = 27$ for the solar thermal capacity). In this way, the simulation study investigates the coverage probability of $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$, with data with diffusion characteristics, intervention functions, and stage of the lifecycle corresponding to those of the time series considered in Section 4.

Given each simulated time series, $j = 1,\dots,N$, we evaluated parameter estimates and both sCRs, $\mathfrak{C}(\Theta)_j$ and $\mathfrak{I}(\Theta)_j$. We then tested whether the true values of the parameters used to generate the N time series were included in $\mathfrak{C}(\Theta)_j$ or $\mathfrak{I}(\Theta)_j$. The proportion of $\mathfrak{C}(\Theta)_j$ and $\mathfrak{I}(\Theta)_j$ containing the true values represents the coverage probability. The coverage probabilities of $\mathfrak{I}(\Theta)$ and $\mathfrak{C}(\Theta)$ for the constrained BM ($k = 2$), BM ($k = 3$), and GBM ($k = 6$) are plotted in Figure 3: the difference between $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$ increases with the model complexity, and it is negligible for $k = 2$ for both energy systems. It emerges that the coverage probability of $\mathfrak{C}(\Theta)$ also decreases as the variability $(\sigma/m)^2$ increases, but its decay is less severe than what happens with $\mathfrak{I}(\Theta)$. This is especially true for $k = 6$. In summary, for both the case studies, the degree of overlap of $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$ decreased as k increased, denoting that the parameter space curvature increased with k, as well as that the shape of $\mathfrak{C}(\Theta)$ was progressively farther from being ellipsoidal. We could conclude that $\mathfrak{I}(\Theta)$ can be satisfactorily used with a low parameter dimension, or with a moderate parameter dimension only if the variability is limited. The validity of this result is limited to data with diffusion characteristics similar to those used in this paper. Further research is required to generalize the effect of the lifecycle stage on the coverage probability.

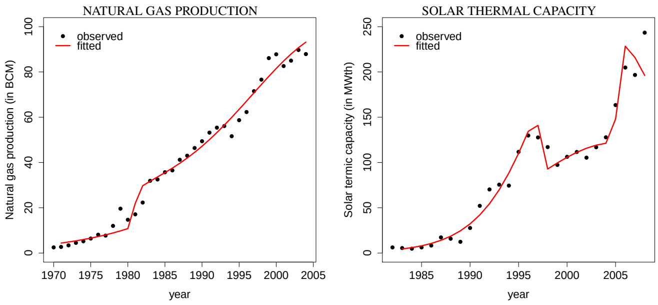


Fig. 1 The lines denote the fitted values (GBM$_{exp}$ and GBM$_{rect}$, respectively).

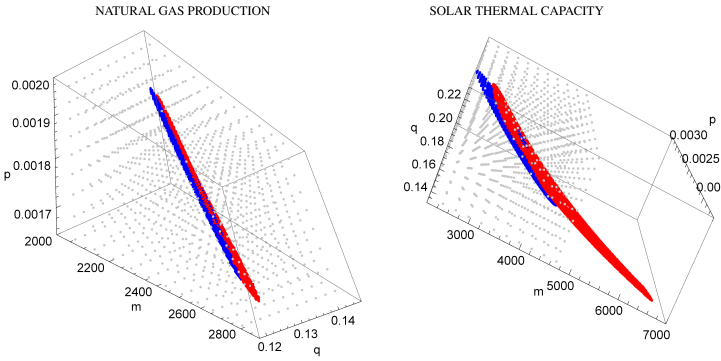


Fig. 2 BM ($k = 3$, $1 - \alpha = 0.95$). $\mathfrak{C}(\Theta)$ is in red and $\mathfrak{I}(\Theta)$ in blue. Grey points represent sCIs.

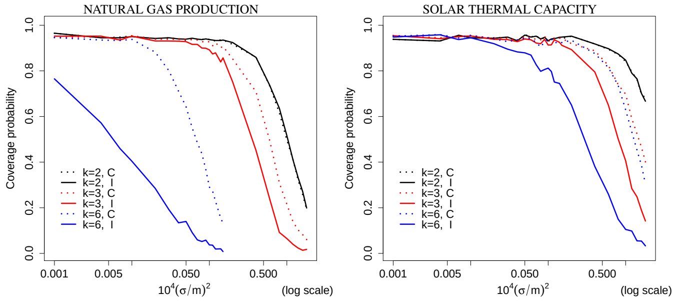


Fig. 3 Coverage probability of $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$. Values of $(\sigma/m)^2$ are in the log scale.

References

Bass FM, Krishnan TV, Jain DC (1994) Why the Bass model fits without decision variables. Marketing Science 13(3):203–223

Demidenko E (2017) Exact and approximate statistical inference for nonlinear regression and the estimating equation approach. Scandinavian Journal of Statistics 44(3):636–665

Guseo R (1983) Confidence regions in non linear regression. Proceedings of the 44th Session of International Statistical Institute, Madrid, 12-22/9/83 pp 333–336

Guseo R, Dalla Valle A (2005) Oil and gas depletion: diffusion models and forecasting under strategic intervention. Statistical Methods and Applications 14(3):375–387

Hartley HO (1964) Exact confidence regions for the parameters in non-linear regression laws. Biometrika 51(3/4):347–353

Lee A, Nyangoma S, Seber G (2002) Confidence regions for multinomial parameters. Computational Statistics & Data Analysis 39(3):329–342

Seber G, Wild C (1989) Nonlinear Regression. Wiley: New York

Tail analysis of a distribution by means of an inequality curve

Analisi della coda di una distribuzione attraverso una curva di concentrazione

E. Taufer, F. Santi, G. Espa and M. M. Dickson

Abstract The Zenga (1984) inequality curve $\lambda(p)$ is constant in p for Type I Pareto distributions. We show that this property holds exactly only for the Pareto distribution and, asymptotically, for distributions with power tail with index α, with $\alpha > 1$. Exploiting these properties one can develop powerful tools to analyze and estimate the tail of a distribution. An estimator for α is discussed. Inference is based on an estimator of $\lambda(p)$ which utilizes all sample information for all values of p. The properties of the proposed estimation strategy is analyzed theoretically and by means of simulations.

Abstract *La curva di concentrazione $\lambda(p)$ di Zenga (1984) è costante in p per le distribuzioni di Pareto di tipo I. Questa proprietà vale esattamente solo per la distribuzione di Pareto e, asintoticamente, per le distribuzioni con indice di coda $-\alpha$, con $\alpha > 1$. Sfruttando queste proprietà si possono sviluppare metodi molto efficaci per analizzare e stimare la coda di una distribuzione. Si discute uno stimatore per α. L'inferenza si basa su uno stimatore di $\lambda(p)$. Le proprietà della strategia di stima proposta sono analizzate teoricamente e mediante simulazioni*

Key words: Tail index, inequality curve, non-parametric estimation

E. Taufer
University of Trento, Trento e-mail: emanuele.taufer@unitn.it

F. Santi
University of Trento, Trento e-mail: flavio.santi@unitn.it

G. Espa
University of Trento, Trento e-mail: giuseppe.espa@unitn.it

M. M. Dickson
University of Padua, e-mail: dickson@stat.unipd.it

1 Introduction

Consider an iid random sample $X_1, X_2, \dots, X_n$ drawn from a random variable with distribution function F satisfying

$$\bar{F}(x) = x^{-\alpha} L(x), \tag{1}$$

where $\bar{F} = 1 - F$, and $L(x)$ is a slowly varying function, that is $L(tx)/L(x) \to 1$ as $x \to \infty$, for any $t > 0$. We will say that $\bar{F}$ is regularly varying (RV) at infinity with index $-\alpha$, denoted as $\bar{F} \in RV_{-\alpha}$. The parameter $\alpha > 0$ is usually referred to as *tail index*; alternatively, in the extreme value (EV) literature it is typical to refer to the EV index $\gamma > 0$ with $\alpha = 1/\gamma$ (see e.g. [10]).

The paper proposes an estimator of the tail index α which relies on Zenga inequality curve $\lambda(p)$, $p \in (0,1)$ [12]. The curve $\lambda(p)$ has the property of being constant for Type I Pareto distributions and, as it will be shown, this property holds for distributions satisfying (1). See [1], [12] , [13] for a general introduction and analysis of $\lambda(p)$.

Probably the most well-known estimator of the tail index is the Hill [6] estimator, which exploits the k upper order statistics. The Hill estimator may suffer from high bias and is heavily dependent on the choice of k (see e.g. [2]). It has been thoroughly studied and several generalization have appeared in the literature. For recent review of estimation procedures for the tail index of a distribution see [3].

The approach to estimation proposed here, directly connected to the inequality curve $\lambda(p)$ has a nice graphical interpretation and could be used to develop graphical tools for tail analysis. Another graph-based method is to be found in [9], which exploits properties of the QQ-plot; while a recent approach based on the asymptotic properties of the partition function, a moment statistic generally employed in the analysis of multi-fractality, has been introduced by [5]; see also [8] which analyzes the real part of the characteristic function at the origin.

2 The curve $\lambda(p)$ and estimation strategy

Let X be a positive random variable with finite mean μ, distribution function F, and probability density f. The inequality curve $\lambda(p)$ is defined as:

$$\lambda(p) = 1 - \frac{\log(1 - Q(F^{-1}(p)))}{\log(1-p)}, \quad 0 < p < 1, \tag{2}$$

where $F^{-1}(p) = \inf\{x\colon F(x) \geq p\}$ is the generalized inverse of F and $Q(x) = \int_0^x t f(t) dt/\mu$ is the first incomplete moment. Q can be defined as a function of p via the Lorenz curve

$$L(p) = Q(F^{-1}(p)) = \frac{1}{\mu} \int_0^p F^{-1}(t) dt. \tag{3}$$

For a Type I Pareto distribution [7, 573 ff.] with

$$F(x) = 1 - (x/x_0)^{-\alpha}, \quad x \geq x_0 \tag{4}$$

it holds that $\lambda(p) = 1/\alpha$, i.e. $\lambda(p)$ is constant in p. This is actually an if-and-only-if result, as we formalize in the following lemma:

Lemma 1. *The curve* $\lambda(p)$ *defined in* (2) *is constant in* p *if, and only if,* F *satisfies* (4).

The following result can also be stated, asymptotically for the case where $\bar{F}$ satisfies (1) as it is stated in the next lemma. For this purpose write

$$\lambda(x) = 1 - \frac{\log(1 - Q(x))}{\log(1 - F(x))}, \tag{5}$$

Lemma 2. *If* $\bar{F}$ *satisfies* (1)*, then* $\lim_{x\to\infty} \lambda(x) = 1/\alpha$.

A tail property of Pareto type I distribution is worth of being noted. Let X be a random variable distributed according to (4) – that is, $X \sim \text{Pareto}(\alpha, x_0)$ –, the following property holds for any $x_1 > x_2 > x_0$:

$$\mathbb{P}[X > x_1 | X > x_2] = \left(\frac{x_1}{x_2}\right)^{-\alpha},$$

hence, the truncated random variable $(X|X > x_2)$ is distributed as $\text{Pareto}(\alpha, x_2)$.

The implications of this property are twofold. Firstly, the truncated random variable is still distributed according to (4), thus Lemma 1 still applies. Secondly, the tail index α is the same both for original and for truncated random variable, thus function $\lambda(p)$ can be used for the estimation of α regardless of the truncation threshold x_2.

The same property we have just outlined holds asymptotically for distribution functions satisfying (1).

Figure 1 reports the empirical curve $\hat{\lambda}(p)$ as a function of p for a Pareto distribution defined by (4) with $\alpha = 2$ and $x_0 = 1$, denoted with $\text{Pareto}(2, 1)$ and a Fréchet distribution with $F(x) = \exp(-x^{-\alpha})$ for $x \geq 0$ and $\alpha = 2$, denoted by Fréchet(2) at different truncation thresholds. Note the remarkably regular behavior or the curves and the closeness to the theoretical form for the Fréchet case already for low levels of truncation.

In this paper the above properties are exploited for devising an estimation method of the tail index α for distributions of class (1).

Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of the sample, $\mathbb{I}_{(A)}$ the indicator function of the event A. To estimate $\lambda(p)$, define the preliminary estimates

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(X_i \leq x)} \qquad Q_n(x) = \frac{\sum_{i=1}^{n} X_i \mathbb{I}_{(X_i \leq x)}}{\sum_{i=1}^{n} X_i} \tag{6}$$

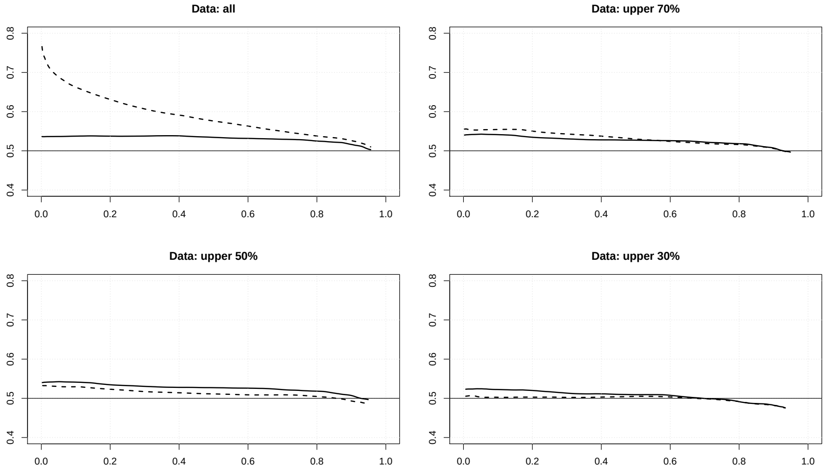


Fig. 1 Plot of $\hat{\lambda}(p)$ and p for Pareto$(2,1)$ (solid line) and Fréchet(2) (dashed line) at various levels of truncation. Sample size $n = 500$. Horizontal line at $1/\alpha = 0.5$

Under the Glivenko-Cantelli theorem (see e.g. [11]) it holds that $F_n(x) \to F(x)$ almost surely and uniformly in $0 < x < \infty$; under the assumption that $E(X) < \infty$, it holds that $Q_n(x) \to Q(x)$ almost surely and uniformly in $0 < x < \infty$. F_n and Q_n are both step functions with jumps at $X_{(1)}, \dots, X_{(n)}$. The jumps of F_n are of size $1/n$ while the jumps of Q_n are of size $X_{(i)}/T$ where $T = \sum_{i=1}^{n} X_{(i)}$. Define the empirical counterpart of L as follows:

$$L_n(p) = Q_n(F_n^{-1}(p)) = \frac{\sum_{j=1}^{i} X_{(j)}}{T}, \quad \frac{i}{n} \le p < \frac{i+1}{n}, \quad i = 1, 2, \dots, n-1, \tag{7}$$

where $F_n^{-1}(p) = \inf\{x : F_n(x) \ge p\}$. To estimate α define

$$\hat{\lambda}_i = 1 - \frac{\log(1 - L_n(p_i))}{\log(1 - p_i)}, \quad p_i = \frac{i}{n}, \quad i = 1, 2, \dots n - \lfloor \sqrt{n} \rfloor. \tag{8}$$

and let $\hat{\alpha} = 1/\bar{\lambda}$ where $\bar{\lambda}$ is the mean of the $\hat{\lambda}_i$'s. The choice of $i = 1, \dots, n - \lfloor \sqrt{n} \rfloor$ guarantees that $\hat{\lambda}_i$ is consistent for λ_i for each $p_i = i/n$ as $n \to \infty$.

3 Simulations

To evaluate the performance of $\hat{\alpha}$, some numerical comparisons are carried out with respect to some reduced bias competitors optimized with respect to the choice of k, the number of largest order statistics used in estimation, as discussed in [4]. The class of moment of order p estimators [4], which reduce to the Hill estimator when $p = 0$ is considered; in the tables they are indicated as $Mop(p)$.

As far as the estimator $\hat{\alpha}$ is concerned, different levels of truncation of the data are considered. In the tables $Ze(q)$ indicates the estimator $\hat{\alpha}$ with q indicating the fraction of upper order statistics used in estimation.

For the comparisons, the Pareto and the Fréchet distributions, as defined in the previous section, are used. Random numbers for the Pareto distribution are simply generated in `R` using the function `runif()` and inversion of F; random numbers from the Fréchet are simulated using the function `rfrechet()` from the library `evd` with shape parameter set equal to α.

Tables 1 and 2 contain the results of simulations. For each sample size $n = 100, 200$, 500, 1000, 2000, $M = 1000$ Monte-Carlo replicates were generated. Computations have been carried out with `R` version 3.3.1 and each experiment, i.e. given a chosen distribution and a chosen n, has been initialized using `set.seed(1)`.

n	*Hill*	$Mop(0.5)$	$Mop(1)$	$Ze(1)$	$Ze(0.7)$	$Ze(0.5)$	$Ze(0.3)$
100	3.41	1.01	1.02	6.44	5.93	5.46	4.68
200	3.97	1.01	1.02	8.67	8.11	7.38	6.53
500	1.99	1.00	0.99	5.33	4.84	4.58	4.07
1000	2.75	1.00	1.00	8.31	7.64	7.11	6.49
2000	1.10	1.00	0.99	3.67	3.48	3.18	2.84

Table 1 Hill estimator: RMSE. Other estimators: relative RMSE *wrt* to the Hill estimator. Pareto$(2,1)$ distribution. Results based on 1000 replications.

n	*Hill*	$Mop(0.5)$	$Mop(1)$	$Ze(1)$	$Ze(0.7)$	$Ze(0.5)$	$Ze(0.3)$
100	0.72	0.99	0.97	1.19	1.41	1.34	1.11
200	0.62	0.99	0.96	1.01	1.26	1.29	1.13
500	0.50	0.98	0.94	0.81	1.07	1.15	1.09
1000	0.44	1.00	0.93	0.71	0.93	1.06	1.06
2000	0.37	1.00	0.91	0.61	0.82	0.94	1.00

Table 2 Hill estimator: RMSE. Other estimators: relative RMSE *wrt* to the Hill estimator. Fréchet(2) distribution. Results based on 1000 replications.

From the tables one can note that the performance of $\hat{\alpha}$ is brilliant for the Pareto and slightly better of Mop estimators for the Fréchet. Truncation seems to have only a small effect on the performance of the estimator.

References

1. Arcagni A., Porro F. (2016) . A comparison of income distributions models throught inequality curves. Statistica & Applicazioni. XIV (2), 123–144
2. Embrechts, P., C. Klüppelberg, T. Mikosch (1997). Modelling Extremal Events. Springer.

3. Gomes, M. I., & Guillou, A. (2015). Extreme value theory and statistics of univariate extremes: a review. *International Statistical Review*, 83(2), 263–292.
4. Gomes, M. I., Brilhante, M. F., & Pestana, D. (2016). New reduced-bias estimators of a positive extreme value index. *Communications in Statistics-Simulation and Computation*, 45(3), 833–862.
5. Grahovac, D., Jia, M., Leonenko, N. N., Taufer, E. (2015) Asymptotic properties of the partition function and applications in tail index inference of heavy-tailed data. *Statistics: A Journal of Theoretical and Applied Statistics* 49, 1221–1242.
6. Hill, B. M. (1975) A simple general approach to inference about the tail of a distribution. *The Annuals of Statistics* 3(5), 1163–1174.
7. Johnson N. L., S. Kotz, N. Balakrishnan (1995) *Continuous Univariate Distributions, Vol. 2*, 2nd ed, Wiley.
8. Jia, M., Taufer, E., Dickson, M. (2018). Semi-parametric regression estimation of the tail index. *Electronic Journal of Statistics* 12, 224–248.
9. Kratz, M. F., Resnick, S. I. (1996) The QQ-estimator and heavy tails. *Comm. Statist. Stochastic Models* 12 (4), 699–724.
10. McNeil, A. J., R. Frey, P. Embrechts (2005) *Quantitative Risk Management*, Princeton University Press.
11. Resnik, S. I. (1999) *A probability path*, Birkhäuser.
12. Zenga, M. (1984). Proposta per un indice di concentrazione basato sui rapporti fra quantili di popolazione e quantili di reddito. *Giornale degli Economisti e Annali di Economia* 5/6, 301–326
13. Zenga M.(1990). Concentration curves and Concentration indexes derived from them. In *Income and Wealth Distribution, Inequality and Poverty*, Dagum, C., Zenga, M. (Ed.), Springher -Verlag, 94–110.

Nonparametric penalized likelihood for density estimation

Stima della densità non-parametrica basata su verosimiglianza penalizzata

Federico Ferraccioli, Laura M. Sangalli and Livio Finos

Abstract In this work we consider a nonparametric likelihood approach to multivariate density estimation with a regularization based on the Laplace operator. The complexity of the estimation problem is tackled by means of a finite element formulation, that allows great flexibility and computational tractability. The model is suitable for any type of bounded planar domain and can be generalized to the non-Euclidean settings. Within this framework, we as well discuss a new approach to clustering based on the concept of diffusion in a potential field, and a permutation-based procedure for one and two samples hypothesis testing.

Abstract *In questo lavoro viene considerato un metodo di stima della densità tramite verosimiglianza non parametrica con un termine di regolarizzazione basato sull'operatore di Laplace. Il problema di stima è risolto attraverso l'uso di una formulazione ad elementi finiti, che assicura elevata flessibilità e trattabilità computazionale. Il modello è adatto per qualsivoglia tipo di dominio planare chiuso e può essere generalizzato al caso non Euclideo. Si propone un approccio al clustering basato sul concetto di diffusione, insieme ad una procedura di test di ipotesi per uno e due campioni basata su permutazione.*

Key words: finite elements, Laplace operator, mode clustering.

Federico Ferraccioli
Dipartimento di Scienze Statistiche, Via Cesare Battisti, 241, 35121 Padova (Italy)
e-mail: ferraccioli@stat.unipd.it

Laura M. Sangalli
MOX-Dipartimento di Matematica, Piazza L. da Vinci, 32, 20133 Milano (Italy)
e-mail: laura.sangalli@polimi.it

Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Via Venezia, 8, 35131 Padova (Italy) e-mail: livio.finos@unipd.it

1 Introduction

The problem of density estimation plays a central role in statistics. It is a fundamental tool for the visualization of structure in exploratory data analysis, and it may be used as intermediate procedure in classification and custering problems. The last decades have seen enormous amount of research focused on kernel density estimation [10]. Simplicity of use and elegant analytic results are the key features of the success of the kernel approach. However, bandwidth selection remains a crucial problem for this method. Moreover, despite recent progress on the asymptotic convergence of the errors, good finite sample performance is by no means guaranteed and many practical challenges remain. The problem get even worse in the multidimensional setting, where the specification of a symmetric, positive definite bandwith matrix is needed, although it's common practice to use diagonal matrices.

Beside the class of kernel density estimators, many other smoothing methods for density estimation have been proposed. All these estimators are based on the idea of reducing the complexity of the problem with some type of approximations or some form of constraint on the space of solutions. In the former case, the approximation is given by basis expasion such as wavelets [4] or splines [9]. In the latter case, the two most prominent approaches are based on regularization of the likelihood functional [8] or shape constraints on the density, e.g. log-concavity [3]. The latter is a parameter-free method but at the cost of a severe restriction on model flexibility. Regularized likelihood methods are extremely flexible but because of the computational complexity they never reached popularity.

In this work we present a new nonparametric likelihood approach to density estimation. The model is based on a finite element formulation and can deal with data distributed over non-regular planar domains. We also briefly discuss a density based clustering method and a permutation based procedure for goodness of fit and for two-samples hypothesis testing.

2 Methodology

2.1 Classical approach

The problem of nonparametric maximum likelihood estimation, in the univariate case, has been considered for the first time in [6]. Let $X_1, \dots, X_n$ i.i.d. observations with distribution function F and density f on a bounded domain $\Omega \in \mathbb{R}$. Without further assumptions, the maximum likelihood estimator for f is not well defined. The likelihood function is unbounded above and the maximization procedure returns the trivial solution of sum of delta functions at the observations. Unlike classical parametric likelihood estimation, where the parameter space is finite, the estimator belongs to an infinite class of functions and some type of regularization becomes necessary to obtain a non-degenerate solution. The basic approach is to maximize a score ω, depending on f and on the observations, defined by

$$\omega = \omega(f) = L - \alpha R(f)\,, \tag{1}$$

where $L = \sum_i \log f(x_i)$ is the log-likelihood, $R(f)$ is the roughness penalty, and the parameter $\alpha > 0$ controls the amount of smoothness. The authors consider, as measure of the roughness or complexity, the functional $R(f) = ||(\sqrt{f})^{(1)}||_2^2$, where the square root permits to avoid the positive contraints on the density. Further developments of this model are presented in [8], where the author considers a regularization functional of the form $R(f) = ||(\log f)^{(3)}||_2^2$. In this case the limiting estimate, as α tends to infinity, is the normal density with the same mean and variance as the data. Note that in this case the positive constraint is avoided by means of the logarithm transformation. Although both models could be generalized to the multivariate setting, consistency results and implementation are given only in the univariate case.

2.2 Model and estimation procedure

In this work we propose a generalization to the estimation of density defined over bounded planar domains. Suppose we observe $X_1, \dots, X_n$ i.i.d. observations drawn from a distribution F on a bounded planar domain $\Omega \in \mathbb{R}^2$. Instead of considering the density f, let us define the log density $g = \log f$, where g is a real function on Ω. This transformation is particularly convenient from the theoretical as well as the practical point of view.
We are interested in a penalized maximum likelihood estimation for g. As previously stated, some types of regularization are necessary, in order to restrict the class of possible solutions. More formally, we consider the estimator that is a solution of the optimization problem

$$\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int_{\Omega} \exp(g(x))\,dx + \lambda R(g) \tag{2}$$

$$\text{subject to} \quad g \in \mathscr{H}^2(\Omega)\,, \tag{3}$$

where $\mathscr{H}^2(\Omega)$ is Sobolev space of functions with continuous weak derivatives up to the second order. As pointed out by [8], the second term of the functional ensures the unitary contraint on the density. We consider here the penalization functional $R(f) = \int_{\Omega} (\Delta \log f)^2\, dx$, where Δ is the Laplace operator. The Laplacian is a measure of local curvature that is invariant with respect to Euclidean transformations of spatial coordinates, and therefore ensures that the concept of smoothness does not depend on the orientation of the coordinate system.
A more complex prior knowledge concerning the domain, which can be translated into a partial differential operator, could be incorporated in the regularization term. We shall consider linear second order elliptic operators of the form

$$Lg = -\mathrm{div}(K\nabla g) + b\nabla g + cf \tag{4}$$

The diffusion term $-\mathrm{div}(K\nabla g)$ induces a smoothing with a preferential direction that corresponds to the first eigenvector of the diffusion tensor K. The degree of anisotropy is controlled by the ratio between its first and second eigenvalue. The transport term $b\nabla g$ induces a smoothing only in the direction specified by the transport vector b. Finally, the reaction term cf has instead a shrinkage effect towards a uniform density on the domain.

Likewise in [7] and [1], the estimation problem is tackled by means of finite element method (FEM), a methodology mainly developed and used in engineering applications, to solve partial differential equations. The strategy of finite element analysis is very similar in spirit to that of univariate splines, and consists of partitioning the problem domain into small disjoint sub-domains and defining polynomial functions on each of these sub-domains in such a way that the union of these pieces closely approximates the solution. Convenient domain partitions are given for instance by triangular meshes. The simplified problem is made computationally tractable by the choice of the basis functions for the space of piecewise polynomials on the domain partition. Each piece of the partition, equipped with the basis functions defined over it, is named a finite element.

Unlike kernel density estimation, the proposed approach admits a likelihood formulation and it's well defined on any domain Ω, without necessity for boundary correction. The absence of constraints on f allows the estimation of extremely complex structures, a fundamental feature in research areas such as density based clustering. Based on the proposed method, we shall in particular discuss a clustering procedure stimulated by Morse theory [2]. We shall also introduce one and two-sample nonparametric tests, based on a permutation approach.

3 Simulation study

Let us consider a complex domain, defined by the closed annulus $ann(a;r,R) = \{x \in \mathbb{R}^2 : r \leq ||x-a|| \leq R\}$, where a is the center and (r,R) the internal and external radii. In our case we consider the annulus centered at the origin, with internal radius 3 and external radius 5. The distribution we have defined on the annulus is the joint probability $\theta \sim \mathrm{Unif}(0,2\pi)$ and a truncated Gaussian distribution in the interval $[1,1]$ with zero mean and standard deviation $\sigma = 0.3$. The Gaussian defines a random distance from the circle with center the origin and radius 4, in the direction normal to the perimeter. This domain includes complex characteristics such as nonlinear boundaries and holes. Despite the presence of complex domains in real data, none of the standard methods in density estimation is appropriate for these problems.

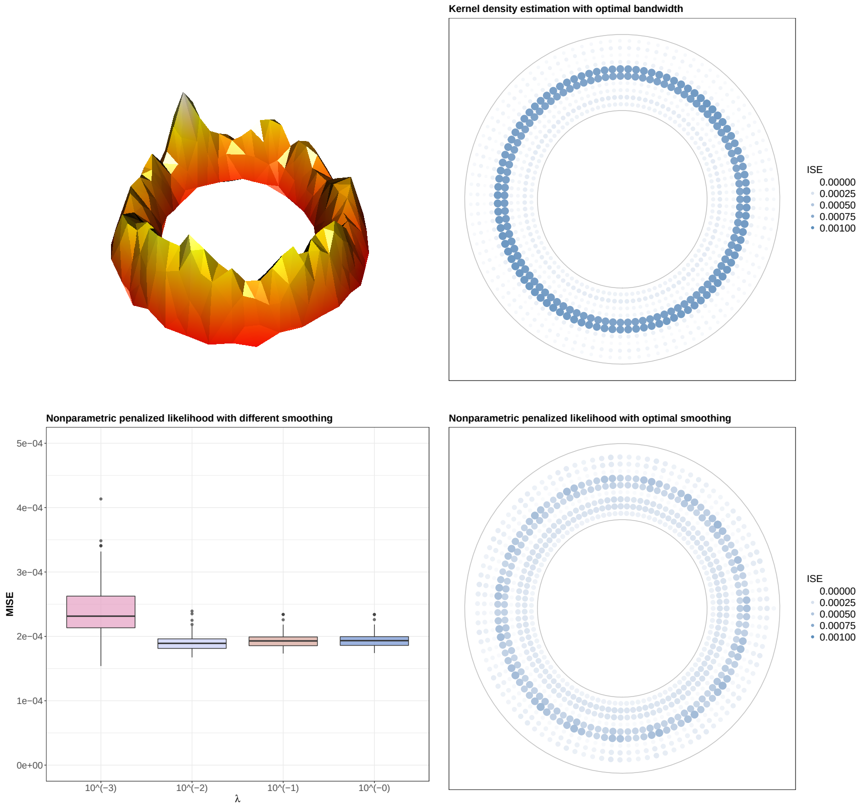


Fig. 1 On the left, an example of estimated density (top) and the distrubution of the MISE of 1000 simulation for different values of the smoothing parameters (bottom). On the right, the MSE surface of 1000 simulation for the nonparametric likelihood (top) and the KDE estimator (bottom), respectively.

4 Model extensions and conclusions

The proposed model performs well with respect to the state of the art of density estimators, while reducing the number of parameters to be selected. The estimator is also well defined on every bounded planar domain. The model can be generalized to non-euclidean setting, e.g. manifolds [5], and a time dependency can be included. These two features are extremely important in applications such as the study of brain activity, where the distribution of the signals over an highly convoluted domain, the cerebral cortex, changes over time. To the best of the authors knowledge, none of the existing methods is appropriate for this type of problems. Clustering procedures and two-samples testing based on the proposed estimator are also presented. Future works will consider convergence of the estimators, consistency in the multivariate case and time-dependent generalizations.

References

[1] Laura Azzimonti et al. “Mixed finite elements for spatial regression with PDE penalization”. In: *SIAM/ASA Journal on Uncertainty Quantification* 2.1 (2014), pp. 305–335.

[2] José E Chacón et al. “A population background for nonparametric density-based clustering”. In: *Statistical Science* 30.4 (2015), pp. 518–532.

[3] Madeleine Cule, Richard Samworth, and Michael Stewart. “Maximum likelihood estimation of a multi-dimensional log-concave density”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.5 (2010), pp. 545–607.

[4] David L Donoho et al. “Density estimation by wavelet thresholding”. In: *The Annals of Statistics* (1996), pp. 508–539.

[5] Bree Ettinger, Simona Perotto, and Laura M Sangalli. “Spatial regression models over two-dimensional manifolds”. In: *Biometrika* 103.1 (2016), pp. 71–88.

[6] IJ Good and RA Gaskins. “Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data”. In: *Journal of the American Statistical Association* 75.369 (1980), pp. 42–56.

[7] Laura M Sangalli, James O Ramsay, and Timothy O Ramsay. “Spatial spline regression models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.4 (2013), pp. 681–703.

[8] Bernard W Silverman. “On the estimation of a probability density function by the maximum penalized likelihood method”. In: *The Annals of Statistics* (1982), pp. 795–810.

[9] Grace Wahba. *Spline models for observational data*. Vol. 59. Siam, 1990.

[10] Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

Rethinking the Kolmogorov-Smirnov Test of Goodness of Fit in a Compositional Way

Il test di bontá di adattamento di Kolmogorov-Smirnov ripensato in chiave composizionale

G.S. Monti, G. Mateu-Figueras, M. I. Ortego, V. Pawlowsky-Glahn and J. J. Egozcue

Abstract The Kolmogorov Smirnov test (KS) is a well known test used to asses how a set of observations is significantly different from the probability model specified under the null hypothesis. The KS test statistic quantifies the distance between the empirical distribution function and the hypothetical one. The modification introduced in Monti et al. (2017) consists of computing the mentioned distances as Aitchison distances. In this contribution, we suggest a further modification of the latter test and investigate, by simulation, the asymptotic distribution of the proposed test statistic, checking the appropriateness of a Generalized Extreme Value (GEV) Distribution. The properties of the asymptotic distribution are studied via Monte Carlo simulations.

Abstract *Il test di Kolmogorov Smirnov (KS) é tra i piú noti test di bontá di adattamento di un modello ai dati. Il test KS é una funzione della distanza tra la distribuzione empirica dei dati e quella ipotizzata sotto l'ipotesi nulla. La modifica del test proposta in Monti et al. (2017) consiste nell'impiego della distanza di Aitchison come misura di tale scostamento. In questo contributo proponiamo una leggera modifica di quest'ultima statistica test, per la quale, attraverso simulazioni Monte Carlo, studieremo la distribuzione asintotica valutando l'accuratezza di una distribuzione generalizzata per valori estremi (GEV).*

Key words: Generalized Extreme Value Distribution, Aitchison distance, Monte Carlo Simulations

G.S. Monti
Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy , e-mail: gianna.monti@unimib.it

G. Mateu-Figueras and V. Pawlowsky-Glahn
Department of Computer Science, Applied Mathematics, and Statistics, University of Girona, Spain

M. I. Ortego and J. J. Egozcue
Department of Civil and Environmental Engineering, Technical University of Catalonia-BarcelonaTECH, Spain

1 Modified Kolmogorov-Smirnov Test

Consider a random sample, denoted $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_n)$, coming from a continuous variable X. Let the hypothesized CDF be $F(x|\theta)$, where θ is the vector of parameters of F. We formulate the hypothesis $H_0 : \; X \sim F(\cdot|\theta)$, against the alternative that the random variable does not follow the claimed distribution.

The Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933) consists of rejecting H_0 when the statistic

$$D_{KS} = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|\,,$$

exceeds a critical value — which depends on the sample size n and on the significance level α — where, for all x, $F_n(x) = \frac{1}{n}\{\text{the number of } X_i\text{'s which are} \leq x\}$ is the empirical distribution function (EDF) of the sample. D_{KS} can be computed calculating first

$$D_{KS}^{+} = \max_{i=1,\ldots,n} \left\{ \frac{i}{n} - F(X_{(i)}) \right\} \quad \text{and} \quad D_{KS}^{-} = \max_{i=1,\ldots,n} \left\{ F(X_{(i)}) - \frac{(i-1)}{n} \right\}, \qquad (1)$$

where $X_{(i)}$ is the ith order statistic; then the KS test statistic is $D_{KS} = \max\left\{D_{KS}^{+}, D_{KS}^{-}\right\}$. The distribution of this statistic is known, even for finite samples (Darling, 1957), and tables are available.

Here we consider a slight variation of the modified KS test statistic, denoted D_a, which has been defined and discussed previously (Monti et al., 2017). D_a consists in replacing the absolute difference between the sample and the hypothetical CDF, with the Aitchison distance (Aitchison, 1983) between two part compositions

$$\begin{aligned} \mathbf{Z}_\ell(i) &= \left(\frac{i}{n+1}, 1 - \frac{i}{n+1}\right) = \left(\frac{i}{n+1}, \frac{n+1-i}{n+1}\right), \\ \mathbf{Z}_u(i) &= \left(\frac{i-1}{n+1}, 1 - \frac{i-1}{n+1}\right) = \left(\frac{i-1}{n+1}, \frac{n+2-i}{n+1}\right), \\ \mathbf{Z}_0(i) &= \left(F(x_{(i)}), 1 - F(x_{(i)})\right), \end{aligned}$$

that is $D_a = \max\left\{D_a^{+}, D_a^{-}\right\}$, where

$$D_a^{+} = \max_{i=1,\ldots,n} \left\{ \mathrm{d}_a\Big(\mathbf{Z}_\ell(i), \mathbf{Z}_0(i))\Big) \right\}, \quad D_a^{-} = \max_{i=1,\ldots,n} \left\{ \mathrm{d}_a\Big(\mathbf{Z}_0(i), \mathbf{Z}_u(i)\Big) \right\}. \qquad (2)$$

Whereas in the previous version we considered the ratios $\frac{i}{n}$ in formula (2), in this version we adopt the median rank or the Weibull plotting position which are slightly more accurate than mean ranks.

D_a is motivated by the fact that probabilities, like for instance $i/(n+1)$ and $F(x_{(i)})$ as well as $(i/(n+1), 1 - i/(n+1))$ and $(F(x_{(i)}), 1 - F(x_{(i)}))$, can be considered as two part compositions, and then the Aitchison distance (Aitchison, 1983; Aitchison et al., 2001) can be adopted as a natural similarity measure. We recall that for 2-part compositions, $\mathbf{p}_1 = (p_1, 1 - p_1)$ and $\mathbf{p}_2 = (p_2, 1 - p_2)$, the Aitchison square distance between them is

$$\mathrm{d}_a^2(\mathbf{p}_1,\mathbf{p}_2) = \left(\frac{1}{\sqrt{2}} \ln \frac{p_1}{1-p_1} - \frac{1}{\sqrt{2}} \ln \frac{p_2}{1-p_2} \right)^2 .$$

It has been shown that D_a, as a test statistic, is invariant under a reversion of the orientation of the axis of the data (Monti et al., 2017).

Supported by a large number of Monte Carlo simulations, in Section 2 it will be shown that D_a follows reasonably well a Generalized Extreme Value Distribution (GEVD) for maxima and its location and scale parameters depend approximately on the sample size.

Recall that a random variable Z has a GEVD if its probability function can be written as

$$F_Z(z|\mu,\sigma,\xi) = \exp\left[-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi} \right] , \; 1+\frac{\xi}{\sigma}(z-\mu) > 0 \, , \qquad (3)$$

where $\mu \in \mathbb{R}$ is a location parameter, $\sigma > 0$ is a scale parameter, and $\xi \in \mathbb{R}$ is a shape parameter. The values of the shape parameter ξ define the three families of asymptotic distribution: type II for $\xi > 0$, type III for $\xi < 0$ and Gumbel in the limiting case $\xi = 0$ in this parameterization (Fisher and Tippett, 1928; Embrechts et al., 1997).

2 Simulation results

In order to assess the accuracy of the GEV model to the D_a statistic defined in (2), we have conducted an intensive Monte Carlo (MC) simulation.

For each reference model – Normal, Uniform, Gamma, Beta, Exponential and lognormal with random parameters, i.e. we consider only the all-parameters-known case – and for each sample size – 1,000 different sample size values, ranging from 5 to 50,000 – we have simulated 1,000 random samples. For each simulated sample we have computed the D_a statistic in order to test the goodness of fit of the theoretical distribution. All the computations were carried out using the R statistical software program (R Core Team, 2017).

For each reference model and for each fixed sample size we have estimated the parameters of the Gumbel distribution, a subfamily of the GEV for $\xi = 0$, and of the GEV model for the $1,000$ D_a values by maximum likelihood method.

Two linear regression models and three linear regression models of the 1,000 MC estimates of the Gumbel, μ (location) and σ (scale), and GEV parameters, μ (location), σ (scale) and ξ (shape), were estimated as a function of the log-size of the sample. The regression outputs are summarized in Table 1, which reports estimates, standards errors and p-values.

Table 1 Regression output for the different linear regression models.

Reference distribution: Normal			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1772 (0.0088; 0.0000)	0.7975 (0.0009; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7206 (0.0054; 0.0000)	-0.0009 (0.0006; 0.0911)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1817 (0.0092; 0.0000)	0.7973 (0.0009; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7226 (0.0057; 0.0000)	-0.0011 (0.0006; 0.0623)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0117 (0.0066; 0.076)	0.0008 (0.0007; 0.235)
Reference distribution: Uniform			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1782 (0.0095; 0.0000)	0.7974 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7184 (0.0054; 0.0000)	-0.0008 (0.0006; 0.165)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1851 (0.01; 0.0000)	0.7968 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7219 (0.006; 0.0000)	-0.0011 (0.0006; 0.058)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0174 (0.0066; 0.0089)	0.0015 (0.0007; 0.028)
Reference distribution: Gamma			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1798 (0.0091; 0.0000)	0.7971 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7205 (0.0053; 0.0000)	-0.001 (0.0005; 0.0656)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1858 (0.0094; 0.0000)	0.7966 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7233 (0.0055; 0.0000)	-0.0012 (0.0006; 0.0275)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0155 (0.0067; 0.0206)	0.0013 (0.0007; 0.0639)
Reference distribution: Beta			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1366 (0.0126; 0.0000)	0.8018 (0.0013; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.718 (0.0064; 0.0000)	-0.0007 (0.0007; 0.292)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1437 (0.0132; 0.0000)	0.8013 (0.0014; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7217 (0.0067; 0.0000)	-0.001 (0.0007; 0.148)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0183 (0.0082; 0.0262)	0.0014 (0.0009; 0.0982)
Reference distribution: Exponential			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1704 (0.0093; 0.0000)	0.7982 (0.0009; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7112 (0.0055; 0.0000)	-0.0001 (0.0006; 0.861)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1706 (0.0097; 0.0000)	0.7983 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.711 (0.0058; 0.0000)	-0.0001 (0.0006; 0.919)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0003 (0.0071; 0.964)	-0.0003 (0.0007; 0.726)
Reference distribution: lognormal			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1459 (0.0134; 0.0000)	0.8007 (0.0014; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7106 (0.007; 0.0000)	-0.00002 (0.0007; 0.975)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1507 (0.0141; 0.0000)	0.8003 (0.0014; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7128 (0.0074; 0.0000)	-0.0002 (0.0008; 0.809)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0121 (0.0088; 0.169)	0.0008 (0.0009; 0.368)

Likelihood ratio tests have been used to compare the two nested models for all simulation settings, and the proportions of simulated p-values less than 0.05 are reported in Table 2.

Looking at the simulations results we can deduce that the D_a statistic follows a Gumbel distribution, whose location parameter μ is related to the logarithm of the

Reference Model	#p-values < 0.05/1000
Normal	0.046
Uniform	0.048
Gamma	0.049
Beta	0.052
Exp	0.067
lognormal	0.048

Table 2 Proportions of simulated p-values less than 0.05 for comparisons of Gumbel and GEV models via asymptotic likelihood ratio tests for each reference distribution.

sample size by a linear relationship. Furthermore, the estimated parameter values are stable with rather small variations among models.

To complete the work, a further Monte Carlo investigation was made on the size (type I error) and on the power of the test. 2,000 samples of fixed size $n = 10, 50, 100, 200, 500, 1000, 1500, 2000, 5000, 10000$, were drawn from each of several distributions. Figure 1 reports six different plots. In the first column the probability of rejecting the null hypothesis using the D_a statistic considering three underlying distributions are reported. The second column reports the probability of rejecting hypothesis $H_0 : X \sim N(1,4)$ against $H_1 : X \sim T(2)$ using the D_a statistic (case (a)); $H_0 : X \sim Ga(2,3)$ against $H_1 : X \sim Exp(2)$ (case (b)) and $H_0 : X \sim Unif(0,1)$ against $H_1 : X \sim Exp(2)$ (case (c)).

References

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika 70*(1), 57–65.

Aitchison, J., C. Barceló-Vidal, A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Reply to letter to the editor by S. Rehder and U. Zier on Logratio analysis and compositional distance. *Mathematical Geology 33*(7), 849–860.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises Tests. *The Annals of Mathematical Statistics 28*(4), 823–838.

Embrechts, P., T. Mikosch, and C. Klüppelberg (1997). *Modelling Extremal Events: For Insurance and Finance*. London, UK, UK: Springer-Verlag.

Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society 24*(2), 180190.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari 4*, 83–91.

Monti, G. S., G. Mateu-Figueras, M. I. Ortego, V. Pawlowsky-Glahn, and J. J. Egozcue (2017). Modified Kolmogorov-Smirnov test of goodness of fit. In K. Hron and R. Tolosana-Delgado (Eds.), *Proceedings of CoDaWork 2017*, pp. 151–158. CoDA, http://www.coda-association.org/en/.

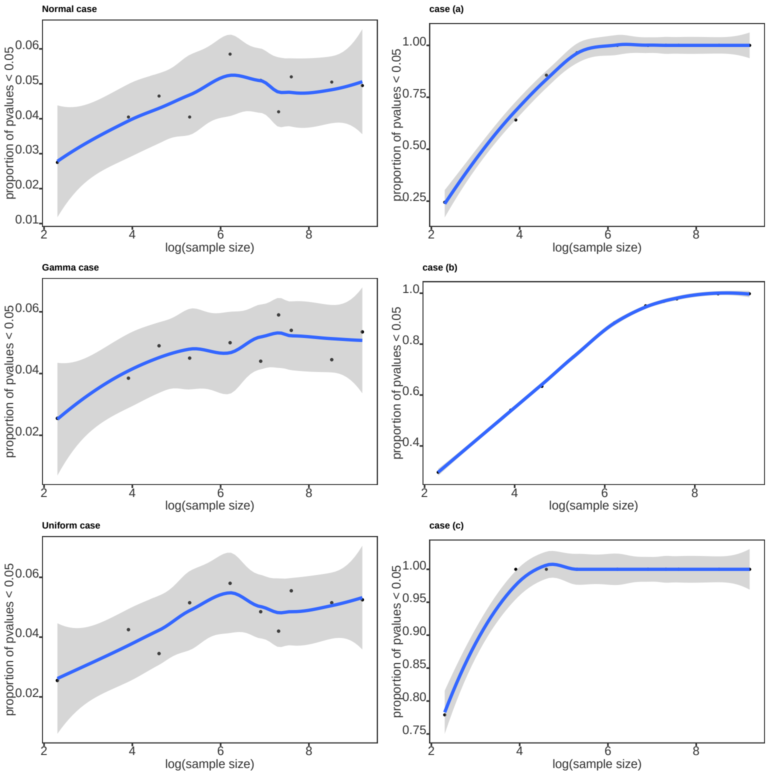


Fig. 1 MC results for probability of type I error (first column) and power of the test (second column). The blu lines represent a smoothing spline fitted to the data.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Stochastic Dominance for Generalized Parametric Families

Dominanza Stocastica per Famiglie Parametriche Generalizzate

Tommaso Lando and Lucio Bertoli-Barsotti

Abstract The T-X family is a recent method for generating distributions by composing probability distributions and quantile functions. Such an approach makes it possible to obtain a large number of flexible families of parametric distributions, new or already existing, most of which are typically used to model phenomena in different areas, such as economics and finance. We present a general method to derive sufficient conditions for the second-order stochastic dominance, within T-X families of distributions.

Abstract *La famiglia T-X permette di generare distribuzioni di probabilità tramite la composizione di funzioni di ripartizione e funzioni quantile. Tale approccio permette di ottenere un gran numero di famiglie parametriche molto flessibili, nuove o già esistenti, gran parte delle quali sono utilizzate per descrivere fenomeni di tipo economico, finanziario, e non solo. Si presenta un metodo generale che permette di ricavare condizioni sufficienti per la dominanza del secondo ordine all'interno di famiglie T-X di distribuzioni.*

Key words: stochastic dominance, T-X family, generalized distributions

1. Introduction

[1] Tommaso Lando, Università degli studi di Bergamo; email: tommaso.lando@unibg.it

Lucio Bertoli-Barsotti, Università degli studi di Bergamo; email: lucio.bertoli-barsotti@unibg.it

Generalized parametric distributions have a wide range of application in several different scientific fields, due to the flexibility offered by a quite high number of parameters. Ranking generalized distribution with respect to a dominance relation represents a major issue in different areas, such as economics and finance (see e.g. Wilfling 1996, Kleiber and Kotz, 2003). On the other hand, it is generally difficult to rank generalized models with stochastic dominance rules, because their functional forms are generally not easily tractable.

In this framework, the so-called T-X family, recently introduced by Alzatreeh et al. (2013), provides an interesting method for generating distributions by composing a "baseline" distribution with a quantile function and finally a "transformer" distribution (see Aljarrah et al. 2014). This approach can be used to generate new or already existing families of distributions. It can be shown (Lando and Bertoli-Barsotti) that, by studying the composition of the T-X family, it is possible to derive sufficient conditions for the first order stochastic dominance (FSD) and the second order stochastic dominance (SSD). In particular, under some circumstances, a pair of T-X models still preserve the SSD order if the corresponding baseline and transformer distributions are ranked by FSD and SSD, respectively.

2. Preliminaries

In this paper, we refer only to continuous random variables (RVs). We denote with CDF the cumulative distribution function and with PDF the probability density function. Therefore, a RV U has CDF F_U and PDF f_U. We recall the basic definitions of the first order stochastic dominance (FSD) and the SSD.

Definition 1. We say that U_1 FSD dominates U_2 and write $U_1 \geq_1 U_2$ iff

$$F_{U_1}(u) \leq F_{U_2}(u), \forall u \in \mathbb{R}$$

Definition 2. We say that U_1 SSD dominates U_2 and write $U_1 \geq_2 U_2$ iff

$$\int_{-\infty}^{u} F_{U_1}(t)dt \leq \int_{-\infty}^{u} F_{U_2}(t)dt, \forall u \in \mathbb{R}$$

It is clear that FSD holds iff CDFs do not cross. Moreover, SSD can be related to the number of crossings between CDFs or PDFs. In particular, when the integral conditions of Def. 2 is difficult to verify for some parametric distributions, we may use an alternative method for deriving sufficient conditions for SSD, which requires CDFs to cross (at most) once (see Hanoch and Lèvi, 1969). Moreover, when it is not possible to verify the crossing condition on CDFs (e.g. when a closed form expression for the CDF is not available), we can rely on some closely related results, which involve densities. In particular, it is sufficient to prove that PDFs cross at most twice (Shaked 1982; see also Ramos et al. 2000, for some related conditions for non-negative RVs). Let us denote with $S^{-}(h)$ the number of sign changes of a function h. We summarize some important results in the theorem below.

Theorem 1.

Let Z_1 and Z_2 have finite means.

1) If $S^{-}(F_{U_1} - F_{U_2}) = 1$ and the sign sequence is $-,+$, then $U_1 \geq_2 U_2$ iff $E(U_1) \geq E(U_2)$.
2) Let $S^{-}(f_{U_1} - f_{U_2}) = 2$ with sign sequence $-,+,-$ Then, $U_1 \geq_2 U_2$ iff $E(U_1) \geq E(U_2)$.
3) Let U_1, U_2 be non-negative RVs and let $f_{U_1}(z)/f_{U_2}(z)$ be unimodal, where the mode is a supremum. Then, $E(U_1) \geq E(U_2)$ implies $U_1 \geq_2 U_2$.

The T-X{Y} method, originally introduced by Alzaatreh et al. (2013) and then studied by also Aljarrah et al. (2014), is based on the composition of the CDFs of two RVs, X and T, with the quantile function (QF) of a third RV Y. Given three RVs X, Y and T, where Y and T must have the same support, a new RV Z is defined by means of its CDF

$$F_Z(z) = F_T{}^{\circ}Q_Y{}^{\circ}F_X(z) = \int_{-\infty}^{Q_Y(F_X(z))} dF_T(t), \tag{1}$$

where Q_Y is the QF of Y. The corresponding PDF is

$$f_Z(z) = \left\{\frac{d}{dz} Q_Y{}^{\circ}F_X(z)\right\} f_T\{Q_Y{}^{\circ}F_X(z)\} = \frac{f_X(z)}{f_Y\{Q_Y{}^{\circ}F_X(z)\}} f_T\{Q_Y{}^{\circ}F_X(z)\}, \tag{2}$$

where f_X, f_Y, f_T are the PDFs of X, Y, T, respectively (Aljarrah et al. 2014). In this formula, F_T plays the role of the generator distribution (transformer) and F_X represents a baseline distribution (transformed).

Many continuous RVs have closed-form expressions for the QF, than can be used as the RV Y in (1), to generate T-X{Y} families. For instance, the original paper of Alzaatreh et al. (2013) focuses on the T-X{exponential} family, which is obtained by taking Q_Y to be the QF of an exponential RV with scale parameter equal to 1, i.e.:

$$Q_Y(p) = -\ln(1-p). \tag{3}$$

The T-X{Y} family makes it possible to generate a large number of new families of distributions, as well as many existing parametric models of noticeable practical relevance because of their several applications, such as: the generalized beta of the first and the second kind, and the generalized gamma distributions (McDonald 1984).

3. Sufficient conditions for SSD

Because many existing parametric distributions belong to the T-X family, we are concerned with finding the sufficient conditions for ranking distributions of such family with FSD and SSD (in particular). It can be shown (Lando ans Bertoli-Barsotti) T-X families obtained by composition of CDFs and QFs ranked by FSD or SSD preserve some kind of order.

In this study, we are interested in studying dominance relations among pairs of distributions within the same T-X family. Put otherwise, we compare pairs of distributions with CDF given by (1), but with different parameters. In particular, we assume that X and T, taken individually, are parametric families of distributions, say,

$$F_X(x) = F_X(x, \boldsymbol{\pi}), F_T(t) = F_T(t, \boldsymbol{\lambda}), \tag{3}$$

Thus, the new distribution defined by (1) depends on the parameters of F_X and F_T:

$$F_Z(z) = F_Z(z, \boldsymbol{\pi}, \boldsymbol{\lambda}). \tag{4}$$

We aim at comparing the RVs Z_1 and Z_2, where, for $i = 1,2$:

$$F_{Z_i}(z) = F_Z(z, \boldsymbol{\pi}_i, \boldsymbol{\lambda}_i) = F_{T_i}{}^\circ Q_{\mathrm{Y}}{}^\circ F_{X_i}(z), \tag{5}$$

with $F_{X_i}(x) = F_X(x, \boldsymbol{\pi}_i)$ and $F_{T_i}(t) = F_{T_i}(t, \boldsymbol{\lambda}_i)$.

Theorem 3. Let $F_{Z_i}(z) = F_{T_i}{}^\circ Q_Y{}^\circ F_{X_i}(z)$, for $i = 1,2$. Let Z_1 and Z_2 have finite means, and let $Q_Y{}^\circ F_{X_2}$ be convex . If $X_1 \geq_1 X_2$, and $T_1 \geq_2 T_2$ with $S^-(F_{T_1} - F_{T_2}) \leq 1$, then $Z_1 \geq_2 Z_2$.

Proof. This theorem has been proved by Lando and Bertoli-Barsotti (submitted manuscript).

It can be shown that there is a wide class of distributions that can be ranked using the sufficient condition of Theorem 2, although, in some cases, the proposed method is not applicable, because, for some T-X families, $Q_Y{}^\circ F_{X_2}$ is not convex.

Theorem 2 establishes that a strong dominance (FSD) between baseline distributions and a weak dominance (SSD) between generators may be sufficient for the weak dominance among the T-X family. Now, it is also worth noting that, for $Q_{\mathrm{Y}}{}^\circ F_{X_2}$ convex, $X_1 \geq_2 X_2$, with single-crossing CDFs, and $T_1 \geq_1 T_2$ do not imply $Z_1 \geq_2 Z_2$. This can be shown with a counter-example. A fortiori, $X_1 \geq_2 X_2$ and $T_1 \geq_2 T_2$ do not imply $Z_1 \geq_2 Z_2$ as well. Therefore, if we wish to rank Z_1, Z_2 by SSD with this method, it is generally required that the baseline distributions are ranked by FSD.

4. Conclusions

This study is aimed at deriving sufficient conditions for stochastic dominance within the T-X family of distributions. This approach can be extended to other types of dominance. In particular, we shall analyse the interesting case of the Lorenz order, which is especially relevant in the field of economics.

Funding
The research was supported through the Czech Science Foundation (GACR) under project 17-23411Y (to T.L.).

References

1. Alexander, C., Cordeiro, G. M., Ortega, E. M., Sarabia, J. M.: Generalized beta-generated distributions. Computational Statistics & Data Analysis, 56(6), 1880-1897 (2012).
2. Aljarrah, M. A., Lee, C., Famoye, F.: On generating T-X family of distributions using quantile functions. Journal of Statistical Distributions and Applications, 1(1), 2 (2014).
3. Alzaatreh, A., Lee, C., Famoye, F.: A new method for generating families of continuous distributions. Metron, 71(1), 63-79 (2013).
4. Hanoch, G., Levy, H.: The efficiency analysis of choices involving risk. The Review of Economic Studies, 36(3), 335-346 (1969).
5. Kleiber, C., Kotz, S.: Statistical size distributions in economics and actuarial sciences (Vol. 470). John Wiley & Sons (2003).
6. Lando, T. & Bertoli-Barsotti, L. Stochastic dominance relations for T-X families of distributions. Submitted manuscript.
7. McDonald, J. B.: Some generalized functions for the size distribution of income. Econometrica: Journal of the Econometric Society, 647-663 (1984).
8. Ramos, H. M., Ollero, J., Sordo, M. A.: A sufficient condition for generalized Lorenz order. Journal of Economic Theory, 90(2), 286-292 (2000).
9. Shaked, M.: Dispersive ordering of distributions. Journal of Applied Probability, 19(2), 310-320 (1982).
10. Wilfling, B.: Lorenz ordering of generalized beta-II income distributions. Journal of Econometrics, 71(1), 381-388 (1996).

Statistical Models for Ordinal Data

A comparative study of benchmarking procedures for interrater and intrarater agreement studies

Valutazione comparativa di procedure di benchmarking per l'analisi dell'accordo inter e intra valutatore

Amalia Vanacore[1] and Maria Sole Pellegrino[2]

Abstract Decision making processes typically rely on subjective evaluations provided by human raters. In the absence of a gold standard against which check evaluation trueness, the magnitude of inter/intra-rater agreement coefficients is commonly interpreted as a measure of the rater's evaluative performance. In this study some benchmarking procedures for characterizing the extent of agreement are discussed and compared via a Monte Carlo simulation.

Abstract *In numerosi contesti, le decisioni strategiche sono affidate a valutazioni soggettive, fornite da valutatori umani, per le quali non esiste un gold standard che permetta di valutarne la veridicita'. L'affidabilita' del valutatore viene quindi spesso misurata in termini di precisione attraverso coefficienti di accordo inter- e intra-valutatore, che risultanto utili se interpretabili. Nel lavoro proponiamo uno studio Monte Carlo per analizzare e confrontare le prestazioni di alcune procedure di benchmarking.*

Key words: rater agreement, kappa-type coefficient, benchmarking procedures, Monte Carlo simulation

1 Introduction

Agreement coefficients are widely adopted for assessing the precision of subjective evaluations provided by human raters to support strategic and operational decisions in several fields (e.g. manufacturing and service industries, food, healthcare and risk management). Specifically, the agreement between the evaluations provided on the same sample of items by two or more raters (i.e. inter-rater agreement) or by

[1] Dept. of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples; email: amalia.vanacore@unina.it

[2] Dept. of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples; e-mail: mariasole.pellegrino@unina.it

the same rater in two or more occasions (i.e. intra-rater agreement) is commonly measured using a kappa-type agreement coefficient.

In order to qualify the extent of agreement as good or poor the computed coefficient is compared against an arbitrary benchmark scale. However, the magnitude of an agreement coefficient may strongly depend on some experimental factors such as number of rated items, rating scale dimension, trait prevalence and marginal probabilities [13, 9]. Thus, interpretation based on the straightforward benchmarking should be treated with caution especially for comparison across studies when experimental conditions are not the same.

A proper characterization of the extent of rater agreement should rely upon a benchmarking procedure that allows to identify a suitable neighborhood of the true value of rater agreement by taking into account sampling uncertainty. The most simple and intuitive way to accomplish this task is by building a confidence interval of the agreement coefficient and comparing its lower bound against an adopted benchmark scale. A different approach is the one recently proposed by Gwet [9] which, under the assumption of asymptotically normal distribution, evaluates the likelihood that the estimated agreement coefficient belongs to each benchmark category.

The above benchmarking approaches will be in the following discussed and their performances will be evaluated and compared in terms of weighted misclassification rate via a Monte Carlo simulation study.

The remainder of the paper is organized as follows: in Section 2 two well-known paradox-resistant kappa-type agreement coefficients are discussed; the commonly adopted benchmark scales and some characterization procedures based on parametric and non-parametric approaches to benchmarking are presented and discussed in Section 3; in Section 4 the simulation design is described and the main simulation results are fully discussed; finally, conclusions are summarized in Section 5.

2 Paradox-resistant agreement coefficients

The kappa-type agreement coefficients are rescaled measures of the observed agreement corrected with the probability of agreement expected by chance. The most common kappa-type coefficient is that proposed by Cohen [5]. Despite its popularity, it is affected by two paradoxes [4]: the degree to which raters disagree (bias problem) and the marginal distribution of the evaluations independently provided by each rater (prevalence problem). A solution to face the above paradoxes is to adopt the uniform distribution for chance measurements, which given a certain rating scale can be defended as representing the maximally non-informative measurement system [6].

Specifically, let n be the number of items rated by two raters on an ordinal k-point rating scale (with $k > 2$), n_{ij} the number of items classified into i^{th} category by the first rater and into j^{th} category by the second rater and w_{ij} the corresponding symmetrical weight, introduced in order to consider that, on an ordinal rating scale, disagreement on two distant categories is more serious than disagreement on neighboring categories. The weighted version of the uniform kappa, often referred to as Brennan-Prediger coefficient [9], is formulated as:

$$\widehat{BP}_w = \frac{p_{a_w} - p_{a|c}^{BP_w}}{1 - p_{a|c}^{BP_w}} \tag{1}$$

where p_{a_w}, the weighted observed proportion of agreement, and $p_{a|c}^{BP_w}$, the weighted proportion of agreement expected under the assumption of uniform chance measurements, are respectively given by:

$$p_{a_w} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\frac{n_{ij}}{n}; \quad p_{a|c}^{BP_w} = \frac{T_w}{k^2} \tag{2}$$

being T_w the sum over all weight values w_{ij}.

Another well-known paradox-resistant agreement coefficient alternative to Cohen's Kappa is the AC_1 coefficient proposed by Gwet [8], whose weighted version AC_2 [9] is formulated as:

$$\widehat{AC}_2 = \frac{p_{a_w} - p_{a|c}^{AC_2}}{1 - p_{a|c}^{AC_2}} \tag{3}$$

where the probability of chance agreement $p_{a|c}^{AC_2}$ is given by:

$$p_{a|c}^{AC_2} = \frac{T_w}{k(k-1)} \cdot \sum_{i=1}^{k} \pi_i(1-\pi_i) \tag{4}$$

Specifically, $p_{a|c}^{AC_2}$ is defined as the probability of the simultaneous occurrence of two events, one rater provides random rating (R) and the two raters agree (G):

$$p_{a|c}^{AC_2} = P(G \cap R) = P(G|R) \cdot P(R) \tag{5}$$

where $P(G|R) = T_w/k^2$ and $P(R)$ is approximated with a normalized measure of randomness defined by the ratio of the observed variance to the variance expected under the assumption of totally random ratings:

$$P(R) = \frac{\sum_{i=1}^{k} p_i(1-p_i)}{(k-1)/k} \tag{6}$$

with p_i denoting the propensity that a rater assigns score i to an item which is estimated by $p_i = (n_{i\cdot} + n_{\cdot i})/2n$ being $n_{i\cdot}$ (resp. $n_{\cdot i}$) the total number of items classified into i^{th} category by the first (resp. second) rater.

3 Benchmarking procedures for characterizing the extent of agreement

After computing an agreement coefficient, a common question is ”how good is the extent of agreement?” As a general rule kappa values greater than 0.6 are generally considered acceptable [10]. In order to provide an aid to qualify the magnitude of kappa-type coefficients, a number of benchmark scales have been proposed mainly in social and medical sciences over the years. The scale proposed by Landis and Koch [11] is by far the most widely adopted benchmark scale; it consists of six ranges of values corresponding to as many categories of agreement: poor, slight, fair, moderate, substantial and almost perfect agreement for coefficient values ranging between -1 and 0, 0 and 0.2, 0.21 and 0.4, 0.41 and 0.6, 0.61 and 0.8 and 0.81 and 1.0, respectively. This scale was then simplified by Fleiss [7] and Altman [1], with three and five ranges, respectively, and by Shrout [12] who collapsed the first three ranges of values into two agreement categories.

Despite its popularity, the straightforward benchmarking can be misleading because it does not associate the interpretation of the extent of agreement with a degree of uncertainty and it does not allow to compare the extent of agreement across different studies, unless they are carried out under the same experimental conditions. In order to have a fair characterization of the extent of rater agreement, it is necessary to associate a degree of uncertainty to the interpretation of the coefficient.

Under asymptotic conditions, the magnitude of the kappa type coefficient can be related to the notion of extent of agreement by benchmarking the lower bound of its asymptotic $(1-2\alpha)\%$ confidence interval (CI). Recently, Gwet [9] proposed a parametric benchmarking procedure based on Interval Membership Probability (IMP) that is the probability that the coefficient falls into each benchmark category.

Under non-asymptotic conditions, two non-parametric CIs based on bootstrap resampling are the percentile (p) CI and, for severely skewed distribution, the Bias-Corrected and Accelerated (BCa) CI [2]. Being free from distributional assumptions, the benchmarking procedure based on bootstrap CIs fits also the cases of moderate and small sample sizes.

4 Simulation study

The above-discussed benchmarking procedures have been applied to characterize the extent of both BP_w and AC_2 across 72 different settings. Their statistical properties have been investigated via a Monte Carlo simulation study developed considering two raters classifying $n = 10, 30, 50, 100$ items into one of $k = 2, 5, 7$ possible ordinal rating categories. The data have been simulated by sampling $r = 2000$ Monte Carlo data sets from a multinomial distribution with parameters n and $\mathbf{p} = (\pi_{11}, \ldots, \pi_{ij}, \ldots, \pi_{ik})$; the π_{ij} values have been set so as to obtain six true population values of agreement (viz. 0.4, 0.5, 0.6, 0.7, 0.8, 0.9), assuming a linear weighting scheme [3].

The performances of the benchmarking procedures under comparison have been evaluated in terms of weighted misclassification rate (hereafter, $\mathbf{M}_{\mathrm{w}}$). Specifically, let $\{X_h; r\}$ be a Monte Carlo data set containing r benchmarks X_h obtained for a population value taken as reference for a specific agreement category ω. $\mathbf{M}_{\mathrm{w}}$ is evaluated as the weighted proportion of misclassified X_h:

$$\mathbf{M}_{\mathrm{w}} = \frac{1}{r} \sum_{\omega=1,\Omega} w_{\omega\omega'} \cdot I\left[X_{h|\omega} \in \omega'\right]; \quad \omega' \neq \omega \tag{7}$$

where $I[\cdot]$ is an indicator taking value 1 if the argument is true and 0 otherwise and $w_{\omega\omega'}$ is a linear misclassification weight adopted to account that on an ordinal benchmarking scale some misclassifications are more serious than others. The best and worst values of $\mathbf{M}_{\mathrm{w}}$ obtained across the analysed benchmarking procedures for BP_w and AC_2 are reported in Table 1 for each combination of n and k values. Specifically, while the benchmarking procedure based on bootstrap CIs are suitable for all the analysed sample sizes, the parametric procedures work only under asymptotic conditions being thus applied only to large samples of $n \geq 50$; therefore the parametric and non-parametric procedures are compared each other only for $n \geq 50$.

Table 1 Best and worst $\mathbf{M}_{\mathrm{w}}$ across the four benchmarking procedures (Standard: Parametric CI; $\underline{\text{Underlined}}$: IMP; *Italics*: p CI; **Bold**: BCa CI) for BP_w and AC_2 for different n and k values

(a) Best $\mathbf{M}_{\mathrm{w}}$ for BP_w

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	*0.102*	**0.096**	*0.068*	$\underline{0.049}$
$k=5$	**0.123**	*0.081*	0.056	0.034
$k=7$	**0.087**	*0.066*	0.048	0.027

(b) Worst $\mathbf{M}_{\mathrm{w}}$ for BP_w

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	**0.160**	*0.118*	$\underline{0.080}$	*0.058*
$k=5$	*0.131*	**0.088**	$\underline{0.072}$	$\underline{0.051}$
$k=7$	*0.089*	0.072	$\underline{0.060}$	$\underline{0.044}$

(c) Best $\mathbf{M}_{\mathrm{w}}$ for AC_2

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	*0.159*	*0.098*	*0.072*	0.046
$k=5$	*0.111*	*0.073*	0.051	0.030
$k=7$	*0.085*	**0.031**	0.044	*0.026*

(d) Worst $\mathbf{M}_{\mathrm{w}}$ for AC_2

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	**0.193**	**0.099**	**0.084**	**0.055**
$k=5$	**0.130**	**0.092**	**0.066**	$\underline{0.046}$
$k=7$	**0.092**	*0.058*	**0.056**	$\underline{0.042}$

For small and moderate samples (i.e. $n \leq 30$), $\mathbf{M}_{\mathrm{w}}$ slightly differs across non-parametric benchmarking procedures and agreement coefficients: specifically, the highest difference in $\mathbf{M}_{\mathrm{w}}$ is 9%, observed for $n = 10$ and $k = 2$. Moreover, for increasing sample sizes, $\mathbf{M}_{\mathrm{w}}$ becomes quite indistinguishable across procedures and coefficients with a difference always no more than 2%. It is worthwhile to pinpoint that the differences in $\mathbf{M}_{\mathrm{w}}$ across non-parametric benchmarking procedures and agreement coefficients get smaller as n increases because of the decreasing skewness in the distributions of the coefficients: if the distribution is symmetric, the BCa and *p* CIs agree.

5 Conclusions

The results of the Monte Carlo simulation suggest that for small samples the non-parametric benchmarking procedures based on bootstrap resampling have satisfactory and comparable properties in terms of weighted misclassification rate. Moreover, with $n \geq 30$ the performances of the procedures based on bootstrap CIs differ from each other at most for 2%, therefore benchmarking the lower bound of the percentile bootstrap confidence interval could be suggested — because of its less computational burden — for characterizing the extent of rater agreement, both for BP_w and AC_2. For large samples, the performances are indistinguishable across all benchmarking procedures, thus benchmarking the lower bound of the parametric confidence interval would be preferred being the easiest method to implement.

References

1. Altman, D.G.: Practical statistics for medical research. CRC press (1990)
2. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med **19**(9), 1141–1164 (2000)
3. Cicchetti, D.V., Allison, T.: A new procedure for assessing reliability of scoring EEG sleep recordings. Am J EEG Technol **11**(3), 101–110 (1971)
4. Cicchetti, D.V., Feinstein, A.R.: High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. **43**(6), 551–558 (1990)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educ Psychol Meas **20**(1), 37–46 (1960)
6. De Mast, J., Van Wieringen, W.N.: Measurement system analysis for categorical measurements: agreement and kappa-type indices. J Qual Technol **39**(3), 191–202 (2007)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol Bull **76**(5), 378–382 (1971)
8. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol **61**(1), 29–48 (2008)
9. Gwet, K.L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
10. Hartmann, D.P.: Considerations in the choice of interobserver reliability estimates. J. Appl. Behav. Anal. **10**(1), 103–116 (1977)
11. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics pp. 159–174 (1977)
12. Shrout, P.E.: Measurement reliability and agreement in psychiatry. Stat. Methods Med. Res. **7**(3), 301–317 (1998)
13. Thompson, W.D., Walter, S.D.: A reappraisal of the kappa coefficient. J Clin Epidemiol **41**(10), 949–958 (1988)

Measuring the multiple facets of tolerance using survey data.

Misurare le molteplici attitudini alla tolleranza usando dati di survey.

Caterina Liberati and Riccarda Longaretti and Alessandra Michelangeli

Abstract In recent studies, there has been a growing interest toward tolerance and its implications in social and economic systems. So far, the openness of people to homosexuals or to foreign-born people has been considered the best indicator of tolerant attitudes. In this paper, we consider tolerance as a multi-faced phenomenon involving several different social domains. The aim is to provide some recommendations on how to develop a multidimensional index for tolerance in the case of survey's items measured by Likert-scale. Our discussion relies on a case study about a student survey carried out at Milan-Bicocca University

Abstract *In studi recenti, c'è stato un crescente interesse verso la tolleranza e le sue implicazioni nei sistemi sociali ed economici. Finora, l'apertura delle persone agli omosessuali o alle persone nate all'estero è stata considerata il miglior indicatore di atteggiamenti tolleranti. In questo lavoro, consideriamo la tolleranza come un fenomeno dalle molteplici sfaccettature che coinvolge diversi domini sociali. Presentiamo alcune raccomandazioni da tenere in considerazione quando si sviluppa un indice composito per valutare il livello di tolleranza in caso di items di un questionario misurati con una scala Likert. La nostra discussione si basa su un caso studio su un'indagine studentesca svolta all'Università Milano-Bicocca.*

Key words: Tolerance Measure; Multidimensional Index; Ordinal data; Survey

Caterina Liberati
DEMS Universitá di Milano-Bicocca, p.zza Ateneo Nuovo 1, e-mail: caterina.liberati@unimib.it

Riccarda Longaretti
DEMS Universitá di Milano-Bicocca, p.zza Ateneo Nuovo 1, e-mail: riccarda.longaretti@unimib.it

Alessandra Michelangeli
DEMS Universitá di Milano-Bicocca, p.zza Ateneo Nuovo 1,
e-mail: alessandra.michelangeli@unimib.it

1 Introduction

Most studies about tolerance focus on individuals' attitudes towards homosexual people, ethnic minorities and migrants. In these works, the analysis is usually based on surveys with questions such as "*Would you like to have homosexuals as your neighbors?*" or "*Would you like to have people of a different race as your neighbors?*"[1] They are simple yes/no questions, formally known as polar questions, that give a clear-cut understanding of people attitudes. The fraction of respondents giving a positive answer is the measure of tolerance (see, for example ref. [1]). References [3] and [4] argue that openness to homosexuals is the best available indicator of tolerant attitudes.

In this paper, we adopt a wider perspective and look at tolerance as a multi-faced phenomenon involving several different social domains, so that attitudes towards homosexuals and foreign-born people are only a partial aspect of this phenomenon. We provide some recommendations of statistical nature on how to develop a multidimensional index when ordinal data are used. This is the case of surveys with Likert scale type questions that, usually, uses statements such as "*Please rate the extent to which you agree/disagree with the following*" and the response scales use anchors such as 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree, etc. One of the advantages of Likert scale survey questions is that they allow to measure attitudes and opinions with a greater degree of nuance than a simple yes/no question.

Our attempt assesses different domains of the social environment and at the same time raises considerations about how to merge all the aspects together. We rely our analysis on a survey administered to a sample of students at University of Milan-Bicocca. The multidimensional approach aggregates different dimensions of tolerance into an index. The aggregation procedure requires an explicit choice on construction of variables and weighting of dimensions. We discuss these aspects considering a student survey carried out at the University of Milan-Bicocca.

The paper is organized as follows: Section 2 describes the data and in Section 3 we make preliminary considerations about a synthetic measure of tolerance.

2 Data

To study tolerance, we use a survey addressed to university students on religiosity and inter-religious dialogue. The survey has been developed by the University of Milan-Bicocca, and it is part of a framework agreement signed by 30 Italian Universities and 25 Research Centers. The objective of the study is to investigate the relationship between "Gender and Religions" among university students. Its purpose is to survey the opinion of university students about different topics, such as

[1] Both questions are from the World Value Survey [6].

inter-religious dialogue, women/religion relationship, multicultural society, homosexuality. Questions, measured by 7-points Likert scale [2], are listed in Table 1.

Table 1 Questionnaire themes

Social Domain	**Item topic**
Inter-religious dialogue	Inter-religious dialogue mitigates conflicts in the Italian society Same freedom of religious practice for all religions in Italy Same freedom of religious practice for all religions in Italy
Women/religion relationship	Women priests into Catholic Church Women can pray together with Muslim men Women rabbis into Jewish religion
Death/religion relationship	Voluntary interruption of pregnancy Eutanasia is socially acceptable
Multicultural society	The marriage between people of different religions The marriage between people of different ethnic communities Italian society is enriched by foreign people
Homosexuality	All States should legally recognize same-sex marriage. All States should legally recognize the child adoption by same-sex couples.

Because of the sensitivity of topics surveyed and in order to preserve the privacy of respondents, the questionnaire has been administered via Pencil And Paper Interview (PAPI) without saving any identification details. The sample interviewed is composed of 3,386 bachelor and master students from different fields: humanistic studies, legal, political, economic and statistical sciences .

The gender distribution is very unbalanced showing a severe prevalence of females (73.10%) respect to males (25.90%). However, we observe the same proportion for the students' population at the university of Milano-Bicocca. Also nationality and residence show an unimodal frequency distribution, with a maximum in correspondence of labels "Italian" (81.13%) and "Milan metropolitan area" (41.04%), respectively.

3 Measuring Tolerance: considerations for a quantitative synthesis

As mentioned in the Introduction, the main challenge of this work is to point out some issues arising when tolerance is analyzed from a multidimensional perspective.

First, the metric of variables included in the aggregation procedure should be homogeneous. Combining data measured by different scales or mixed data types

[2] The answers ranged between 1 = Strongly Disagree to 7=Strongly Agree.

means transforming one or more variables, to allow for statistical or thematic analysis on all information available [2].

Second, the functional form used to aggregate items should take into account the type of relationship between items. For example, a linear additive form assumes that items are independent. It should be verified that they are actually independent and, if so, this would be a necessary and sufficient condition for a proper composite indicator [5]. On the other hand, the additive aggregation function permits the assessment of the marginal contribution of each variable separately, therefore it would be optimal if a measure of tolerance could be defined into an algebraic sum of alternative domains. If items are correlated, then a geometric functional form should be preferred the linear one.

Third, the aggregation procedure should take into account the variability in the distribution of each item. Items with a uniform distribution across individuals should have a lower weight than items characterized by a dominant rating-value. Indeed a uniform distribution implies a strong heterogeneity in the subjects' responses.

In conclusion, an ideal composite index should be able to deal with all the issues listed above or, in alternative, should be the result of a compromise between different requirements.

References

1. Berggren, N. and Elinder: Is tolerance good or bad for growth? Public Choice.**150**, 283–308 (2012)
2. Caracelli, V. J. and Greene, J.C.: Data Analysis Strategies for Mixed-Method Evaluation Designs. Educational Evaluation and Policy Analysis. **15**, pp. 195- 207 (1993)
3. Inglehart, R. and Welzel, C. Modernization, cultural change and democracy. Cambridge University Press, Cambridge (2005).
4. Mellander, C., and Florida, R. . The creative class or human capital? Explaining regional development in Sweden. Working Paper No. 79. Stockholm: CESIS, Royal Institute of Technology (2007)
5. Nardo M., Saisana M., Saltelli A., Tarantola S., Hoffman A., Giovannini E.: Handbook on constructing composite indicators: methodology and user guide. OECD Statistics Working Paper, Paris (2005)
6. WORLD VALUES SURVEY 1981-2014 LONGITUDINAL AGGREGATE v.20150418. World Values Survey Association (www.worldvaluessurvey.org). Aggregate File Producer: JDSystems, Madrid SPAIN.

Modified profile likelihood in models for clustered data with missing values

Verosimiglianza profilo modificata in modelli per dati raggruppati con valori mancanti

Claudia Di Caterina and Nicola Sartori

Abstract Clustered data are frequently subject to missing values, especially those collected from longitudinal studies. The main focus of the analysts is usually not on the clustering variables, hence the group-specific parameters are treated as nuisance. If a fixed effects formulation is preferred and the total number of clusters is large relative to the single-group sizes, classical frequentist techniques are often misleading. We propose here to combine multiple imputation and the modified profile likelihood function to obtain accurate inferences on a parameter of interest under models with incidental parameters for incomplete grouped observations. Such solution is examined via simulation studies which shed light on the convenience for the imputation model to take into account the clustered structure of the data.

Abstract *Nei dati raggruppati si registrano abitualmente valori mancanti, soprattutto in quelli raccolti per studi longitudinali. L'attenzione degli analisti di solito non è rivolta alle variabili di raggruppamento, dunque i parametri specifici dei vari cluster sono considerati di disturbo. Se si preferisce adottare una formulazione ad effetti fissi ed il numero totale di gruppi è grande rispetto alle singole dimensioni di questi, le classiche tecniche frequentiste risultano spesso inadeguate. Qui proponiamo di combinare l'imputazione multipla e la verosimiglianza profilo modificata per ottenere un'inferenza accurata sul parametro d'interesse in modelli con parametri incidentali per osservazioni incomplete organizzate in cluster. Tale soluzione viene esaminata attraverso studi di simulazione che fanno luce sull'opportunità che il modello di imputazione tenga conto della struttura raggruppata dei dati.*

Key words: fixed effects, incidental parameters, missing at random, multiple imputation.

Claudia Di Caterina
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova, Italy; e-mail: dicaterina@stat.unipd.it

Nicola Sartori
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova, Italy; e-mail: sartori@stat.unipd.it

1 Introduction

Clustered, stratified or grouped data are either cross-sectional or longitudinal observations that can be arranged in groups. Missing values are ubiquitous in quantitative research analysis, particularly in clustered data resulting from clinical trials or panel surveys. Depending on the pattern and mechanism of missingness, a variety of techniques for handling inference in the presence of incomplete datasets can be used (see, e.g., Little and Rubin, 2002). When observations are organized in many groups of small to moderate size, statistical models which capture the unobserved heterogeneity across clusters via group-specific nuisance parameters are likely to suffer from the incidental parameters problem (Neyman and Scott, 1948). Such specifications are referred to as fixed effects models, in opposition to the random effects models which require to assume a distribution for the group features and their incorrelation with the covariates in the model.

Here we focus our attention on clustered observations characterized by both aspects, and propose a twofold strategy. On the one hand, tackling the incompleteness of the data by means of multiple imputation, and on the other, dealing with the incidental parameters assumed by the model through the modified profile likelihood function. More details on the two approaches can be found in Sections 2 and 3, respectively. Section 4 shows simulation results that help to investigate how the considered inferential tools should be combined in order to draw reliable conclusions on the parameter of interest.

2 Multiple imputation

The basic rationale behind multiple imputation (MI) is to exploit the distribution of the observed data in order to estimate a set of plausible values for the unobserved data. In particular, m multiply imputed datasets are created by substituting the missing observations in the original sample with draws from the posterior predictive distribution of the unobserved data conditional on the observed data. These completed datasets are then separately analyzed and the m results are pooled into overall estimates and standard errors using Rubin's rules (Rubin, 1987).

Various methods can be adopted to generating imputations (Little and Rubin, 2002, Section 10.2). Among those drawing from pragmatic conditional distributions when more variables are incomplete, multiple imputation by chained equations (MICE) (van Buuren and Oudshoorn, 1999) provides considerable flexibility in customizing imputation models for different data characteristics (Ji et al., 2018). For a thorough overview of the standard procedure and a helpful guidance for practice in case of data missing at random (MAR), we refer to White et al. (2011).

A well-known matter in MI inference is uncongeniality (Meng, 1994), which occurs when the imputer's model class and the ultimate analyst's model class are incompatible. Recently, Xie and Meng (2017) have pointed out many open problems connected with this topic. The general prescription is to include in the imputation model all variables that are related to the missing data, so that to make the MAR assumption more plausible. This should reduce the need to make special adjustments for mechanisms that are not MAR (van Buuren and Oudshoorn, 1999). With specific reference to models for clustered observations with incidental parameters, a typical question concerns whether and how accounting for the groups when imputing the missing values. White et al. (2011) suggest to disregard the clustering in this phase, if this is not of direct interest. Results in Andridge (2011) also highlight the inadequate inferential performance due to the inclusion of the fixed effects in the imputation model, contrary to what happens with random effects. On the opposite, Reiter et al. (2006) conclude that completely ignoring the sampling design during MI can be a risky practice. Further evidence is surely needed in this area.

3 Modified profile likelihood

In fixed effects models for clustered data where the number of groups is much larger than the single group sizes, the incidental parameters problem descends from the magnitude of the bias of the profile score function (McCullagh and Tibshirani, 1990). Correcting for the presence of the nuisance components, Barndorff-Nielsen (1980, 1983) proposed to rely on the modified profile likelihood (MPL) for making adequate inference on the parameter of interest. In fact, its superiority with respect to the ordinary profile likelihood (PL) within the two-index asymptotic setting can be proved for independent clustered sample units (Sartori, 2003).

For observations y_{it} subdivided in N groups of sizes T_i, suppose the model

$$Y_{it} \sim f(y_{it};x_{it},\psi,\lambda_i)\,, \qquad i=1,\dots,N, \quad t=1,\dots,T_i, \tag{1}$$

where x_{it} is a p-dimensional vector of covariates. The global parameter is $\theta = (\psi,\lambda)$, where $\psi \in \Psi \subseteq \mathrm{I\!R}^k$ denotes the component of interest and $\lambda = (\lambda_1,\dots,\lambda_N) \in \Lambda$ indicates the vector with incidental parameters. Note that, here and henceforth, to avoid clutter the transpose symbol acting on vectors is omitted. Moreover, we assume $T_i = T$ and $\dim(\lambda_i) = 1$ $(i=1,\dots,N)$ for the sake of notational simplicity. With independent groups, the log-likelihood function about θ can be expressed by

$$l(\theta) = \sum_{i=1}^{N} l^i(\theta) = \sum_{i=1}^{N} l^i(\psi,\lambda_i)\,,$$

with $l^i(\psi,\lambda_i) = \sum_{t=1}^{T} \log p(y_{it};x_{it},\psi,\lambda_i)$. Let us define the full maximum likelihood (ML) estimate for model (1) as $\hat\theta = (\hat\psi,\hat\lambda) = \arg\max_\theta l(\theta)$. Standard inference on the parameter of interest is typically based on the profile log-likelihood

$$l_P(\psi) = \sum_{i=1}^{N} l^i\big(\psi,\hat\lambda_{i\psi}\big) = \sum_{i=1}^{N} l_P^i(\psi)\,,$$

where $\hat{\lambda}_{i\psi}$ is the constrained ML estimate of λ_i for fixed ψ obtained, under usual regularity conditions, by equating to zero the score $l_{\lambda_i}(\theta) = \partial l^i(\psi,\lambda_i)/\partial\lambda_i$ and solving for λ_i $(i = 1,\dots,N)$. Given $\hat{\lambda}_\psi = (\hat{\lambda}_{1\psi},\dots,\hat{\lambda}_{N\psi})$, the full constrained ML estimate for fixed ψ is denoted by $\hat{\theta}_\psi = (\psi,\hat{\lambda}_\psi)$.

The general expression taken by the logarithmic version of the MPL is $l_M(\psi) = l_P(\psi) + M(\psi)$, and one computationally convenient formulation of $M(\psi)$ is owed to Severini (1998). Specifically, using the additive form $M(\psi) = \sum_{i=1}^N M_i(\psi)$, the ith summand in Severini's modification term equals

$$M_i(\psi) = \frac{1}{2}\log j_{\lambda_i\lambda_i}(\hat{\theta}_\psi) - \log I_{\lambda_i\lambda_i}(\hat{\theta};\hat{\theta}_\psi), \qquad i = 1,\dots,N.$$

In the above equation, we have $j_{\lambda_i\lambda_i}(\theta) = -\partial^2 l^i(\psi,\lambda_i)/(\partial\lambda_i\partial\lambda_i)$ and $I_{\lambda_i\lambda_i}(\hat{\theta};\hat{\theta}_\psi) = E_{\theta_0}\{l_{\lambda_i}(\theta_0)l_{\lambda_i}(\theta_1)\}\big|_{\theta_0=\hat{\theta},\theta_1=\hat{\theta}_\psi}$ indicating the scalar expected value calculated with regard to $\hat{\theta}$ of the product of partial scores evaluated at two different points in the parameter space.

4 Simulation studies

Monte Carlo experiments based on 1000 iterations can be run to study the effectiveness of the approach which incorporates MI and MPL inferences. The cluster size and number of groups considered are $T = 6$ and $N = 50,100,250$, respectively. For each couple (T,N), $p = 2$ covariates are randomly generated. The first, x_{1it}, is sampled from a Bernoulli(0.5) distribution. The second, x_{2it}, is drawn from the $N(0,1)$ random variable. We then simulate the binary clustered outcomes as independent realizations of $Y_{it} \sim Bern(\pi_{it})$ $(i = 1,\dots,N, t = 1,\dots,T)$. In particular $\pi_{it} = e^{\lambda_i+\beta x_{it}}/(1+e^{\lambda_i+\beta x_{it}})$, where $x_{it} = (x_{1it},x_{2it})$, $\beta = (\beta_1,\beta_2) = (1,1)$ and each λ_i is independently generated from the standard normal distribution. Here our interest is confined to datasets with completely observed response and MAR predictors, yet the same methodology could be applied in different contexts of incompleteness. Missing entries are thus created by deleting x_{1it} with probability ω_{1it} and x_{2it} with probability ω_{2it}. According to the dependence of the missingness on the stratification, two main scenarios may be distinguished. In case of missingness unrelated to the groups (setting I), we suppose $\omega_{1it} = e^{-2+y_{it}}/(1+e^{-2+y_{it}})$ and $\omega_{2it} = e^{0.5-y_{it}}/(1+e^{0.5-y_{it}})$. When instead the clustered structure plays a role in the probability of observing a covariate (setting II), we use $\omega_{1it} = e^{\lambda_i-2+0.2y_{it}}/(1+e^{\lambda_i-2+y_{it}})$ and $\omega_{2it} = e^{\lambda_i-y_{it}}/(1+e^{\lambda_i-y_{it}})$. Such values are chosen in order for the fraction of missing data in the datasets to be around 35%. The procedure starts by obtaining $m = 5$ complete samples through MICE, using a logistic regression for imputing x_{1it} and a Bayesian linear regression for x_{2it}, as implemented by the R package `mice` (van Buuren and Groothuis-Oudshoorn, 2010). In both imputation models, the dummy variables indicating the groups are either included or not. Inference on each completed dataset is then conducted using PL and MPL for the parame-

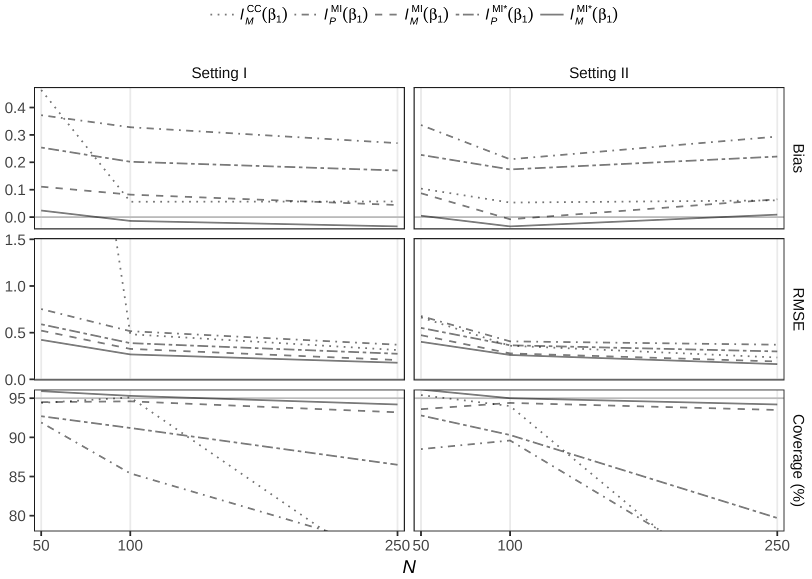


Fig. 1 Comparison between inferences on β_1 obtained via complete-case analysis $\left(l_M^{\text{CC}}\right)$ and multiple imputation, either taking the groups into consideration when imputing the missing values $\left(l_P^{\text{MI}}, l_M^{\text{MI}}\right)$ or not $\left(l_P^{\text{MI}^*}, l_M^{\text{MI}^*}\right)$. Results based on 1000 clustered datasets simulated with $T = 6$ observations per group and $N = 50, 100, 250$ number of groups.

ter of interest β in the logistic regression with outcomes y_{it}, using the R package `panelMPL` (Bellio and Sartori, 2015). Rubin's rules are finally applied to pool together estimates and variances derived by the two likelihood functions. A complete-case analysis, which disregards units with missing values, is also carried out via both methods. Due to space constraints, just partial results of the experiments referred to β_1 are shown in Figure 1. Therein, empirical bias and root mean squared error (RMSE) of the various estimators can be compared, along with coverage of 95% Wald confidence intervals. Note that the performance of the complete-case PL is not reported, as it was found to be poorer than any other method in all scenarios. The output indicates that the solution combining MI and MPL outperforms the complete-case analysis, in terms of point and interval estimation. In addition, it seems that to neglect the clustering while imputing the unobserved covariates is recommendable, whether the incompleteness depends on the specific group features or not. One plausible reason is the incidental parameters problem observed under the imputation model. An improved fit of the latter might be achieved, for instance, by adopting bias reduction (Firth, 1993). This can be the subject of future research, as well as developments of the present work that consider other values for T and different patterns and mechanisms of missingness in the data.

References

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical journal 53*, 57–74.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika 67*, 293–310.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika 70*, 343–365.

Bellio, R. and N. Sartori (2015). panelMPL: *Modified profile likelihood estimation for fixed-effects panel data models*. http://ruggerobellio.weebly.com/software.html.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*, 27–38.

Ji, L., S.-M. Chow, A. C. Schermerhorn, N. C. Jacobson, and E. M. Cummings (2018). Handling missing data in the modeling of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–22.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley, New York.

McCullagh, P. and R. Tibshirani (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological) 52*, 325–344.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science 9*, 538–558.

Neyman, J. and E. Scott (1948, January). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Reiter, J. P., T. E. Raghunathan, and S. K. Kinney (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology 32*, 143.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika 90*, 533–549.

Severini, T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika 85*, 403–411.

van Buuren, S. and K. Groothuis-Oudshoorn (2010). MICE: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1–68.

van Buuren, S. and K. Oudshoorn (1999). Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054. For associated software see http://www.multiple-imputation.com.

White, I. R., P. Royston, and A. M. Wood (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine 30*, 377–399.

Xie, X. and X.-L. Meng (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when gods, imputers and analysts models are uncongenial. *Statistica Sinica 27*, 1485–1594.

Worthiness Based Social Scaling

Giulio D'Epifanio

Abstract

The construction of a set of scales is delineated, for evaluating the performance of social agents (e.g. providers of services as hospitals, schools, etc.) conditionally on "reference states" $x := X \in \{x_1, \ldots, x_R\}$ of the governed individuals. Each scale is associated to an index which uses conditional "worthiness increases" $\omega_{l|x}$, between the levels of an ordinal outcome indicator $Y := l \in (0, 1, .., L)$. This indicator was been defined on a scheduled, by the policy-maker (PM), chain of hierarchically ordered goals. The "worthiness increases" are interpreted by modeling interrelated latent evolutionary processes, on the scheduled goal chain, up to hyperparameters γ which are driven by conditions x. Then, to standardize the set of scales on a given "reference behavior", a pseudo-Bayesian (see [1]) method is used which elicits value γ^* by minimizing "residual from updating" (see [4]). It norms the model specifications on the "reference data" of the (chosen a priori) "standard agent". Finally, adhering to general requirements in rational choices from the decision theory, a standardized worthiness-based index can be implemented, which takes into input the agents actual data.

Key words: performance index, ordinal scaling, worthiness

1 Indexing worthiness

The performance of any social agent u is associated to the "social behavior", described by the set of distributions (e.g. see table 1) $p_{|x}[u] := (p_{0|x}, p_{1|x}, \cdots, p_{L|x})[u]$, which were realized on the set of the individuals that u governs, upon the levels of

G. D'Epifanio
Depart. of Political Sciences, via A. Pascoli 06123 Perugia Italy, e-mail: ggiuliodd@gmail.com.it

an ordinal classifier of outcome Y varying the status $x := X \in \{x_1, \ldots, x_R\}$ of the governed individuals.

agent A1	performance level (Y)					**agent A2**	performance level(Y)					**agent A3**	performance level(Y)					**agent A4**	performance level (Y)				
status (X)	I	II	III	IV	V	status (X)	I	II	III	IV	V	status (X)	I	II	III	IV	V	status (X)	I	II	III	IV	V
x1	0	0	0	0	0	x1	2	2	0	0	0	x1	0	3	1	0	0	x1	1	3	0	0	0
x2	2	9	5	0	0	x2	3	13	9	1	0	x2	2	24	16	0	0	x2	16	37	18	0	0
x3	0	3	18	0	0	x3	1	20	31	4	1	x3	0	20	48	1	0	x3	0	36	59	4	1
x4	0	3	17	3	0	x4	1	3	59	8	1	x4	0	2	53	3	0	x4	0	12	107	10	0
x5	0	0	14	9	4	x5	0	6	48	18	3	x5	0	0	49	30	2	x5	0	0	87	43	4

Table 1 Example. Actual data of the social agents to be evaluated

reference agent A0	performance level (Y)				
status (X)	I	II	III	V	V
x1	3	8	1	0	0
x2	23	83	48	1	0
x3	1	79	156	9	2
x4	1	20	236	24	1
x5	0	6	198	100	13

Table 2 Example. Reference data of the standard-agent A_0

Suppose that the PM has specified a certain chain of, increasingly challenging, binary-outcome goals

$$O_0 \preceq O_1 \preceq O_2 \preceq \ldots \preceq O_l \preceq \ldots \preceq \ldots \preceq O_{L-1} \preceq O_L := O_{Full}, \tag{1}$$

which are hierarchically (i.e. Guttman like) ordered. Then, the (nominally recoded on $\{0, 1, \ldots, L\}$) ordinal outcome indicator Y is defined so that the event occurrence "$Y \geq l$" identifies the achieving of the l-th scheduled goal $O_l := (Y \geq l)$, $l := 0, \ldots, L$. Therefore, the pursued "full purpose" could be realized at different degree of achieving, from the "tautological" (i.e. alway achieved) goal $O_0 := (Y \geq 0)$ toward the final goal O_L. Let $\mathscr{P}^*$denote the population of the (real or perhaps virtual) individuals which are governed by the reference agent *(*e.g. a recognized "best practice" for standardization*)* A_0 (e.g. see table 2). Then, the criterion of intrinsic worthiness (see [3]) may be interpreted[1] on a goal-based probabilistic setup as follows.

For any actual individual i, having achieved goal O_{l-1} on chain of goals (1), the higher "the risk of failing the next goal O_l", referring such a risk on the population $\mathscr{P}^*$, the greater the "increase of worthiness", due to the performance of the agent which governs i "as if" i was in $\mathscr{P}^*$, whenever it actually achieves goal O_l.

[1] Consider hierarchical chain of goals *(1)*. Given that a certain goal O_{l-1} has been achieved, the greater the resistance, with reference to the evaluation framework, to also achieve the next pursued goal O_l*,* by continuing to improve, the greater the increment of value due to the intrinsic worthiness of who, effectively, is able to achieve it.

Thus, the $\mathscr{P}^*-$standardized, conditionally on status $x := X \in \{x_1, \dots, x_R\}$, worthiness increase between any two adjacent levels of $Y := l \in (0, 1, .., L)$ is provided[2] (for $l := 1, .., L$) by:

$$\omega^*_{l|x} := \Delta_{l-1} Val_{|x} := Val_{|x}(O_l) - Val_{|x}(O_{l-1}) =$$
$$= \varphi_l\left(\frac{Pr\{Y = l-1|x;\, \mathscr{P}^*\}}{Pr\{Y \geq l-1|x;\, \mathscr{P}^*\}}\right) = \varphi_l\left(\frac{p_{l-1|x}}{p_{l-1|x} + p_{l|x} + \cdots + p_{L|x}}\right) \geq 0 \qquad (2)$$

Here, continuous monotone functions $\varphi_l(.)$ (e.g. set here the identity) of the conditional probability rates may be chosen (see [3]) for specifying some characteristics (e.g. the additivity) of the scale. Formally re-interpreting "worthiness increases" as "utility increases", functionals of the "rank dependent expected utility", adhering to requirements of rational choices (e.g. see [2], pp. 559), leads to the following instance of conditional-expectation-based index[3]:

$$u :\longmapsto W[p_x[u];\, \omega^*_x, x] := \sum_{l:=1}^{L} \varphi_l\left(\frac{Pr\{Y = l-1\,|\,x;\, \mathscr{P}^*\}}{Pr\{Y \geq l-1\,|\,x;\, \mathscr{P}^*\}}\right) \cdot (1 - F_{Y|x}[p[u]](l)) \qquad (3)$$

Here, $F_{Y|x}[p]$ denotes the cumulative distribution such that $F_{Y|x}[p](l) = p_{0|x} + p_{1|x} + \cdots + p_{(l-1)|x}$. Thus, through $x \in \{x_1, \dots, x_R\}$, it may be defined the global evaluation index: $u \longmapsto \sum_{r:=1}^{R} q_r \cdot W[p_{x_r}([u];\, \omega_{x_r}(\mathscr{P}^*)]$. It uses the actual agents data (e.g. see table 1)), standardized on the reference-agent's data (e.g. see table 2). Here, $q_r \geq 0$ ($\sum_{l:=1}^{R} q_r = 1$) weights[4] the reference domain for the status x_r.

2 Eliciting standardized worthiness increases

To justify differences in "worthiness increases" (2), through reference conditions $x := X \in \{x_1, \dots, x_R\}$, the PM may adopt some "reference evaluation criterion" and working assumptions formally specified by means of a structural probabilistic model (4)-(7). Here[5], the conditional rates (1-ν_{rl}) (which enter "worthiness increases"

[2] It is the worthiness credit which is gained by any social agent in improving the condition of a "*standard individual*", in the reference condition $x \in \{x_1, \dots, x_R\}$, from the current level $(l-1)$ to the next l on the scale of Y which was constructed on goal chain (1).

[3] for any agent u, given x, it takes into input the distribution realized (e.g. see table 1), by the individuals that u governs in condition x, on the standardized worthiness-quantified levels of Y.

[4] these weights should represent the political relevancy of the "social reference domains" to the main aim of the PM.

[5] On the stratum of the n_r individuals in the condition x_r, the manifest outcome $(Y_{r0}, \dots, Y_{rL})$ is distributed as a multinomial (eq.4) where the expectation-parameters $\psi_r := (\psi_{r0}, \psi_{r1}, \dots, \psi_{rL})$ are normed, within the container Dirichlet model (eq. 5), on a set of constraints on the latent evolutionary processes undertaken the levels of outcome scale Y (eqs (6)-(7)).

(2)) are represented as latent parameters of interrelated evolutionary-processes behind the goals chain (1), which are driven by manifest conditions x up to hyper-parameters[6] $\gamma := (\mu_0, \delta, \beta^X)$ to be regulated. Then, the methodological question arises in automatic eliciting of values γ^* so that "worthiness increases" $\omega_{l|x}(\mathscr{P}^*; \gamma^*)$ enter evaluation indexes (3). To norm the model on the reference-agent's data table (2), recalling a "minimum information principle"[7], hyper-parameters γ may be regulated (e.g. see [5],[4]) to that value γ^* such that the *"residual from updating"*[8]$\| \; Vec\,(\, E(\Psi \mid y,x;\gamma,w) - E(\Psi \mid x;\gamma,\, w)\,) \; \|$ is minimized subject to specifications of constraints (6)-(7).

$$Y_r{:=}\{Y_{r0},...,Y_{rL}\}|\psi_r \overset{ind.}{\sim}{}_{r:=1,...,R} Mult(y_{r0},...,y_{rL};\psi_{r0},\psi_{r1},...,\psi_{rL},n_r) \tag{4}$$

$$\psi_r{:=}(\psi_{r0},\psi_{r1},...,\psi_{rL})|m_r,a_r \overset{ind.}{\sim}{}_{r:=1,...,R} Dirichlet(\psi_{r0},...,\psi_{rL};m_r,a_r) \tag{5}$$

$$m_r{:=}(m_{r0},\, m_{r1},...,m_{rL}),\, 0{<}m_{rl}{:=}E[\psi_{rl}]{<}1,\, \textstyle\sum_{s:=0}^{L} m_{rs}{=}1, \quad a_r{:=}w_r, w_r{>}0$$

$$v_{r1}{:=}\frac{m_{r0}}{m_{r0}+m_{r1}}=\frac{e^{\eta_{r1}}}{1+e^{\eta_{r1}}}\,, \tag{6}$$

$$v_{rl}{:=}\frac{m_{r0}+...+m_{r(l-1)}}{m_{r0}+m_{r1}+...+m_{rl}}=\frac{e^{\eta_{irl}}}{1+e^{\eta_{rl}}}\,,$$

$$\cdots$$

$$v_{rL}{:=}\frac{m_{r0}+...+m_{r(L-1)}}{m_{r0}+m_{r1}+...+m_{rL}}=\frac{e^{\eta_{rL}}}{1+e^{\eta_{rL}}}\,,$$

$$\eta_{rl}{=}\mu_0{+}\textstyle\sum_{s:=1}^{L} \delta_l{\cdot}I_{(s=l)}{+}\sum_{w:=2}^{R} \beta^X_{w(l-1)}{\cdot}I_{(X(r)=w)} \tag{7}$$

$$reference\,condition\,r{:=}1,...,R{:=}5;\; scale\,level\,transitions\,l{:=}1,...,L{:=}4$$

References

[1] Casella G, Robert C P (2002) Monte Carlo Statistical Methods (third printing). Springer, New York

[2] Chateauneuf A., Cohen M., Meilijson I. (2004), Four Notions of Mean-preserving Increase in Risk, Risk attitudes and Applications to the Rank-dependent Expected Utility Model, Journal of Mathematical Economics **40**, 547-571

[3] D'Epifanio G (2011) Sviluppo di un Indice Multi-attributo per la Valutazione del Merito. In "Criteri e indicatori per misurare l'efficacia delle attivit universitarie", vol I, pp. 279, CLEUP, Padova (for previous versions see: http://www.ec.unipg.it/DEFS/depifanio.html?lang=it)

[4] D'Epifanio G (2005) Data Dependent Prior Modeling and Estimation in Contingency Tables. In: Vichi M. et al (eds) Studies in Classification Data Analysis and Knowledge Organization, Springer-Verlag, NewYork http://www.springerlink.com/content/j73233521955n624/?p=6505e914841943a6bbeaf8cbd144238b&pi=3

[6] Hyper-parameters μ_0, δ represent, respectively, the common base-line and increments in the level scores of the scale; instead, the parameters of profileβ^X represent the crossed effects of condition and levels, in transition processes. Here, $I_{(.)}$ denote a binary indicator function.

[7] "The less a prior representation of knowledge is updated by current data, the more intrinsically it already was accounted for by the *intrinsic information* added by such data"

[8] Here, $E(\Psi \mid y,x;\gamma,w)$ denotes the predictive expectation of full parameter profile $\Psi := (\psi_1, .., \psi_R)$ which is updated by outcome y, whereas $E(\Psi \mid x;\gamma,w)$ is his non updated counterpart, over the design-point x

[5] D'Epifanio G (1996) Notes on A Recursive Procedure for Point Estimation. Test, **5**, **1**, 1-24, http://www.springerlink.com/content/m0458041m043/?p=6505e914841943a6bbeaf8cbd144238b&pi=0

[6] D'Epifanio G (2017) Indexing the Normalized Worthiness of Social Agents In: Perna C. et al (eds) Studies in Theoretical and Applied Statistics, SIS 2016, Springer-Verlag (forthcoming)

Direct Individual Differences Scaling for Evaluation of Research Quality

DINDSCAL per la Valutazione della qualità della Ricerca

Gallo M., Trendafilov N., and Simonacci V.

Abstract *The eValuation of Research Quality (VQR) is one of the most important assessment processes achieved by the National Agency for the Evaluation of Universities and Research Institutes. Its main task is to provide information on the status of the Italian research system by assessing the performance of universities in various scientific areas. The basic evaluation criteria were defined by panels of experts according to the specific characteristics of each subject area and through a synthetic statement on the products submitted by researchers. With the aim of studying this phenomenon in depth, DINDSCAL (Direct Individual Differences Scaling) model is proposed for a compositional analysis of VQR dataset.*

Abstract *La Valutazione della Qualità della Ricerca (VQR) è uno dei processi di valutazione più importanti realizzati dall'Agenzia Nazionale di Valutazione del Sistema Universitario. Il suo compito principale è fornire informazioni sullo stato del sistema di ricerca italiano valutando le prestazioni delle universitın varie aree scientifiche. I criteri di valutazione di base sono stati definiti da gruppi di esperti in base alle caratteristiche specifiche di ciascuna area tematica e attraverso una dichiarazione sintetica sui prodotti presentati dai ricercatori. Con l'obiettivo di studiare approfonditamente tale fenomeno, il modello DINDSCAL è proposto per l'analisi composizionale dei dati VQR.*

Key words: compositional data, log-ratios, DINDSCAL, Stiefel mainifold, VQR data.

Michele Gallo
Department of Human and Social Sciences, University of Naples "L'Orientale", Italy.
e-mail: mgallo@unior.it

Nickolay Trendafilov
School of Mathematics and Statistics, Open University, UK.
e-mail: nickolay.trendafilov@open.ac.uk

Violetta Simonacci
Department of Human and Social Sciences, University of Naples "L'Orientale", Italy.
e-mail: mgallo@unior.it

1 Introduction

The National Agency for the Evaluation of the University and Research Systems (ANVUR), in the framework of an evaluation project (VQR), collected research outputs from 96 Italian universities, including 18 research Institutes. Sixteen panel of experts in evaluation (GEV), one for each scientific area of research, classified the products in specific merit classes. According to Ministerial Decree no. 458 dated 27 June 2015, common guidelines were defined for all GEVs. A new approach based on DINDSCAL (Direct Individual Differences Scaling [2]) model and compositional analysis is proposed in this work to extract information regarding the criteria used by the GEV, if the dimension of structure and the geographic location influence the quality of research outputs. In literature the INDSCAL (Individual Differences Scaling) model is used to study the individual differences in three-way data by doubly centered set of matrices of squared dissimilarity measures between a range of stimuli [1]. A direct approach as DINDSCAL is here preferred, in order to directly analyse simultaneous slices of squared dissimilarity matrices organized as compositional data.
The new approach is shortly described in Sect. 2. Sect. 3 summarizes the analysis of the VQR data.

2 Theory

2.1 Compositional data

Let $\underline{V}$ $(n \times p \times m)$ be a three-way array where each row v_{ik} $(i = 1,\cdots,n, k = 1,\cdots,m)$ is a compositional vector of p parts observed after the kth treatment (or occasion). From a geometrical point of view the sample space for all vectors v_{ik} is the simplex. There is a rich literature for CoDa proprieties and how is possible to handle them in simplex space (for a detailed review and references see [3]).
Here the CoDa are transformed in centred log-ratio (clr) in order to move from simplex to real space [4]. Let $\underline{L}$ $(n \times p \times m)$ be a three-way array with the CoDa in logarithm scale $[l_{ijk} = log(v_{ijk})]$. The clr-coordinates for each frontal slice of $\underline{X}$ are defined $X_k = L_kJ_p$, where L_k is a $n \times p$ matrix and J_p is a $p \times p$ centring matrix, $J_p = I_p - \frac{1}{p}E_p$ with I is an identity matrix and E is a matrix of ones. Thus, when the columns of X_k are centred, $X_k = J_nL_kJ_p$, the metric multidimensional scaling (MDS) for each X_k is given by the following identity:

$$-\frac{1}{2}J_n(D_k \odot D_k)J_n = X_kX_k^\top, \tag{1}$$

where '$\odot$' denotes the usual elementwize (Hadamard) matrix product and D_k is a $n \times n$ symmetric matrix containing zero on main diagonal and the dissimilarity measure between the n compositions collected at the kth occasion.

As well as recalled, the measures in D_k are Aitchison distances. Thus, they have all properties necessary for a meaningful interpretation of compositional results. Moreover, the identity (1) shows that $-\frac{1}{2}J_n(D_k \odot D_k)J_n$ is positive semi-definite.
With the aim to a simultaneous metric of m symmetric slices D_k, the INDSCAL model decomposes the each slice as

$$-\frac{1}{2}J_n(D_k \odot D_k)J_n = Q\Lambda_k Q^\top + \Delta_k, \tag{2}$$

where Q is $n \times r$ assumed of full column rank, Λ_k diagonal matrix and Δ_k is $n \times n$ matrix, containing the errors of the model fit. In other words all slices share a common loading matrix Q and differ each other only by the (non-negative) diagonal elements of Λ_k called idiosyncratic saliences.
Unfortunately, in (2) the parameter set of all $n \times r$ matrices Q with full column rank is a non-compact Stiefel manifold. To solve this drawback an approach called direct INDSCAL was proposed by [2].

2.2 DINDSCAL problem

Following the approach proposed by [5], it is easy to show that the squared Aitchison distances are given by the following identity:

$$D_k \odot D_k = (I_n \odot X_k X_k^\top)E_n + E_n(I_n \odot X_k X_k^\top) - 2X_k X_k^\top \quad k = 1, ..., m. \tag{3}$$

Thus, the DINDSCAL fitting problem for CoDa is concerned with the following equality constrained optimization problem:

$$\min_{Q,\Lambda_k} \sum_{k=1}^{m} \|D_k \odot D_k - (I_n \odot Q\Lambda_k^2 Q^\top)E_n - E_n(I_n \odot Q\Lambda_k^2 Q^\top) + 2Q\Lambda_k^2 Q^\top\| \ , \tag{4}$$

subject to $(Q, \Lambda_1, \Lambda_2, \ldots, \Lambda_m) \in \mathscr{O}_0(n,r) \times \mathscr{D}(r)^m$, where $\mathscr{D}(r)^m = \underbrace{\mathscr{D}(r) \times \ldots \times \mathscr{D}(r)}_{m}$, and $\mathscr{D}(r)$ denotes the set of all $r \times r$ diagonal matrices. $\mathscr{O}_0(n,r)$ denotes the set of all $n \times r$ orthonormal matrices with zero column-sums, i.e.: $\mathrm{O}_0(n,r) := \{Q \in \Re^{n \times r} | \ Q^\top Q = I_r$ and $E_n Q = 0_n\}$. It is easy to observe that the additional constraint of clr-transformation zero column- rows sums, i.e. $E_n X_k = 0_n$ and $X_k E_p = 0_p$ does not introduce new constrains. Thus the problem of minima defined in (4) can be solved by a non-linear conjugate gradient algorithm, which leads to globally convergent algorithms.
For a formal introduction to the method and its numerical integration see [6], while all information about the gradient dynamical system for the DINDSCAL problem is given in [2].

3 Case study

The research outputs submitted from each university was calculated considering the number of university staff members. Each product was classified in a specific merit classes, that is, Excellent, Good, Fair, Acceptable, and Limited. Products classified as not eligible are assigned to a specific merit class. According to the kind of research outputs (articles, monographs, book chapters, etc.) bibliometric algorithm or peer-review methodology were used to evaluate the 117,079 research outputs submitted in the 16 MIUR scientific areas.
The data are preprocessed according to the procedure described in Sect. 2.1, DINDSCAL is able to find some important differences between two groups of scientific areas and between the scientific structures with different size. In short, there is a tangible contraposition between large-size scientific structures, especially those localized in the Centre of Italy, and the others. Moreover, the medium size structures localized in the Centre and in the South of Italy are characterized only by specific scientific area.

References

[1] Carroll, J.D., Chang J.J.: Analysis of individual differences in multidimensional scaling via an $n-$way generalization of "Eckart-Young" decomposition. *Psychometrika,* 35, 283–319 (1970).

[2] Trendafilov, N.T.: Dindscal: direct INDSCAL. *Statistics and Computing*, 22(2), 445-454 (2012).

[3] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: *Modeling and analysis of compositional data.* John Wiley & Sons (2015).

[4] Gallo, M.: Tucker3 model for compositional data. *Communications in Statistics-Theory and Methods*, pp 4441-4453 (2015).

[5] Browne, M.W.: The Young-Householder algorithm and the least squares multidimensional scaling of squared distances, *Journal of Classification*, 4, 175–219 (1987).

[6] Trendafilov, N.T.: The dynamical system approach to multivariate data analysis, a review, *Journal of Computational and Graphical Statistics*, 15, 628–650 (2006).

A test for variable importance

Sulla scelta di variabili esplicative rilevanti

Rosaria Simone

Abstract Statistical literature is being more and more concerned with debates about hypothesis testing and p-values supporting the significance of a given variable specification. Specifically, if on one hand statistical foundations about significance are not arguable, scholars should be able to distinguish between significance and variable importance. This is a matter of serious concern in questionnaire analysis to derive respondents' profiles and develop targeted marketing strategies, for instance. To this aim, this contribution proposes a hypothesis system that considers the normalized dissimilarity measure to assess the importance of explanatory variables in the setting of mixture models for ordinal data to account for uncertainty of choice.
Abstract *Un dibattito sempre più presente nella letteratura statistica moderna riguarda lo studio della effettiva importanza di covariate significative rispetto ai risultati dei classici test di ipotesi e tecniche di selezione del modello. Se l'applicazione delle procedure standard non è in discussione, d'altra parte la distinzione tra variabili significative e variabili rilevanti assume un ruolo fondamentale ai fini decisionali. Tale problematica è di particolare rilievo nell'dei dati provenienti da questionari, ad esempio per la profilazione dei consumatori al fine di individuare specifiche strategie di marketing. In questo contesto, un sistema di ipotesi basato su una misura di dissimilarità viene proposto per testare la rilevanza di variabili esplicative nel caso di una classe di misture per dati ordinali.*

Key words: Dissimilarity, Variable Selection and Importance, CUB mixture models, Rating Data

Rosaria Simone
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22, 80138 Naples, Italy e-mail: rosaria.simone@unina.it

1 Motivations

The opening lines of [3] invite readers to critically use p-values in the era of big data:

> There is growing frustration with the concept of the p-value. Besides having an ambiguous interpretation, the p-value can be made as small as desired by increasing the sample size, n. The p-value is outdated and does not make sense with big data: Everything becomes statistically significant.

The purpose of this contribution is to investigate such concern and propose a methodology for variable selection in the setting of statistical models for rating data. Our discussion stems from a well consolidated idea to measure separation of probability distributions relying on the concept of Gini's *Transvariation* [5] which has been applied in several circumstances (see [1], for instance). Departing from the identity: $\min(a,b) = \frac{1}{2}\big(a+b-|a-b|\big)$, an inverse indicator of how far apart two (discrete) probability distributions $\mathbf{p} = (p_1,\dots,p_m)'$, $\mathbf{q} = (q_1,\dots,q_m)'$ are, is given by:

$$\sum_{r=1}^{m} \min(p_r,q_r) = 1 - \frac{1}{2}\sum_{r=1}^{m} |p_r - q_r|. \tag{1}$$

This is a measure of their overlapping. For instance, consider the conditional distributions of a discrete response given a dichotomous factor (dashed lines joining mass probabilities are chosen to enhance visualization in Figure 1). Albeit statistically significant differences are found, the discrimination of response patterns becomes more meaningful from left to right as the overlapping gets smaller.

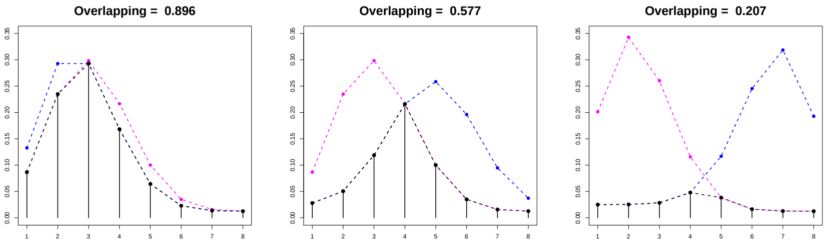


Fig. 1 Visualization of overlapping between discrete probability distributions

As an application of how this measure could work in discriminating significant covariates for a given ordinal response, here the focus will be on CUB models [2]. The original paradigm is the weighted combination of a (shifted) Binomial distribution $g_r(\xi_i)$ for the *feeling* component and a (discrete) Uniform for the *uncertainty* component, meant as the fuzziness derived from the discretization of the continuous latent perception. For a sample $(R_1,\dots,R_n)$ of ordinal responses, say on the support $\{1,\dots,m\}$ for a given $m>3$, a CUB regression model is specified via:

$$Pr(R_i = r|\boldsymbol{y}_i, \boldsymbol{w}_i) = \pi_i\, g_r(\xi_i) + (1-\pi_i)\frac{1}{m}, \qquad (2)$$

where feeling ξ_i and uncertainty π_i parameters are linked to values of subjects' covariates $\boldsymbol{w}_i, \boldsymbol{y}_i$ by a logit link:

$$logit(\xi_i) = \boldsymbol{w}_i\,\boldsymbol{\gamma}, \qquad logit(\pi_i) = \boldsymbol{y}_i\,\boldsymbol{\beta}.$$

This full model specification is customarily abbreviated as CUB (p,q), where $\boldsymbol{\gamma}^{'} = (\gamma_0,\dots,\gamma_q)^{'}, \boldsymbol{\beta}^{'} = (\beta_0,\dots,\beta_p)^{'}$ are the estimable parameters: when $p = q = 0$, then model fitting assumes constant feeling ξ and uncertainty π parameters. Estimation of CUB models relies on likelihood methods and, specifically, on the implementation of the Expectation-Maximization algorithm. Fit improvements yielded by the specification of covariates can be tested via a Likelihood Ratio Test if models are nested: in general, the significance of an explanatory variable can be checked via standard Wald test. In the following, let D be a dichotomous variable included in the model specification to explain patterns of responses in terms of feeling and/or uncertainty, so that $\boldsymbol{\theta}_0 = (\pi_0, \xi_0)$, $\boldsymbol{\theta}_1 = (\pi_1, \xi_1)$ are the parameters of the conditional CUB distributions $\big(R_i|D_i = 0\big) \sim$ CUB (π_0, ξ_0) and $\big(R_i|D_i = 1\big) \sim$ CUB (π_1, ξ_1). If D implies statistically significant differences in model parameters, then two subgroups of respondents are identified and one should establish if the resulting clustering is actually relevant. This issue is particularly common when the sample size n is large. Here we wish to discuss a system of hypothesis:

H_0 : D should not be retained in the model (D is not important)

versus

H_1 : D should be retained in the model (D is important).

Thus, significant differences in model parameters will be investigated in order to disclose to which extent the clustering variable is actually relevant and should be retained in the model.

2 A test for variable importance

In order to test if the inclusion of a significant factor leads to relevant improvement of the fit, the (normalized) dissimilarity measure [6, 8]:

$$Diss(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{2}\sum_{r=1}^{m} |p_r - q_r| \ \in (0,1) \qquad (3)$$

stems quite naturally from the motivating discussion. If $\boldsymbol{p},\boldsymbol{q}$ are two probability distributions, it assesses the proportion of cases in which the two distributions differ. Thus, if a CUB $(1,1)$ is fitted to the data with a dichotomous factor D for both components:

$$logit(\pi_i) = \beta_0 + \beta_1 D_i, \qquad logit(\xi_i) = \gamma_0 + \gamma_1 D_i,$$

and, accordingly, two clusters are identified, the dissimilarity between the estimated conditional response probabilities of $R_i|D=0$, $R_i|D=1$ indicates how far apart the groups $D=0$ and $D=1$ are in terms of the corresponding estimated CUB $(0,0)$ probability distributions. Similar considerations hold if the dichotomous variable D is specified only for one of the components.

3 A simulation experiment

The validation of the proposed approach to test variable importance will be run with a Monte Carlo experiment. For illustrative purposes, we shall consider the simplest case of a dichotomous variable D, with levels 0,1 (for instance, males and females, smokers and non smokers, etc.), able to discriminate feeling, by assuming heterogeneity constant among subjects. Thus, in the end we shall have two separate groups of respondents, corresponding to feeling parameters ξ_0 and ξ_1 if $D=0$ or $D=1$, respectively. We derive the empirical critical values c_α under the null $H_0 : \xi_0 = 0.30,\ \xi_1 = 0.35$ by generating a sample of data in which a dummy covariate is significant but the difference in parameter values is very small, thus it may raise doubts about importance of the implied classification. To this aim, we sample 1000 times from the null distribution for varying $\pi \in (0.2, 0.4, 0.6, 0.8)$, different numbers of categories and sample sizes for the two groups.

Empirical critical values for the dissimilarity statistics corresponding to nominal level $\alpha = 0.05$ are summarized in Table 1: lower and upper bounds (l_b and u_b, resp.) of 80% bootstrap confidence intervals (1000 replicates) are also reported as an instance of a measure of uncertainty of the test statistics. Thus, at level α, a value of dissimilarity between the implied conditional distributions lower than the corresponding critical value indicates that D has a weak importance for the purpose of discrimination of response patterns and its specification in the model could be matter of discussion.

As a by product of the simulation experiment, new evidence is found to support the specification of uncertainty for ordinal data models: indeed, critical values decrease for higher values of heterogeneity (that is, larger weights for the Uniform distribution), indicating that this component has a not-negligible effect in the analysis of variable importance. In order to enhance the purpose of the test, an additional simulation experiment has been planned: for each run and for the chosen parameter values, a sample has been generated with a significant dummy variable splitting the observations into two groups of sizes n_0 and n_1 respectively. Then the dissimilarity test has been applied to check for variable importance according to the proposal. Results are summarized in Table 2 and highlight that, especially for large samples,

Table 1 Empirical critical values with increasing level of uncertainty parameter

	$m = 5$			$m = 7$			$m = 10$		
	c_α	l_b	u_b	c_α	l_b	u_b	c_α	l_b	u_b
	$n_0 = 300, \quad n_1 = 400, \quad \xi_0 = 0.30, \quad \xi_1 = 0.35$								
$\pi = 0.2$	0.104	0.102	0.106	0.103	0.101	0.106	0.102	0.099	0.104
$\pi = 0.4$	0.117	0.115	0.119	0.112	0.109	0.114	0.116	0.112	0.119
$\pi = 0.6$	0.122	0.119	0.130	0.125	0.121	0.129	0.135	0.132	0.138
$\pi = 0.8$	0.137	0.132	0.142	0.135	0.131	0.139	0.147	0.142	0.150
	$n_0 = 1300, \quad n_1 = 1400, \quad \xi_0 = 0.30, \quad \xi_1 = 0.35$								
$\pi = 0.2$	0.058	0.057	0.061	0.053	0.052	0.055	0.054	0.053	0.055
$\pi = 0.4$	0.064	0.063	0.065	0.065	0.063	0.067	0.072	0.070	0.073
$\pi = 0.6$	0.075	0.073	0.076	0.079	0.078	0.080	0.094	0.093	0.096
$\pi = 0.8$	0.090	0.088	0.093	0.099	0.098	0.101	0.119	0.118	0.121
	$n_0 = 13000, \quad n_1 = 14000, \quad \xi_0 = 0.30, \quad \xi_1 = 0.35$								
$\pi = 0.2$	0.024	0.023	0.024	0.026	0.025	0.026	0.030	0.030	0.031
$\pi = 0.4$	0.042	0.041	0.042	0.046	0.046	0.047	0.056	0.055	0.056
$\pi = 0.6$	0.060	0.059	0.060	0.066	0.066	0.067	0.081	0.081	0.081
$\pi = 0.8$	0.077	0.076	0.077	0.086	0.086	0.086	0.106	0.105	0.106

the classical concept of statistical significance has to be accompanied by a more specific analysis of variable importance.

Table 2 Importance rates for a significant dummy variable

	$\pi = 0.2$		$\pi = 0.4$		$\pi = 0.6$		$\pi = 0.8$	
Importance	Yes	No	Yes	No	Yes	No	Yes	No
$n_0 = 300, \quad n_1 = 400, \quad \xi_0 = 0.3, \quad \xi_1 = 0.35$								
$m = 5$	0.09	0.91	0.06	0.94	0.04	0.96	0.04	0.96
$m = 7$	0.08	0.92	0.07	0.93	0.05	0.95	0.05	0.95
$m = 10$	0.06	0.94	0.09	0.91	0.04	0.96	0.11	0.89
$n_0 = 1300, \quad n_1 = 1400, \quad \xi_0 = 0.3, \quad \xi_1 = 0.35$								
$m = 5$	0.03	0.97	0.05	0.95	0.06	0.94	0.11	0.89
$m = 7$	0.07	0.93	0.05	0.95	0.14	0.86	0.08	0.92
$m = 10$	0.08	0.92	0.13	0.87	0.12	0.88	0.12	0.88
$n_0 = 13000, \quad n_1 = 14000, \quad \xi_0 = 0.3, \quad \xi_1 = 0.35$								
$m = 5$	0.16	0.84	0.12	0.88	0.09	0.91	0.17	0.83
$m = 7$	0.14	0.86	0.12	0.88	0.10	0.90	0.13	0.87
$m = 10$	0.14	0.86	0.11	0.89	0.11	0.89	0.14	0.86

The proposed variable-importance test has shown perfect agreement with the corresponding one using the (symmetrized) Kullback-Leibler divergence (here not reported for the sake of brevity), but it is more advantageous since the dissimilarity measure is a proper (normalized) distance and it is able to foster interpretation and visualization of results. Similar conclusions are found when considering $\xi_0 = 0.1$ or $\xi_1 = 0.5$ for the feeling under the null (it is not necessary to test for values $\xi > 0.5$

since CUB distributions are reversible) and for increasing differences between ξ_0 and ξ_1. Dually, the proposed test can be run in case the clustering covariate is tested to explain heterogeneity for samples with homogeneous feeling or both components.

4 On-going developments

The proposed testing procedure for variable importance prescribes that, once a dichotomous factor is specified in the model to explain the response (in terms of feeling and uncertainty in case one assumes the CUB paradigm), then the dissimilarity between the estimated conditional distributions can reveal its discrimination ability in an effective and insightful way. The topics here investigated are being subject to more in-depth analysis stemming from real case-studies; further studies are tailored to the application of the approach to other classes of models, as well as to the study of properties of the dissimilarity estimator -also known as Duncan segregation index in other contexts- in the vein of [4, 10]. Notice that the same approach can be exploited to design a proper test of significance for difference in parameter values:

$$H_0 : \boldsymbol{\theta}_0 = \boldsymbol{\theta}_1 \quad versus \quad H_1 : \boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$$

with test statistics based on the dissimilarity between the conditional distributions. From some preliminary investigations in this perspective one obtains a test that is as powerful as the corresponding one using the Kullback-Leibler divergence to assess distances between distributions. This approach has been investigated in [7, 9] to design a homogeneity test in case of continuous populations and small sample sizes.

References

1. Bragoli, D., Ganugi, P., Ianulardo, G.: Gini's transvariation analysis: an application on financial crises in developing countries. Empirica **40**, 153–174 (2013)
2. D'Elia A., Piccolo D.: A mixture model for preference data analysis. Comput. Stat. Data An. **49**, 917–934 (2005)
3. Demidenko, E.: The p-values You Can't Buy. The American Statistician. **70**, 33–38 (2016)
4. Forcina, A., Galmacci, G.: On the distribution of the Index of Dissimilarity. Metron **32**, 361–374 (1974)
5. Gini, C.: Il concetto di transvariazione e le sue prime applicazioni. in: Transvariazione, Gini, C. ed., Libreria Goliardica, Roma (1916)
6. Gini, C.: La dissomiglianza. Metron, 85–215 (1965)
7. Girone, G., Nannavecchia, A.: The distribution of an Index of Dissimilarity for two samples from a Uniform Population. Applied Mathematics **4**, 1028–1037 (2013)
8. Leti, G. (1983). *Statistica descrittiva*. Il Mulino, Bologna.
9. Manca, F., Marin, C.: Simulated Sample Behaviour of a Dissimilarity Index when Sampling from Populations differing by a location parameter only. Applied Mathematics **5**, 2199–2208 (2014)
10. Mazza, A., Punzo, A.: On the Upward Bias of the Dissimilarity Index and Its Corrections. Sociological Methods and Research **44**(1), 80– 107 (2015)

Statistical Models New Proposals

Decomposing Large Networks: An Approach Based on the MCA based Community Detection

La decomposizione di networks di grosse dimensioni: un approccio basato sull'identificazione di comunitá via MCA

Carlo Drago

Abstract The emergence of the big data has called for considering new methodologies to analyze big networks. In these particular contexts there are many cases in which it is important to take into account not only the single node but groups of nodes which can have the same or similar functions on a defined network. On large networks it is important to represent them in a meaningful way. Interval data seems an adequate representation which can be used to represent these networks. The specific contribution of this work it is to show the way in which is possible to rank the different structural characteristics of the different robust communities represented by the network. The rank applied to the structural characteristics allows the understanding also of the relevant core of the network

Abstract *L'emergere dei big data ha richiesto di considerare nuove metodologie per analizzare le grandi reti. In questi contesti ci sono molti casi in cui è importante prendere in considerazione non solo il singolo nodo ma gruppi di nodi che possono avere le stesse funzioni o funzioni simili su una rete definita. I dati ad intervallo sembrano una rappresentazione adeguata che può essere utilizzata per rappresentare queste reti. Il contributo specifico di questo lavoro è mostrare il modo in cui sia possibile classificare le diverse caratteristiche strutturali delle diverse comunità robuste rappresentate dalla rete. Il rango applicato alle caratteristiche strutturali consente di comprendere il nucleo principale della rete*

Key words: Community Detection, Interval Data, MCA, Network Analysis, Symbolic Data

Carlo Drago
University of Rome "Niccolo Cusano", Via Don Carlo Gnocchi 3, e-mail: carlo.drago@unicusano.it

1 Big Data and Networks

The emergence of the big data has demanded the consideration of new methodologies to analyze big networks. In particular the growth in size of the social networks has called for a new relevant role for the different platforms which have given various new services. At the same time the data related to the different attributes on the network is growing exponentially. In this way it is usually difficult to handle and analyze networks and it is necessary to define an approach which can be useful to deal with these types of networks. Furthermore in these contexts there are many cases in which it is important to take into account not only the single node but also the group of nodes which can have the same functions on a defined network. One strategy is to decompose the networks and to represent it in a manageable way [14]. The different groups of nodes need to be considered as kinds of compartments on the networks and they can have a similar function or role on the networks as a whole [6]. It could be important to consider the groups of the nodes as a relevant entity and it is relevant to analyze the different relationships between the different entities. The challenge is to represent the network by considering their more relevant parts. The approach proposed considers symbolic data [1]. So the proposal is to consider a specific representation for the community or the specific aggregated data and then consider the community also as an entire entity.

2 The Analysis of the Community Structure

The different communities are groups of nodes which tend to be strongly connected to each other and they tend to be loosely connected with nodes of other communities [6]. The identification of the community structure is very important in order to detect groups of nodes which can be part of the same functional structure of the same network. The communities are the relevant elements on the construction of a network. In this sense we consider each different network as based on the communities which can be identified on the network. The first step is to identify the different communities which can be considered inside the network and then represent them. There are various different methodologies with the aim of detecting the different communities inside a network [9]. Each different method can have a different performance [11, 10]. In particular different algorithms can have different biases for the separate network structures and so we have to compare the results we obtain using different community algorithms. In this regard the global optimum as set of nodes of the considered objective can be really discrepant than the one returned by each method [10]. The use of a single methodology can as the Louvain method can be robustified by considering other methods and synthesize the eventually different results obtained. It is useful to consider approaches which can take into account an ensamble of different algorithms or approaches in order to synthesize the results obtained [4]. So we obtain a robust community structure via multiple correspondence analysis (MCA) and we validate them using the Rand Index. At this point it is necessary to represent the different communities in such a way which cannot lead to the loss of relevant information from the original data. The computed Rand Index gives us information on the capacity of the resulting representation to ”capture” the initial results of the different community detection on the network.

3 From the Communities to their Representation

Each different community, can be represented as a different interval data [3]. Differently it is possible to consider the entire network as a symbolic data [7]. In this sense we are able to obtain different interval data for each community. The procedure used is comprised of three steps: identifying the different communities from a network using an approach of community detection [9] (eventually using an MCA approach), and then from the different member community we can obtain the interval data. Following [3] each different community is based on all the single nodes of the network. From the interval data considered it is possible to measure the different attributes which are relevant in order to represent the entire community. Each measure is related to structural characteristics or attributes of the same node. We can have the attribute or structural characteristics for the entire community in addition to the attribute of those of the single n nodes. Then we have the single interval measure for the community based on those of the nodes which are members of the community:

$$X^a = (x_1, x_2, \dots, x_n) \tag{1}$$

Where x is a measure for the nodes belonging to a community X^a (for instance the different betweenness or the degree). The interval data for the single community is:

$$X^{I,a} = [\overline{x}, \underline{x}] \tag{2}$$

Where $\overline{x}$ represents the upper bound of the measure belonging to the community and the $\underline{x}$ the lower bound. At this point we can consider the descriptors of the different communities as intervals [8]. In this way we can consider both the single different observations, but also the different communities by considering the intervals of their measures. It is possible to compare the different communities by their attributes (the upper and the lower bound) but also the centers and the radii [12]. We have the center:

$$X^{I,a}_{center} = \frac{1}{2}(\underline{x} + \overline{x}) \tag{3}$$

we can also consider the range between the upper and the lower bound

$$X^{I,a}_{range} = (\underline{x} - \overline{x}) \tag{4}$$

and the radii

$$X^{I,a}_{radius} = \frac{1}{2}(\underline{x} + \overline{x}) \tag{5}$$

These descriptors allow to take into account the different communities and to compare them.

4 Ranking the Different Representations

At this point it is necessary to identify the different rankings of the representations. In this sense we have to explicitly consider the different intervals and their attributes. In particular each interval can be characterized by their attributes as the lower bound and the upper bound. Starting from their descriptors it is possible to compare the different attributes or structural indicators for each community considered. Following [12] we consider the ranking for the different intervals obtained. The comparison can be conducted by considering the different attributes of the intervals (the upper and the lower bounds, the range and the radii). An application of ranking of the different attributes of the different communities is to detect the centre of the network based on the different communities. In this sense we are interested not in single nodes but in considering the communities as the initial point of the analysis. The ranking of centrality, for instance, is computed by considering the different communities, and at the same time those selected are considered on the final network selected by their structural characteristics. At this point, it is possible to consider the ranking also by taking into account only a number of different communities. The aim is to detect the central part of the network for some relevant structural characteristics. We obtain a stylized structure of the network considering the most relevant communities. The validation phase is performed by observing a graph in which are visualized the changes on some indicators (betweenness and degree for instance). We consider the changes on the center values for each community. A radar plot [13] is a tool to analyze and compare the different measures on the ranking: it could be used as a diagnostic tool in the choice. The final network structure is based on considering only these communities.

5 Simulation Study and Application on Real Data

It is possible to consider different simulated networks in order to evaluate the procedure proposed. In order to test the algorithms then we consider various types of networks and we consider the approach for each different network. In particular we simulate different networks of different typology and different size and then we apply the approach (Barabasi Game, Erdos Renyi and also Forest Fire [2]). We are able to show the community structure by detecting the different communities using the MCA-based community algorithm procedure [4]. Then we represent them as interval data and we compute two descriptors as upper and lower bound for each community. Finally we are able to compute also the center and the radius. The statistical methods considered on the different intervals based on the communities are on [8]. The package RSDA on R allows the performing of different computations based on interval data [15]. We visualize the ranking of the different communities obtained by considering the appropriate methods and we can visualize them by using a radar plot. A radar plot visualizes each attribute of the community and structural indicator expressed as interval. At the same time we can choose the number of communities by observing the change on the relevant center parameters in the different communities (on betweenness and degree in our case). So we are able to visualize the most central communities by considering the highest ranked communities by their betweenness and the degree. At the same time the radar plot is actually showing the ranking considering also the other structural characteristics represented as interval data for each specific data. Finally by choosing the first ranked communities we are able to identify the stylized structure of the network starting from their specific initial structure. In this sense we start from the entire structure and then we are able to rank the different communities by considering the different attributes. Finally we select the first communities and we obtain the most central communities from the network. In the case of application on real data we consider the network of the Zachary karate club [16]. Here we are able to observe and select (see the figure 6) the most relevant part of the network by selection of the most central communities. These communities identify the ”core of the network” rather than other peripherical network structures.

6 Conclusions

The procedure considered determines the different communities of the network and detects the most central different representations by considering some structural indicators as the betweenness or the Freeman degree. Other attributes of the different communities can be considered. The approach followed in this paper is to consider the different communities, representing them as interval data and then ranking them. It is important to emphasize that the analysis is community-based and it is robust allowing to enclosing the results of many community detection algorithms.

References

1. Billard, L., & Diday, E. (2006) *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. England: Wiley & Sons Ltd
2. Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org
3. Drago C. (2017) Identifying Meta Communities on Large Networks. SIS Italian Statistical Society 2017 Conference: Statistics and Data Science: New Challenges, New Generations.

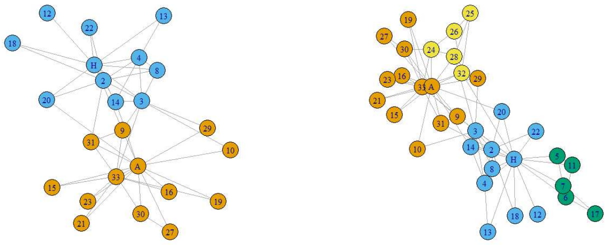


Fig. 1 Zachary Karate Club: selected part of the network (on the left) and the entire network (on the right)

4. Drago C. (2017) MCA Based Community Detection In book: Classification, (Big) Data Analysis and Statistical Learning, Edition: Studies in Classification, Data Analysis, and Knowledge Organization, Publisher: Springer, Editors: Francesco Mola, Claudio Conversano, Maurizio Vichi
5. Duan, L., & Binbasioglu, M. (2017). An ensemble framework for community detection. Journal of Industrial Information Integration, **5**, 1-5.
6. Fortunato, S. (2010). Community detection in graphs. Physics reports, **486** (3), 75-174.
7. Giordano G., Brito M. P. (2014) Social Networks as Symbolic Data, in: Analysis and Modeling of Complex Data in Behavioral and Social Sciences, Edited by Vicari, D, Okada, A, Ragozini, G, Weihs, C. (Eds, 06/2014; Springer Series: Studies in Classification, Data Analysis, and Knowledge Organization.
8. Gioia, F., & Lauro, C. N. (2005). Basic statistical methods for interval data. Statistica applicata, **17** (1), 1-27.
9. Khan, B. S., & Niazi, M. A. (2017). Network Community Detection: A Review and Visual Survey. arXiv preprint arXiv:1708.00977.
10. Leskovec, J., Lang, K. J., & Mahoney, M. (2010, April). Empirical comparison of algorithms for network community detection. In Proceedings of the 19th international conference on World wide web (pp. 631-640). ACM.
11. Mahmoud, H., Masulli, F., Rovetta, S., & Abdullatif, A. (2016, September). Comparison of Methods for Community Detection in Networks. In International Conference on Artificial Neural Networks (pp. 216-224). Springer International Publishing.
12. Mballo, C., & Diday, E. (2005). Decision trees on interval valued variables. The electronic journal of symbolic data analysis, 3(1), 8-18.
13. Noirhomme-Fraiture, M. (2002). Visualization of large data sets: the zoom star solution. International Electronic Journal of Symbolic Data Analysis, 26-39.
14. Richards, W., & Macindoe, O. (2010, August). Decomposing social networks. In Social Computing (SocialCom), 2010 IEEE Second International Conference on (pp. 114-119). IEEE.
15. Rodriguez R.O. (2017) with contributions from Carlos Aguero, Olger Calderon, Roberto Zuniga and Jorge Arce. RSDA: R to Symbolic Data Analysis. R package version 2.0.2. https://CRAN.R-project.org/package=RSDA
16. Zachary W.W. An information flow model for conflict and fission in small groups, Journal of Anthropological Research **33**, 452-473 (1977).

On Bayesian high-dimensional regression with binary predictors: a simulation study

La regressione Bayesiana con previsori binari in contesti ad alta dimensionalità: uno studio di simulazione

Debora Slanzi, Valentina Mameli and Irene Poli

Abstract Aim of this work is to develop a comparative analysis to evaluate the performances of several Bayesian regression approaches in the high-dimensional context where the number of observations is very small with respect to the number of predictors. Moreover in this study we assume that the predictors can be expressed only as binary variables coding the presence or the absence of a particular characteristic of the system. This binary structure is very present in many real studies, in particular in laboratory experimentation.
Abstract *Lo scopo di questo lavoro è quello di sviluppare un'analisi comparativa per valutare il comportamento di alcuni metodi inferenziali di regressione Bayesiana in contesti di alta dimensonalità dove il numero di osservazioni è molto piccolo rispetto al numero dei predittori assunti per il modello. Lo studio considera solo predittori espressi in forma di variabili binarie in grado di codificare la presenza e l'assenza di una particolare caratteristica del sistema. Questa struttura del problema presente in molti studi di fenomeni reali e in particolare in ambito sperimentale.*

Key words: Bayesian regression, Binary predictors, Comparative analysis, High-dimensionality.

Debora Slanzi
Department of Management, Ca' Foscari University of Venice, San Giobbe, Cannaregio 873, Venice (IT) and European Centre for Living Technology, S. Marco 2940, Venice (IT), e-mail: debora.slanzi@unive.it

Valentina Mameli
European Centre for Living Technology, S. Marco 2940, Venice (IT) e-mail: valentina.mameli@unive.it

Irene Poli
Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, via Torino 155, Mestre (IT) and European Centre for Living Technology, S. Marco 2940, Venice (IT), e-mail: irenpoli@unive.it

1 Introduction

Bayesian regression models have been widely studied and adopted in the statistical literature [14, 10]. Many studies regard the development of efficient and effective priors to select the set of relevant variables and derive accurate posterior predictive distributions [6, 4]. Moreover in the context of high-dimensionality, when there are many predictors, sparsity is assumed and many parameters can be set to values very close to zero without affecting the fit of the model [11, 9]. The Bayesian penalized regression techniques for the analysis of high-dimensional data include, among others, the Bayesian Lasso [8, 7], the normal-gamma regression [5], the horse-shoe regression [1] and the Bayesian ridge regression [3, 12]. Generally the setup of the regression considers the standard multiple linear model assuming independent Normal error terms. Moreover it is usual to standardize both the response and the covariates to have zero mean and variance equals to one. While there are several studies conducted to compare the performances of the models when the predictors are continuous, these approaches are not very suited when the predictors are binary variables. This situation frequently occurs in many experimental fields, as for example in biochemical studies where the presence and absence of a component determines the results of the experimentation and affects the success of the study. In this paper we focus on this particular situation, and we conduct a simulation study to compare the performance of several high-dimensional Bayesian regression models when the predictors are expressed as binary variables.

The paper is organized as follows. In Section 2 we present the Bayesian multivariate regression model and we introduce the prior distributions considered in the analysis. Section 3 describes the characteristics of the simulation study and presents the results of the comparison by means of indicators of goodness of fitting and prediction. Finally in Section 4 we derive some concluding remarks.

2 Bayesian regression

Let consider the standard multiple linear regression model which assumes that a vector of responses $y = (y_1, y_2, \dots, y_n)$ can be represented as

$$y = \alpha\mathbf{1} + \mathbf{X}\beta + \varepsilon$$

where the vector of errors $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ are independent with $p(\varepsilon_i) = N(\varepsilon_i|0, \sigma^2)$ and $\mathbf{X}$ is an $n \times p$ matrix of predictor variables. The scalar α is the intercept, $\mathbf{1}$ a $n \times 1$ unit vector and the vector β represents the regression coefficients. In this work we adopt the Bayesian inferential approach which involves a choice of the prior distribution of the $(p \times 1)$-dimensional vector of regression coefficients β. Many approaches are proposed in literature to derive effective and efficient prior distributions with different characteristics. Among them, the sparsity inducting priors are commonly applied in the setting of high-dimensionality, where most of the predictors are assumed to be unassociated with the responses. The hierarchical representation of the full Bayesian regression model as proposed by [8], introduces the distributions of the parameters and the hyperparameters as follows:

$$\begin{aligned} y|\alpha,\mathbf{X},\beta,\sigma^2 &\sim N_n(\alpha\mathbf{1}+\mathbf{X}\beta,\sigma^2\mathbf{I}_n), \\ \beta_j|\lambda_j^2,\sigma^2 &\sim N(0,\lambda_j^2\sigma^2), \\ \sigma^2 &\sim \pi(\sigma^2)d\sigma^2, \\ \lambda_j^2 &\sim \pi(\lambda_j^2)d\lambda_j^2. \end{aligned}$$

Different priors for σ^2 lead to different regression structures in terms of error distribution, while the hyperparameters $\lambda_1,\lambda_2,\dots,\lambda_p$ are used to model the sparsity characteristics and control the amount of shrinkage in the coefficient estimates [11]. Usually σ^2 follows an improper prior distribution proportional to $1/\sigma^2$, while the distribution assumed for $\lambda_1,\lambda_2,\dots,\lambda_p$ leads to different prior distributions for the regression coefficients β. Therefore, depending on the particular choices for the local shrinkage hyperparameters $\lambda_1,\lambda_2,\dots,\lambda_p$ we can consider some of the most frequently used Bayesian regression models:

- *Bayesian Lasso regression*: the hyperparameters $\lambda_1,\lambda_2,\dots,\lambda_p$ follow a joint exponential prior distribution which depends on further hyperparameters. Generally, this assumption is simplified by assuming that $\lambda_j^2 \sim Exp(1)$, $i = 1,\dots,p$ [8, 7];
- *Horseshoe regression*: the prior distribution for the local shrinkage hyperparameters $\lambda_1,\lambda_2,\dots,\lambda_p$ is the zero-mean half-Cauchy distribution [1];
- *Normal-Gamma regression*: the hyperparameters $\lambda_1,\lambda_2,\dots,\lambda_p$ follow a Gamma distribution where both shape and scale parameters have an associated prior distribution. Then the marginal distribution of β_j is generally affected by these choices in a way that smaller value of shape parameter of the Gamma distribution is associated with larger amount of shrinkage for the betas [5];
- *Bayesian Ridge regression*: it can be obtained by assuming $\lambda_1^2 = \lambda_2^2 = \dots = \lambda_p^2 = \lambda^2$ [3].

3 A comparative simulation study

We conduct a comparative simulation study to evaluate the performance of the Bayesian regression models under different prior distributions: Bayesian Lasso, the Bayesian horse-shoe regression, the Bayesian normal-gamma regression and the Bayesian ridge. The study has been conducted considering only binary variable predictors. The statistical analyses were performed using the R-project free software environment for statistical computing. In particular, we use the R-package `monomvn` to fit the regression models [13].

3.1 Experimental setting

The simulation is based on the linear regression model $y_i = X_i\beta + \varepsilon_i$, $i = 1,\dots,n$, where $\varepsilon_i \sim N(0,1)$. Predictors are generated independently from a Bernoulli distribution with probability of success 0.1 to represent sparsity condition[1]. The simulation considers an increasing number of predictors, p =200, 500, 1000, 1500, 2000, 3000. The number of non-zero regression coefficients is assumed to be $p^\star = 10$ with values $\{-1,1\}$.

From this data generative process, we simulate $N = 2000$ observations (the full dataset) and we randomly select n =100, 200 and 500 sample points (corresponding to the 5%, 10% and 25% of the full space) as training set to estimate the models. The remaining data are considered as test set on which to evaluate the performance of the various Bayesian methods. Each simulation is repeated 50 MonteCarlo runs to evaluate the robustness of the approaches and we compute the Predictive Mean Square Error (PMSE) and the Sensitivity (the ratio between the number of selected important variables and the number of actual important variables) as defined in [2]. In particular, values of Sensitivity close to 1 means that the approach is able to select the relevant information for the regression.

3.2 Comparative results

The results for increasing number of predictors are presented, i.e. p= 200, 500, 1000, 1500, 2000, 3000. Note that for n=200, we did not run simulations for p=200, and for n=500 we did not run simulations for p=200 and 500. In Table 1 we report the evaluation of the prediction capacity by means of PMSE for the different regression models. We can see that with regard to the predictive power of the models, they perform almost in the same way producing accurate predictions in particular when the number of covariates don't exceed too much the number of observations. This predictive power tends to decrease when the ratio between the number of observations and variables tends to decrease. We notice that there is the same trend for all the different approaches, however, the Bayesian Ridge regression presents the poorest performance for all the different values of n here considered. A different situation emerges if we consider the sensitivity measure by which the power of selecting important variables of the simulation is expressed. In Figure 1 we show how this measure evolves through the increasing values of p assuming different values of n. We notice that again the approaches show a very similar performance, but the Bayesian Ridge regression presents values of Sensitivity higher than the other regressions. Therefore, comparing the different indicators of performance of these regressions we see that they all have a good prediction accuracy but the Bayesian Ridge regression presents an higher capacity (Sensitivity) to detect the relevant vari-

[1] We plan to develop the simulation also for other usual values of probability of success representing sparsity condition.

Table 1 Predictive Mean Square Errors for regression: BL=Bayesian Lasso; HS= Horseshoe regression; NG= Normal-Gamma regression; BR= Bayesian Ridge (standard errors of MonteCarlo runs in parentheses).

n	p	BL	HS	NG	BR
100	200	1.52 (0.10)	1.55 (0.14)	1.52 (0.10)	1.81 (0.15)
100	500	1.80 (0.11)	1.86 (0.16)	1.81 (0.11)	1.85 (0.16)
100	1000	1.88 (0.09)	1.91(0.14)	1.88 (0.09)	1.87 (0.13)
100	1500	1.99 (0.07)	2.00 (0.10)	1.99 (0.09)	2.08 (0.11)
100	2000	1.87 (0.06)	1.87 (0.07)	1.89 (0.10)	1.99 (0.11)
100	3000	1.99 (0.05)	1.98 (0.05)	1.99 (0.07)	2.20 (0.14)
200	500	1.40 (0.13)	1.42 (0.17)	1.41 (0.14)	1.53 (0.15)
200	1000	1.54 (0.16)	1.57 (0.15)	1.54 (0.14)	1.73 (0.30)
200	1500	1.75 (0.20)	1.77 (0.19)	1.75 (0.17)	1.94 (0.24)
200	2000	1.78 (0.10)	1.77 (0.11)	1.78 (0.10)	2.06 (0.22)
200	3000	1.93 (0.10)	1.91 (0.11)	1.92 (0.12)	2.28 (0.23)
500	1000	1.15 (0.07)	1.12 (0.09)	1.13 (0.08)	1.20 (0.09)
500	1500	1.35 (0.15)	1.34 (0.17)	1.35 (0.16)	1.36 (0.18)
500	2000	1.54 (1.13)	1.52 (0.14)	1.55 (0.14)	1.60 (0.16)
500	3000	1.80 (0.11)	1.76 (0.15)	1.79 (0.16)	1.85 (0.19)

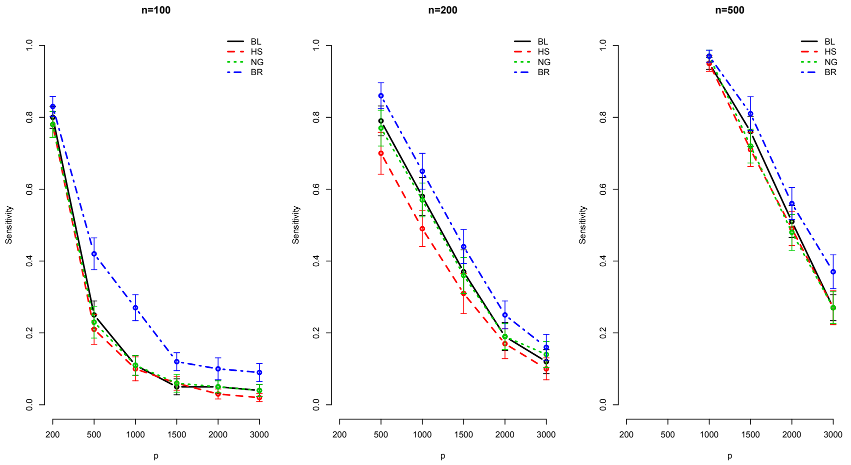


Fig. 1 Sensitivity measures for regression: BL=Bayesian Lasso; HS= Horseshoe regression; NG= Normal-Gamma regression; BR= Bayesian Ridge.

ables of the system. The results of this simulation can be helpful when choosing the structure of the regression model to adopt in a particular study.

4 Concluding remarks

In this work we have developed a comparative analysis to study the performance of several Bayesian regressions with binary predictors in terms of predictive accuracy and variables selection. Further analyses will be conducted to strengthen these preliminary results and to identify the relation between n and p in deriving reliable inferential results in particular when binary predictors are present in the model.

References

1. Carvalho, C., Polson, N., Scott, J. The horseshoe estimator for sparse signals. Biometrika **97**, 465–480 (2010)
2. Geng,Z.,Wang,S.,Yu,M.,Monahan,P.O.,Champion,V.,Wahba,G.: Group variable selection via convex log-exp-sum penalty with application to a breast cancer survivor study. Biometrics **71(1)**, 53–62 (2015)
3. Geweke, J.: Variable selection and model comparison in regression. In: Bernardo,J.M. et al. (eds.) Bayesian Statistics, pp. 609-620. Oxford Press (1996)
4. Griffin, J.E., Brown, P.J.: Hierarchical Shrinkage Priors for Regression Models. Bayesian Analysis **12(1)**,135–159 (2017)
5. Griffin, J.E., Brown, P.J.: Inference with Normal-Gamma prior distributions in regression problems. Bayesian Analysis **5**, 171–188 (2010)
6. Griffin, J.E., Brown, P.J.: Some priors for sparse regression modelling. Bayesian Analysis **8**, 691–702 (2013)
7. Hans, C.: Bayesian lasso regression. Biometrika **96**, 835–845 (2009)
8. Park, T., Casella, G. The Bayesian Lasso. Journal of the American Statistical Association **103**, 672–680 (2008)
9. Piironen, J., Vehtari, A.: Sparsity information and regularization in the horseshoe and other shrinkage priors. Electrononic Journal of Statististics **11(2)**, 5018–5051 (2017)
10. Piironen, J., Vehtari, A.: Comparison of Bayesian predictive methods for model selection. Statistical Computing **27**, 711–735 (2017)
11. Polson, N.G., Scott, J.G.:Local shrinkage rules, L?evy processes and Regularized Regression. Journal of the Royal Statistical Society, Series B **74**, 287–311 (2012)
12. Polson, N.G., Scott, J.G., Windle, J.: The Bayesian bridge. Journal of the Royal Statistical Society: Series B **76**, 713–733 (2014)
13. Robert B., Gramacy: monomvn: Estimation for Multivariate Normal and Student-t Data with Monotone Missingness. R package version 1.9-7 (2017)
14. Sinay, M.S., Hsu, J.S.J.: Bayesian Inference of a Multivariate Regression Model. Journal of Probability and Statistics, vol. 2014, Article ID 673657, 13 pages (2014)

On the estimation of epidemiological parameters from serological survey data using Bayesian mixture modelling

Sulla stima di parametri epidemiologici da dati da indagini sierologiche tramite l'uso di modelli di mistura bayesiani

Emanuele Del Fava, Piero Manfredi, and Ziv Shkedy1

Abstract In the context of serological surveys for the estimation of important epidemiological functions, mixture models offer an alternative and more accurate approach than the conventional one based on the diagnostic assay's cut-off points. In this work, we propose an innovative Bayesian mixture model for the estimation of flexible models for the age-specific seroprevalence and force of infection (or incidence rate). In order to account for the possible waning of immunity by age, we propose a Bayesian variable selection approach to determine the best age-specific model for the mean and the variance of the seropositive subpopulation. Our methodology is applied to a sample of antibody titres to varicella from Italy. Our results confirm that mixture models are a flexible and highly customisable tool, adapt to be systematically used in serological surveys.

Abstract *Nel contesto delle indagini sierologiche per la stima di importanti parametri epidemiologici, i modelli di mistura rappresentano un approccio alternativo e più accurato di quello basato sull'uso dei valori critici associati ai test diagnostici. In questo lavoro, proponiamo un innovativo modello di mistura bayesiano, finalizzato alla stima di modelli flessibili per la sieroprevalenza e la forza dell'infezione (o tasso d'incidenza) dipendenti dall'età. Allo scopo di modellare il possibile declino degli anticorpi con l'età, proponiamo un metodo di selezione bayesiana di variabili per determinare il miglior modello, dipendente dall'età, per la media e la varianza della sottopopolazione dei sieropositivi. La nostra metodologia è applicata ad un campione di titoli anticorpali contro la*

[1] Emanuele Del Fava, Bocconi University, Milan; email: emanuele.delfava@unibocconi.it
Piero Manfredi, University of Pisa, Pisa, Italy; email: piero.manfredi@unipi.it
Ziv Shkedy, Hasselt University, Diepenbeek, Belgium; email: ziv.shkedy@uhasselt.be

varicella dall'Italia. I nostri risultati confermano che i modelli di mistura sono uno strumento flessibile e facilmente configurabile, da utilizzarsi in maniera sistematica nelle indagini sierologiche.

Key words: mixture models, serological data, Bayesian variable selection, MCMC

Introduction

Serological surveys, which are usually employed to quantify the antibody titre against a specific antigen, are among the most direct and informative techniques that are available to investigate the dynamics of the level of immunity protection in a certain population [12]. Despite that, the data resulting from these surveys remain somehow unexploited, for various reasons. One of them is that the assessment of the immunity profile in a population, the so-called seroprevalence of prevalence of immune individuals, is still often performed by dichotomisation of individual antibody titres measured by the diagnostic serological assay. This means that all the information contained in the individual antibody titre, such as the strength and the individual heterogeneity of the immunological response, is lost, as it is replaced by a binary variable giving the infection status, namely, whether the subject is seronegative or seropositive (showing evidence of past infection or vaccination). Moreover, the use of fixed cut-offs in serological surveys may be sub-optimal as the method is prone to misclassification or inconclusive classification [2, 10]. Conversely, mixture models are showing to be a more appropriate tool for the classification of serological antibody titres and for the estimation of age-specific epidemiological parameters, both from the statistical and the epidemiological point of view, in the case of infections in the pre-vaccination and post-vaccination state [3, 4, 14, 15].

The objective of this work is to provide an example of how to accurately estimating flexible models for the age-specific seroprevalence and force of infection (FOI) for varicella, using Bayesian mixture models for the classification of antibody titres and Bayesian variable selection for estimating the parameters associated with the seropositive subpopulation while accounting for possible immunity waning by age.

Data and Methods

Data on antibody titres for varicella zoster virus (VZV) in Italy were collected between 2003 and 2004 at national level [8]. Serological tests for VZV-specific IgG were performed using a commercial enzyme linked immunosorbent assay (ELISA), according to the manufacturer's guidelines. A sample of 2517 subjects, from 1 to 79 years of age, was collected. For each subject, the antibody titre (evaluated quantitatively as an antibody concentration and expressed as an optical density (OD) measured in mUI/mL) and the age were collected for the analysis. Children under 10 years old were oversampled.

We use Bayesian mixture models [5] in order to estimate the age-specific seroprevalence and the age-specific FOI directly from VZV antibody titres. We consider the population to be at demographic and epidemiological equilibrium. Since data were collected in a pre-vaccination period, we can safely assume that each serological sample is drawn from a population consisting of just two subpopulations, one for the seronegative and one for the seropositive individuals. We then assume that the individual antibody titre, after a logarithmic transformation, i.e. $\mathrm{Y_i} = \log_{10}(\mathrm{OD_i} + 1)$, is distributed as a mixture of two skew-normal distributions, with mixture weights depending on the age a of the individuals,

$$Y_i(a) = \left(1 - \pi_i(a)\right)SN(Y_i|\mu_1, \sigma_1^2, \gamma_1) + \pi_i(a)SN(Y_i|\mu_{2i}(a), \sigma_{2i}^2(a), \gamma_2)$$

where μ_k, σ_k^2, γ_k, $k = 1,2$ denote the mean, the variance, and the skewness parameters of the two mixture components, respectively. The mean and the variance of the seronegative component (μ_1, σ_1^2) are assumed to be age-independent, while those for the seropositive component (μ_2, σ_2^2) are allowed to vary by age, in order to account for possible waning of the antibody titre. The skew-normal distribution generalises the normal distribution by allowing for skewness through a specific parameter, γ_k, which we assume to be independent of age [1, 7]. A positive (negative) value of γ_k implies a distribution skewed to the right (left), thus a distribution with an excess of extremely high (low) antibody titres. The mixture weight of the immune component, $\pi(a)$, represents the age-specific seroprevalence, which is the expected proportion of immune individuals at exact age a in the given population [6]. Model estimation and inference are carried out by using Bayesian Markov Chain Monte Carlo (MCMC) methods.

For the parameters $\theta^{SP}(a) = (\mu^{SP}(a), \tau^{SP}(a))$, where τ is the precision, i.e., the reciprocal of the variance σ^2, we specify an age-specific piecewise-constant model, where the parameter θ^{SP} is constant within each of the T considered age groups $(a_{[t-1]}, a_{[t]})$:

$$\theta^{SP}(a) = \theta_1^{SP} + \sum_{t=2}^{T} \theta_{[t]}^{SP}\,, a \in \left(a_{[t-1]}, a_{[t]}\right).$$

Since we do not have any a priori knowledge about the direction of the changes across age of θ^{SP}, we use an unrestricted model, i.e. the age-dependent parameters $\theta_{[t]}^{SP}$ may either increase or decrease with respect to the preceding value $\theta_{[t-1]}^{SP}$. Since it would be too computer-intensive to fit all possible models and then select the best one using a selection criterion, we propose a Bayesian Variable Selection (BVS) approach to estimate the posterior probability of each possible model, and, in particular, the one for the model with the constant parameter [9, 11, 13].

As regards the seroprevalence $\boldsymbol{\pi(a)}$, we specify a Beta prior distribution for each age group $\boldsymbol{j}$, namely, $\boldsymbol{\pi}_{[j]} \sim \boldsymbol{Beta}(\boldsymbol{\alpha}_{[j]}, \boldsymbol{\beta}_{[j]})$, under the monotonically non-decreasing constraint $\boldsymbol{\pi}_{[j-1]} \leq \boldsymbol{\pi}_{[j]} \leq \boldsymbol{\pi}_{[j+1]}$, with the hyperparameters $\boldsymbol{\alpha}_{[j]}$ and $\boldsymbol{\beta}_{[j]}$ being given weakly-informative non-negative prior distributions. The order constraint is necessary to obtain a nonnegative estimate of the FOI. The ensuing posterior distribution of the age-specific seroprevalence is still distributed as a Beta, namely, $\boldsymbol{Beta}(\boldsymbol{Y}_{[j]} + \boldsymbol{\alpha}_{[j]}, \boldsymbol{n}_{[j]} - \boldsymbol{Y}_{[j]} + \boldsymbol{\beta}_{[j]})$, under the same order constraint. Given the estimate of the seroprevalence, the FOI is successively estimated by $\boldsymbol{\pi}'_{[j]}/(\boldsymbol{1} - \boldsymbol{\pi}_{[j]})$, where the first derivative of the prevalence at the numerator is approximated by $\left(\boldsymbol{4\pi}_{[j+1]} - \boldsymbol{3\pi}_{[j]} - \boldsymbol{\pi}_{[j+2]}\right)/\boldsymbol{2}$ and then smoothed.

Results

The VZV antibody titres data show a clear polarised distribution between the seronegative and the seropositive individuals (Figure 1A), with the seronegative cases being concentrated among children and, to a lesser extent, young adults. There is evidence for waning of the antibody protection after 40 years, as implied by the decrease in the mean of the seropositive component, with a posterior probability of 0.47 for this model (Figure 1A). There is also some evidence for a two-step waning model, showing a decrease in the 10-20 age group and again after 40 years of age, with a posterior probability of 0.38. Conversely, for what concerns the variance of the component, we do not reject the null hypothesis of the null model, as its posterior probability is 0.25. The negative estimates of the skewness for both components ($\gamma^{SN} = -1.81, \gamma^{SP} = -2.59$) imply an excess of high-value antibody titres.

The estimate of the seroprevalence (Figure 1B) shows a steep linear increase in the first ten years of age, up to around 80 % (much lower than in other European countries), followed by a slower increase in the following years. This pattern is reflected by the estimated age-specific FOI (Figure 1C), which peaks between 5 and 10 years (during primary school) and then declines to a plateau.

Table 1: Mean and standard deviation of the seronegative (SN) and the seropositive (SP) components of the skew-normal mixture model fitted to VZV data.

Age group	$\boldsymbol{\mu^{SN}}$	$\boldsymbol{\sigma^{SN}}$	$\boldsymbol{\mu^{SP}(a)}$	$\boldsymbol{\sigma^{SP}(a)}$
1-10	1.40 (1.37, 1.43)	0.34 (0.31, 0.37)	3.52 (3.44, 3.58)	0.62 (0.56, 0.70)
10-20	-	-	3.47 (3.42, 3.54)	0.63 (0.58, 0.71)
20-40	-	-	3.47 (3.40, 3.54)	0.63 (0.58, 0.71)
40+	-	-	3.33 (3.27, 3.39)	0.63 (0.58, 0.71)

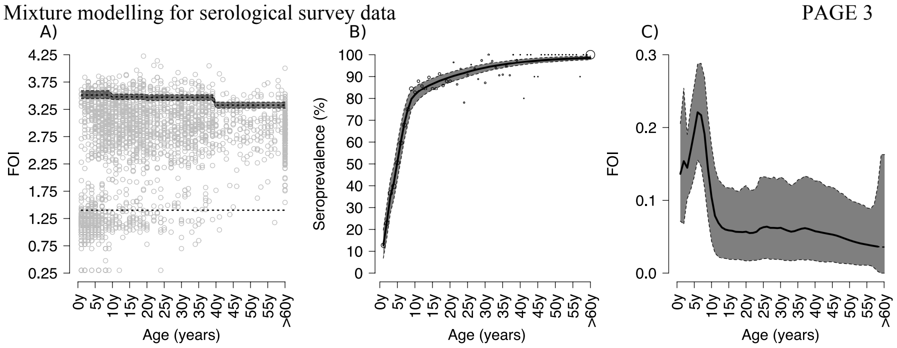


Figure 1: Fit of the mixture model to VZV data for Italy: A) scatter plot by age of antibody titres with over imposed the constant mean of the seronegative component and the age-dependent mean of the seropositive component (with 95% credible interval); B) age-specific seroprevalence with 95% credible interval; C) age-specific FOI with 95% credible interval.

Conclusions

In this work, we employed Bayesian mixture models to estimate key epidemiological parameters, such as the seroprevalence and the force of infection directly from antibody levels. Contrarily to the fixed cut-off approach, which leads to the estimation of these parameters based on binary infection status data, the mixture model adapts directly to the antibody data, without losing the information contained therein. Rather, the method allows for both the classification of individuals among different serological groups (by giving each subject an age-dependent probability of belonging to a specific group), also of those cases that would be classified as inconclusive under the fixed cut-off approach, and the estimation of the epidemiological parameters of interest. Moreover, the employment of a Bayesian MCMC approach allows to derive credible intervals around all the model parameters.

The model is also customisable enough to include a model conditional on age for the mixture parameters. The use of the BVS approach is computationally feasible and allows to show the support for all possible models in terms of posterior probability, even though one must pay attention to the variables under consideration, as the computational time dramatically increases with their number.

For all these reasons, we believe that the mixture modelling approach represents a flexible and highly customisable approach that should be seriously considered as the optimal way of analysing serological survey data. As a consequence, we claim that more attention should be devoted to the design of the serological surveys, both for what concerns the determination of the sample size by age and the measurement of the antibody titres.

References

1. Azzalini, A.: A Class of Distributions Which Includes the Normal Ones. Scand. J. Statist. **12**, 171—178 (1985) doi:10.2307/4615982.
2. Bollaerts, K., Aerts, M., Shkedy, Z., Faes, C., Van der Stede, Y., Beutels, Ph., Hens, N.: Estimating the Population Prevalence and Force of Infection Directly From Antibody Titres. Statist. Mod. **12**, 441—462 (2012) doi:10.1177/1471082X12457495.
3. Del Fava, E., Mirinaviciute, G., Flem, E., Freiesleben de Blasio, B., Scalia Tomba, G., Manfredi, P.: Estimating Age-Specific Immunity and Force of Infection of Varicella Zoster Virus in Norway Using Mixture Models. PLoS ONE **11**, e0163636—12 (2016) doi:10.1371/journal.pone.0163636.
4. Del Fava, E., Shkedy, Z., Bechini, A., Bonanni, P., Manfredi, P.: Towards Measles Elimination in Italy: Monitoring Herd Immunity by Bayesian Mixture Modelling of Serological Data. Epidemics **4**, 124—31 (2012) doi:10.1016/j.epidem.2012.05.001.
5. Diebolt, J., Robert, C.P.: Estimation of Finite Mixture Distributions Through Bayesian Sampling. J. R. Statist. Soc. B **56**, 363—75 (1994)
6. Evans, R.B., Erlandson, K.: Robust Bayesian Prediction of Subject Disease Status and Population Prevalence Using Several Similar Diagnostic Tests. Statist. Med. **23**, 2227—2236 (2004) doi:10.1002/sim.1792.
7. Frühwirth-Schnatter, S., Pyne, S.: Bayesian Inference for Finite Mixtures of Univariate and Multivariate Skew-Normal and Skew-T Distributions. Biostatistics **11**, 317—336 (2010) doi:10.1093/biostatistics/kxp062.
8. Gabutti, G., Rota, M.C., Guido, M., De Donno, A., Bella, A., Ciofi degli Atti, M.L., Crovari, P.: The Epidemiology of Varicella Zoster Virus Infection in Italy. BMC Pub. Health **8**, 372 (2008) doi:10.1186/1471-2458-8-372.
9. George, E., McCulloch, R.E.: Variable Selection via Gibbs Sampling. J. Am. Statist. Assoc. **88**, 881—889 (1993)
10. Hardelid, P., Williams, D., Dezateux, C., Tookey, P.A., Peckham, C.S., Cubitt, W.D., Cortina-Borja, M.: Analysis of Rubella Antibody Distribution From Newborn Dried Blood Spots Using Finite Mixture Models. Epidemiol. Infect. **136**, 1698 (2008) doi:10.1017/S0950268808000393.
11. Kasim, A., Shkedy, Z., Kato, B.S.: Estimation and Inference Under Simple Order Restrictions: Hierarchical Bayesian Approach. In: Dan L., Shkedy, Z., Yekutieli, D., Amaratunga, D., Bijnens (eds.) Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R. Springer, Heidelberg.
12. Metcalf, C.J.E., Farrar, J., Cutts, F.T., Basta, N.E., Graham, A.L, Lessler, J.T., Ferguson, N.M., Burke, D.S., Grenfell, B.T.: Use of Serological Surveys to Generate Key Insights Into the Changing Global Landscape of Infectious Disease. Lancet **388**, 728–730 (2016) doi:10.1016/S0140-6736(16)30164—7.
13. O'Hara, R.B., Sillanpää, M.J.: A Review of Bayesian Variable Selection Methods: What, How and Which. Bayes. Anal. **4**, 85—117 (2009) doi:10.1214/09-BA403.
14. Rota, M.C., Massari, M., Gabutti, G., Guido, M., De Donno, A., Ciofi degli Atti, M.L.: Measles Serological Survey in the Italian Population: Interpretation of Results Using Mixture Model. Vaccine **26**, 4403—4409 (2008) doi:10.1016/j.vaccine.2008.05.094.
15. Vyse, A.J., Gay, N.J., Hesketh, L.M., Morgan-Capner, P., Miller, E.: Seroprevalence of antibody to varicella zoster virus in England and Wales in children and young adults. Epidemiol. Infect. **132**, 1129-1134 (2004)

An evaluation of KL-optimum designs to discriminate between rival copula models

Una valutazione della capacità del disegno KL-ottimo di discriminare tra modelli copula rivali

Laura Deldossi, Silvia Angela Osmetti, Chiara Tommasi

Abstract The problem of model discrimination has prompted a great amount of research over last years. According to the specific characteristics of the rival models (nested, non-nested, linear or not) different optimum criteria have been proposed to select design points with the aim to discriminate between rival models. Ds-, T- and KL-criteria are the most known. Up to our knowledge, in the literature there is not any study to evaluate the performance of these discrimination criteria. In this work, via a simulation study and focusing on rival copula models, we analyze the performance of the KL-optimum design applying the likelihood ratio test for non-nested models.

Abstract *Nel corso degli ultimi anni il problema di discriminare tra modelli rivali ha prodotto una grande quantità di ricerche. A seconda della tipologia di modelli rivali (annidati, non annidati, lineari o non lineari), diversi criteri sono stati proposti con l'obiettivo di selezionare il disegno ottimo per la discriminazione. Tra i più noti ricordiamo i criteri Ds-, T- e KL-. Per quanto ci consta, in letteratura non esistono studi relativi alla valutazione della loro effettiva capacità discriminatoria. In questo lavoro, attraverso uno studio di simulazione in cui abbiamo applicato il test del rapporto di verosimiglianza per modelli non annidati, abbiamo analizzato le prestazioni del disegno KL-ottimo per discriminare tra modelli bivariati la cui struttura di dipendenza è descritta attraverso una funzione copula.*

Key words: Copula model, Cox's test, Optimal experimental design

Laura Deldossi and Silvia Angela Osmetti
Università Cattolica del Sacro Cuore, L.go Gemelli, 1 - Milano, e-mail: laura.deldossi@unicatt.it, silvia.osmetti@unicatt.it

Chiara Tommasi
Università degli studi di Milano, Via Conservatorio 7 - Milano, e-mail: chiara.tommasi@unimi.it

1 Introduction

A major limitation associated with the design of an experiment is that the optimality of a design depends on a priori true model that is not known in advance. Actually, very often, the experimenter has not just one but several possible models for describing a phenomenon. Thus, his/her first goal is to collect data in order to discriminate among rival models. The problem of model discrimination has prompted a great amount of research over last years. To discriminate between nested models (linear or not) [1] propose the D_s-criterion where the models are embedded in a more general one and the design aims at estimating the additional parameters as precisely as possible. A criterion to obtain optimal designs for discriminating between two homoscedastic models for normally distributed observations is T-optimality, which was introduced by [2]. A criterion based on the popular Kullback-Leibler (KL) distance is proposed by [6] for any non-normal assumption.

About discrimination between copula models, [9] apply the D_s-criterion which can be used only for nested models; for this reason, they need to introduce the mixture copula model (which includes the rival copulae as special cases). In this paper, instead, we consider the KL-optimality criterion proposed by [6] which compares directly the competing models without using any other auxiliary reference model. Specifically, we consider a bivariate model with two possible dependence structures: Clayton and Gumbel copulae (the competing models). Since, up to our knowledge, there are no studies to evaluate the performance of a discrimination criterion, in this work we analyze the performance of the KL-optimum design through a simulation study where we apply a version of Cox's test. For comparison purposes, we also describe the performance of the Uniform design, which is very often adopted in real case studies.

The paper is organized as follows. In Section 2 the bivariate copula model is introduced and the main definitions are given. The KL-optimality criterion is introduced in Section 3. Section 4 concerns the simulation study to evaluate the performance of the KL-optimum design.

2 Bivariate Copula-Based Model

Let (Y_1,Y_2) be a bivariate response variable with marginal distributions $F_{Y_1}(y_1;\alpha)$ and $F_{Y_2}(y_2;\beta)$, which depend on the unknown parameter vectors α and β, respectively. If there is an association between Y_1 and Y_2, it is necessary to define a joint model for (Y_1,Y_2).

A bivariate copula is a function $C: I^2 \to I$, with $I^2=[0,1]\times[0,1]$ and $I=[0,1]$, that, with an appropriate extension of the domain in R^2, satisfies all the properties of a cumulative distribution function (cdf). In particular, it is the cdf of a bivariate random variable (U_1,U_2), with uniform marginal distributions in $[0,1]$:

$$C(u_1,u_2;\theta_C) = P(U_1 \le u_1, U_2 \le u_2;\theta_C), \quad 0 \le u_1 \le 1 \quad 0 \le u_2 \le 1, \tag{1}$$

where $\theta_C \in \Theta_C$ is a parameter measuring the dependence between U_1 and U_2.

The importance of copulae in statistical modelling stems from Sklar's theorem [7], which states that a joint distribution can be expressed in terms of marginal distributions and a function $C(\cdot,\cdot;\theta_C)$ that binds them together. In more detail, according to Sklar's theorem, if $F_{Y_1,Y_2}(y_1,y_2;\delta,\theta_C)$ is the joint cdf of (Y_1,Y_2), where $\delta = (\alpha,\beta)$, then there exists a copula function $C\colon I^2 \to I$ such that

$$F_{Y_1,Y_2}(y_1,y_2;\delta,\theta_C) = C\{F_{Y_1}(y_1;\alpha),F_{Y_2}(y_2;\beta);\theta_C\}, \quad y_1,y_2 \in \mathbb{R}. \qquad (2)$$

If $F_{Y_1}(y_1;\alpha)$ and $F_{Y_2}(y_2;\beta)$ are continuous functions then the copula $C(\cdot,\cdot;\theta_C)$ is unique. Conversely, if $C(\cdot,\cdot;\theta_C)$ is a copula function and $F_{Y_1}(y_1;\alpha)$ and $F_{Y_2}(y_2;\beta)$ are marginal cdfs, then $F_{Y_1,Y_2}(y_1,y_2;\delta,\theta_C)$ given in (2) is a joint cdf.

From (2) we have that a copula captures the dependence structure between the marginal probabilities. This idea allows researchers to consider marginal distributions and the dependence between them as two separate but related issues. For each copula there exists a relationship between the parameter θ_C and Kendall's τ coefficient (see [7] pp. 158-170) and between θ_C and the lower and upper tail dependence parameters λ_l and λ_u (which measure the association in the tails of the joint distribution; see [7] pp. 214-216). Several bivariate copulae have been proposed in the literature (see for instance [7]). In this paper we consider only Clayton and Gumbel copulae, which are recalled in Table 1. Both these copulae allow only for positive association between variables ($\tau \geq 0$) but they exhibit strong *left* and strong *right* tail dependence, respectively.

3 KL-Optimality Criterion

An approximate design ξ is a discrete probability measure on a compact experimental domain $\mathscr{X}$; $\xi(x)$ represents (at least approximatively) the proportion of observations to be taken at the experimental condition x. An optimal design maximizes a concave functional of ξ, which is called optimality criterion and reflects an inferential goal.

Let (Cl,G) denote Clayton and Gumbel copulae, respectively and let (θ_{Cl},θ_G) be the corresponding dependence parameters. From now on, we assume that nominal values for δ, θ_{Cl} and θ_G are available; hence, we compute locally optimum designs. In order to discriminate between the two rival copulae, we propose to use

Table 1 Copula functions and related association parameters

Copula	$C(u_1,u_2;\theta)$	$\theta \in \Theta$	$\tau = \tau(\theta)$
Clayton	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	$\theta \in (0,\infty)$	$\tau = \theta/(\theta+2)$
Gumbel	$\exp\left(-\left[\{-\ln(u_1)\}^{\theta} + \{-\ln(u_2)\}^{\theta}\right]^{1/\theta}\right)$	$\theta \in [1,\infty)$	$\tau = 1 - 1/\theta$

the following geometric mean of KL-efficiencies:

$$\Phi_{KL}(\xi;\delta,\theta_{Cl},\theta_G) = \left\{\mathrm{Eff}_{G,Cl}(\xi;\theta_{Cl})\right\}^{\gamma}\cdot\left\{\mathrm{Eff}_{Cl,G}(\xi;\theta_G)\right\}^{1-\gamma} \quad 0\le\gamma\le 1,$$

where γ is a suitably chosen weight which balances the belief in the two competing copulae;

$$\mathrm{Eff}_{i,j}(\xi;\theta_j) = \frac{I_{i,j}(\xi;\theta_j)}{I_{i,j}(\xi^*_{i,j};\theta_j)}, \quad \xi^*_{ij} = \arg\max_{\xi} I_{i,j}(\xi;\theta_j), \quad i,j = Cl,G \tag{3}$$

and

$$I_{i,j}(\xi;\theta_j) = \inf_{\theta_i} \sum_{x\in\mathscr{X}} \mathscr{I}\left\{f^{j}_{y_1y_2}(x;\delta,\theta_j), f^{i}_{y_1y_2}(x;\delta,\theta_i)\right\}\,\xi(x), \tag{4}$$

is the KL-criterion proposed by [6]. Here $\mathscr{I}\left\{f^{j}_{y_1y_2}(x;\delta,\theta_j), f^{i}_{y_1y_2}(x;\delta,\theta_i)\right\}$ denotes the Kullback-Leibler divergence between the true density function $f^{j}_{y_1y_2}(x;\delta,\theta_j)$ and the rival one $f^{i}_{y_1y_2}(x;\delta,\theta_i)$, with $i,j = Cl,G$.

4 Evaluation of the performance of the KL-optimum design: an example with bivariate binary logistic model

In order to assess the ability of the KL-optimum design to discriminate between two competing copula models we employ a version of Cox's test (see [3] and [4]).

Given δ, τ and a design ξ, let (y_{1i}, y_{2i}) for $i = 1,2,...n$ be a sample of outcomes from one of the two rival models. For a specific Scenario δ and for a specific value of Kendall's τ coefficient, we generate M samples of size n, at a design ξ. Then, we check how many times the likelihood ratio test provides an evidence in favour of each model. Following [8] we have to test both the following systems of hypotheses:

$$A)\ \begin{cases} H_{Cl}:\ \mathscr{F}_{Cl} = \{f^{Cl}_{y_1y_2}(x;\delta,\theta_{Cl}),\ \theta_{Cl}\in\Theta_{Cl}\} \\ H_G:\ \mathscr{F}_G = \{f^{G}_{y_1y_2}(x;\delta,\theta_G),\ \theta_G\in\Theta_G\} \end{cases}$$

$$B)\ \begin{cases} H_G:\ \mathscr{F}_G = \{f^{G}_{y_1y_2}(x;\delta,\theta_G),\ \theta_G\in\Theta_G\} \\ H_{Cl}:\ \mathscr{F}_{Cl} = \{f^{Cl}_{y_1y_2}(x;\delta,\theta_{Cl}),\ \theta_{Cl}\in\Theta_{Cl}\} \end{cases}$$

From now on, we omit the argument x and δ for ease of notation. As test statistics, we consider the log-likelihood ratios:

$$T_{ClG} = L_{Cl}(\widehat{\theta}_{Cl}) - L_G(\widehat{\theta}_G) \quad \text{and} \quad T_{GCl} = L_G(\widehat{\theta}_G) - L_{Cl}(\widehat{\theta}_{Cl}), \tag{5}$$

where $L_{Cl}(\theta_{Cl})$ and $L_G(\theta_G)$ are the log-likelihood functions under H_{Cl} and H_G, respectively, and $\widehat{\theta}_{Cl}$ and $\widehat{\theta}_G$ are the corresponding maximum likelihood estimators.

Let p_{ClG} and p_{GCl} be the p-values of T_{ClG} and T_{GCl}, respectively. Whenever $p_{ClG} > p_{GCl}$ (or $p_{GCl} > p_{ClG}$) we accept Clayton (or Gumbel) model.

In the case of non-nested models the log-likelihood ratio is not (asymptotically) distributed as a Chi-squared random variable (see for instance [4, 8]). Hence, we implement a Monte Carlo procedure to approximate the sample distribution of T_{ClG} and T_{GCl} and to compute the corresponding p-values, $\hat{p}_{ClG}$ and $\hat{p}_{GCl}$ under H_{Cl} and H_G, respectively. Differently, [3, 4] proposed the asymptotic distribution of the log-likelihood ratio suitably standardized.

Consider now an example in dose finding study. Let (Y_1, Y_2) be a binary response variable where both Y_1 and Y_2 take values in $\{0, 1\}$ (1 denotes occurrence and 0 denotes no occurrence). We consider (see [5]) the following logistic models for the marginal success probabilities of efficacy and toxicity:

$$\pi_1(x;\alpha) = P(Y_1 = 1|x;\alpha) = \frac{e^{\alpha_0+\alpha_1 x+\alpha_2 x^2}}{1+e^{\alpha_0+\alpha_1 x+\alpha_2 x^2}}, \quad \alpha = (\alpha_0, \alpha_1, \alpha_2),$$
$$\pi_2(x;\beta) = P(Y_2 = 1|x;\beta) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}, \quad \beta = (\beta_0, \beta_1),$$

where $x \in \mathscr{X} = [-1, 1]$ denotes the standardized dose of a drug.

If $C(\cdot,\cdot;\theta_C)$ is a copula function which models the dependence between $\pi_1(x;\alpha)$ and $\pi_2(x;\beta)$, then the joint probability of $(Y_1$,$Y_2)$ at the dose x is

$$p^C_{11}(x;\delta,\theta_C) = P(Y_1 = 1, Y_2 = 1|x;\delta,\theta_C) = C\{\pi_1(x;\alpha), \pi_2(x;\beta);\theta_C\}. \qquad (6)$$

Given $\delta = (1, 1.5, -3, 2.5, 5)$ and $\tau = 0.8$, we perform two Monte Carlo simulations, based on the generation of $M = 5000$ samples of size n from model (6) using (in the data generating model) the Clayton copula with $\theta_{Cl} = 8$ and the Gumbel copula with $\theta_G = 5$, respectively ($\theta_{Cl} = 8$ and $\theta_G = 5$ correspond to the same value of the association parameter $\tau = 0.8$). The doses and the proportions of observations to be taken at each dose are given by the KL-optimum design, which is reported in the first column of Table 2.

Table 2 KL-optimal design ξ_{KL} for $(\theta_{Cl};\theta_G) = (8;5)$ and Uniform design ξ_{Unif}

ξ_{KL}	ξ_{Unif}
$\left\{\begin{array}{cc} -0.793 & -0.050 \\ 0.470 & 0.530 \end{array}\right\}$	$\left\{\begin{array}{ccccc} -1 & -0.5 & 0 & 0.5 & 1 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{array}\right\}$

We apply the likelihood test and compute the Monte Carlo p-values of T_{GCl} and T_{ClG}: p^m_{ClG} and p^m_{ClG} for $m = 1, \ldots, M$. We calculate the percentages of correct selection of the true model, i.e. the percentage of times that $p^m_{ClG} > p^m_{GCl}$ for $m = 1, \ldots, M$, when the data are generated from the Clayton copula, and the percentage of times that $p^m_{GCl} > p^m_{ClG}$ for $m = 1, \ldots, M$, when the data are generated from the Gumbel copula. The results are reported in the third and the fourth columns of Table 3.

We can observe that using the KL-optimum design the percentage of correct decision is around 72% from $n = 100$ and it exceed 90% from $n = 500$. Furthermore, the percentage of wrong decision decreases substantially as n increases. Taking into

Table 3 Monte Carlo simulation of the likelihood ratio test ($M = 5000$): data generated from Clayton and Gumbel copulae for $\tau = 0.8$ at the KL-optimum design ξ_{KL} (columns 3-4) and at the Uniform design ξ_{Unif} (columns 5-6)

n	**Test decision**	ξ_{KL} **True copula model (%)**		ξ_{Unif} **True copula model (%)**	
		Clayton	Gumbel	Clayton	Gumbel
100	Correct decision	72	71.88	60.54	32.64
	Wrong decision	28	28.12	39.46	67.36
200	Correct decision	82.28	84.04	67.50	43.12
	Wrong decision	17.72	15.96	32.50	56.88
500	Correct decision	95.56	95.4	81.20	63.38
	Wrong decision	4.44	4.6	18.80	36.62
1000	Correct decision	99.5	99.2	92.84	81.14
	Wrong decision	0.5	0.8	7.16	18.86

account that the competing models differ only for the tail dependence, the obtained results are excellent. Finally, for comparison purposes, we analyze the performance of the Uniform design defined in the second column of Table 2. The corresponding percentages of correct decision and wrong decision are listed in the fifth and sixth columns of Table 3. We can observe that the percentage of correct selections obtained with the KL-optimum design is substantially better than those corresponding to the uniform design, especially for $n < 500$.

References

1. Atkinson, A. C., Cox, D.R. : Planning experiments for discriminating between models. J.R. Statist. Soc. B, **36**, 321–348 (1974)
2. Atkinson A. C. and Fedorov V. V.: The Design of Experiments for Discriminating Between two Rival Models. Biometrika, **62**, 57–70 (1975)
3. Cox, D.R.: Tests of separate families of hypotheses, In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistic and Probability, University of California Press: Berkeley, 105–123 (1961)
4. Cox, D.R.: Further results on tests of separate families of hypotheses.Journal of the Royal Statistical Society B, **24**, 406–424 (1962)
5. Deldossi, L. and Osmetti, S. A. and Tommasi, C.: PKL-Optimality Criterion in Copula Models for efficacy-toxicity response, In: mODa 11 - Advances in Model-Oriented Design and Analysis, Kunert, J., Muller, C.H., Atkinson, A.C. (eds), Springer International Publishing: Heidelberg, 79–86 (2016)
6. López-Fidalgo, L.J., Tommasi, C., Trandafir, P.C.: An optimal experimental design criterion for discriminating between non-Normal models. Journal of the Royal Statistical Society B **69**(2), 231–242 (2007)
7. Nelsen, R.B., An Introduction to Copulas. Springer, New York (2006)
8. Pesaran, H. and Weeks, M.: Non-nested Hypothesis Testing: An Overview. In: A Companion to Theoretical Econometrics,BH Baltagi (eds), 279–309 (2001).
9. Perrone, E. and Müller, W.G.: Optimal designs for copula models. Statistics, **50**, 917–929, (2016)

Variational Approximations for Frequentist and Bayesian Inference

Approssimazioni Variazionali per Inferenza Frequentista e Bayesiana

Luca Maestrini and Matt P. Wand

Abstract Variational approximations are a flexible instrument for deterministic approximate inference in complex statistical models. We illustrate the concept of variational approximation from both frequentist and Bayesian perspectives, providing methodological examples that take advantage of the classical concepts of exponential families.
Sommario *Le approssimazioni variazionali sono uno strumento flessibile per l'inferenza su modelli statistici complessi. Viene illustrato il concetto di approssimazione variazionale da punti di vista frequentista e Bayesiano, proponendo esempi metodologici che fanno leva sulla classica teoria delle famiglie esponenziali.*

Key words: Bayesian inference, frequentist inference, Gaussian variational approximation, variational message passing.

1 Introduction

Variational approximations is a class of techniques for deterministic approximations which is now part of mainstream computer science and machine learning methodology. Applications cover a wide area of elaborate problems such as those arising in speech recognition, graphical models, document retrieval or genetic linkage analysis [2]. These methods are also widening their presence in statistics as a response to the increasing complexity of models in modern statistical applications [4].

We describe the concept of variational approximation referring to a Bayesian model. In keeping with the statistics literature on variational approximations, let p

Luca Maestrini
Department of Statistical Sciences, University of Padova, Italy
e-mail: luca.maestrini@phd.unipd.it

Matt P. Wand
School of Mathematical and Physical Sciences, University of Technology Sydney, Australia

be the generic symbol for a density function, denote with $\mathbf{y}$ the observed data, $\theta \in \Theta$ the parameters to be inferred and let q be an arbitrary density function over Θ.

The logarithm of the marginal likelihood satisfies

$$\begin{aligned}\log p(\mathbf{y}) &= \int q(\theta)\log\left\{\frac{p(\mathbf{y},\theta)}{q(\theta)}\right\}d\theta + \int q(\theta)\log\left\{\frac{q(\theta)}{p(\theta|\mathbf{y})}\right\}d\theta \\ &\geq \int q(\theta)\log\left\{\frac{p(\mathbf{y},\theta)}{q(\theta)}\right\}d\theta, \qquad (1)\end{aligned}$$

giving a lower bound $\underline{p}(\mathbf{y};q)$ on the marginal likelihood such that

$$p(\mathbf{y}) \geq \underline{p}(\mathbf{y};q) \equiv \exp\int q(\theta)\log\left\{\frac{p(\mathbf{y},\theta)}{q(\theta)}\right\}d\theta. \qquad (2)$$

Maximization of $\underline{p}(\mathbf{y};q)$ is equivalent to minimization of the Kullback–Leibler divergence between $q(\theta)$ and $p(\theta|\mathbf{y})$. The key idea of variational approximations is to approximate the posterior density $p(\theta|\mathbf{y})$, or the likelihood function itself in the frequentist case, by a $q(\theta)$ for which $\underline{p}(\mathbf{y};q)$ is more tractable than $p(\mathbf{y})$ and obtain approximate estimates through lower bound maximization. Tractability is achieved by restricting $q(\theta)$ to a more manageable class of densities. Common restrictions for the approximating density are:

a. $q(\theta)$ is a member of a parametric family of density functions;
b. $q(\theta)$ factorizes into $\prod_{i=1}^{M} q_i(\theta_i)$, for some partition $\{\theta_1,\ldots,\theta_M\}$ of θ.

We apply restrictions (a) and (b) to describe frequentist and Bayesian methodologies respectively which are known as Gaussian variational approximation and variational message passing.

2 Gaussian Variational Approximation

Frequentist models that stand to benefit from variational approximations are those for which the likelihood specification involves conditioning on a vector of latent variables $\mathbf{u}$. Given a log-likelihood of the model parameter vector θ

$$\ell(\theta) \equiv \log p(\mathbf{y};\theta) = \log\int p(\mathbf{y}|\mathbf{u};\theta)\,p(\mathbf{u};\theta)\,d\mathbf{u},$$

interest is in obtaining $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\,\ell(\theta)$, maximum likelihood estimate of θ.

In practice, $\ell(\theta)$ may not be available in closed form because of analytically intractable integration. In such circumstances and depending on the forms of $p(\mathbf{y}|\mathbf{u};\theta)$ and $p(\mathbf{u};\theta)$, variational approximations can provide a more amenable approximation. However, nontrivial frequentist examples where an explicit solution arises by applying a product density methodology as in (b) are not known.

Suppose instead to restrict q to a parametric family of densities $\{q(\mathbf{u};\xi) : \xi \in \boldsymbol{\Xi}\}$, similarly to (a). Then, similarly to (1) we can define the log-likelihood lower bound

$$\underline{\ell}(\theta,\xi;q) \equiv \int q(\mathbf{u};\xi)\log\left\{\frac{p(\mathbf{y},\mathbf{u};\theta)}{q(\mathbf{u};\xi)}\right\}d\mathbf{u}$$

and a new maximization problem

$$\left(\hat{\underline{\theta}},\hat{\underline{\xi}}\right) = \underset{\theta,\xi}{\operatorname{argmax}}\underline{\ell}(\theta,\xi;q),$$

with $\hat{\underline{\theta}}$ variational approximation to the maximum likelihood estimator. Furthermore, standard error estimates can be obtained by plugging in $\hat{\underline{\theta}}$ for θ and $\hat{\underline{\xi}}$ for ξ in the variational approximate Fisher information matrix arising from replacement of $\ell(\theta)$ by $\underline{\ell}(\theta,\xi;q)$.

In *Gaussian variational approximations* (GVA), $q(\mathbf{u};\xi)$ is assumed to be a multivariate normal density [5]. We investigate the application of GVA to generalized linear mixed models (GLMMs) for semiparametric regression.

Consider GLMMs within one-parameter exponential family

$$\mathbf{y}|\mathbf{u} \sim \exp\left\{\mathbf{y}^{\mathrm{T}}(\mathbf{X}\beta+\mathbf{Zu}) - \mathbf{1}^{\mathrm{T}}b(\mathbf{X}\beta+\mathbf{Zu}) + \mathbf{1}^{\mathrm{T}}c(\mathbf{y})\right\}, \quad \mathbf{u} \sim N(\mathbf{0},\mathbf{G})$$

where $\mathbf{X}$ and $\mathbf{Z}$ are general design matrices. Matrix $\mathbf{G}$ models random effect covariance, while $\mathbf{Z}$ can include, for instance, spline basis functions. The functions b and c characterize members of the exponential family.

Setting $q(\mathbf{u};\xi)$ to be the $N(\mu,\boldsymbol{\Lambda})$ we can derive variational lower bound

$$\begin{aligned}\underline{\ell}(\beta,\mathbf{G},\mu,\boldsymbol{\Lambda}) = {} & \tfrac{n}{2} + \mathbf{y}^{\mathrm{T}}(\mathbf{X}\beta+\mathbf{Z}\mu) - \mathbf{1}^{\mathrm{T}}B\left(\mathbf{X}\beta+\mathbf{Z}\mu, \mathrm{dg}\left(\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^{\mathrm{T}}\right)\right) + \mathbf{1}^{\mathrm{T}}c(\mathbf{y}) \\ & -\tfrac{1}{2}\left\{\mu^{\mathrm{T}}\mathbf{G}^{-1}\mu + \mathrm{tr}\left(\mathbf{G}^{-1}\boldsymbol{\Lambda}\right)\right\} + \tfrac{1}{2}\log\left|\mathbf{G}^{-1}\boldsymbol{\Lambda}\right|, \qquad (3)\end{aligned}$$

where n is the number of rows in $\mathbf{y}$, $B\left(\mu,\sigma^2\right) \equiv \int_{-\infty}^{\infty} b(\mu+\sigma x)\,\phi(x)\,dx$, $\phi(x)$ is the $N(0,1)$ density function and, for a square matrix $\mathbf{A}$, $\mathrm{dg}(\mathbf{A})$ is the column vector containing the diagonal entries of $\mathbf{A}$. For vector arguments, function B is applied in element-wise fashion. Inference and prediction on nonparametric, additive or general semiparametric models in GLMM form follow directly from the lower bound optimization.

3 Variational Message Passing

In Bayesian inference, a mean field variational approximation $q^*(\theta)$ is the maximizer of expression (2) subject to a product density restriction as in (b).

It can be shown that the optimal q-density functions satisfy

$$q^*(\theta_i) \propto E_{q(\theta\backslash\theta_i)}\{p(\theta_i|\mathbf{y},\theta\backslash\theta_i)\}, \quad 1 \le i \le M, \qquad (4)$$

where $\theta \backslash \theta_i$ denotes the entries of θ with θ_i omitted. Expression (4) gives rise to an iterative scheme for obtaining the parameters of the optimal density functions $q^*(\theta_i)$ which is known as mean field variational Bayes. *Variational message passing* (VMP) arrives at the same approximation via message passing on an appropriate factor graph. Among the several variants of VMP in the literature, we consider the factor graph fragment approach introduced in [8] and based on [3], whose major advantage is that calculations only need to be done once for a certain distribution family and can be easily adapted to accommodate more complex model structures. The use of conjugates exponential families streamlines the algebraic and computational effort in deriving messages between factor graph components at the base of VMP algorithms. A listing of such a procedure can be found in Sect. 2.5 of [8].

This framework gives rise to a class of VMP algorithms to approximate fitting and inference for a wide range of common and non-standard likelihoods.

4 Illustrations

The next two illustrative examples are applications in frequentist and Bayesian settings that witness the flexibility of variational approximations. The former provides approximate estimates for a simulated Poisson spline regression model, the latter concerns a skew t response regression model on real data.

4.1 GVA for Poisson Spline Regression

We simulate 500 observations from a Poisson process as a function of a $\text{Uniform}(0,1)$ covariate and estimate a Poisson semiparametric regression model with canonical link using O'Sullivan penalized splines [6] on 50 interior knots via GVA.

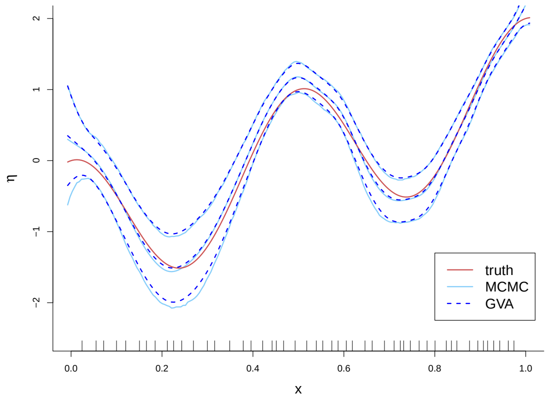


Fig. 1 Data generating process for the Poisson semiparametric regression. 95% MCMC credible intervals and GVA confidence bands are compared to the true η function that generates data according to a $\text{Poisson}(e^{\eta})$ distribution. Knot positions are also displayed.

Let $\hat{\boldsymbol{\Lambda}}_{\text{GVA}}$ be the estimate of $\boldsymbol{\Lambda}$ obtained maximizing the Gaussian variational lower bound (3) adapted to the current model. Given $\text{H}_{\mu\mu}\underline{\ell}$ Hessian matrix of $\underline{\ell}(\beta, \mathbf{G}, \mu, \boldsymbol{\Lambda})$ with respect to μ, one can prove that $\hat{\boldsymbol{\Lambda}}_{\text{GVA}} = \left(-\text{H}_{\mu\mu}\underline{\ell}\right)^{-1}$. We use this result and the estimates obtained through the lower bound optimization to derive the plot in Fig. 1, which concerns the true generating process. For a rough performance evaluation we plot GVA results as in (2) with those from Markov chain Monte Carlo (MCMC) samples obtained using the `R` package `rstan` [7] setting priors $N\left(0, 10^5\right)$ on β and Half-Cauchy$\left(10^5\right)$ on the scale parameter appearing from defining $\mathbf{G} \equiv \sigma^2\mathbf{I}$. GVA seems to adequately approximate the MCMC process prediction and credible intervals.

4.2 VMP for Skew t Regression

We illustrate the parameter estimation of a skew t regression model via VMP.

Consider the dataset examined in [1] with the linear model

$$y_i = \beta_0 + \beta_1 \text{CRSP}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Skew-t}\left(0, \sigma^2, \lambda, \nu\right), \quad 1 \leq i \leq 60$$

with λ parameter of symmetry and $\nu > 0$ degrees of freedom. The variables y_i and CRSP_i denote the Martin Marietta company excess rate and the return excess index for the New York Stock Exchange respectively. We adopt the skew t distribution described in [1] and write it in terms of standard normal and inverse χ^2 auxiliary variables to limit the complexity of algebraic derivations and numerical integration appearing in the derivation of a VMP algorithm. We choose a product density restriction on $q\left(\theta\right)$ which is a compromise between approximation performances and algebraic complexity. We approximate the parameter posterior densities with VMP

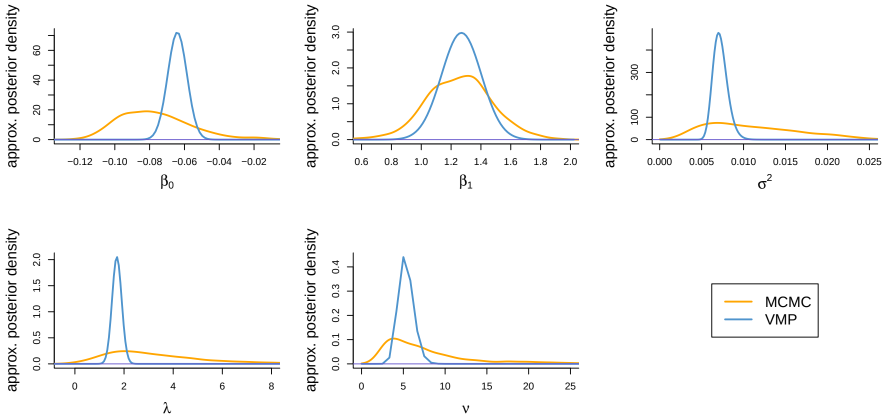


Fig. 2 Martin Marietta data: posterior density plots via MCMC and VMP.

and compare them to MCMC density estimation via `rstan`. The hyperparameters for β are fixed to $\mu_\beta = \mathbf{0}$ and $\boldsymbol{\Sigma}_\beta = 10^5\mathbf{I}$ over a prior $N\left(\mu_\beta, \boldsymbol{\Sigma}_\beta\right)$ while those on the shape parameters are Inverse-$\chi^2(0.01, 0.01)$ on the squared scale, $N\left(0, 10^5\right)$ on λ and $\Gamma\left(1, 0.01\right)$ on ν. Posterior density plots are shown in Fig. 2. VMP curves apparently underestimate the variance of MCMC posterior densities but locate around their modes.

Acknowledgements Luca Maestrini carried out his research activity during a visiting period at the School of Mathematical and Physical Sciences, University of Technology Sydney, Australia. We are grateful to Alessandra Salvan and Nicola Sartori for their comments on this research.

References

1. Azzalini, A., Capitanio, A.: Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. J. R. Stat. Soc. Ser. B. **65**, 367–389 (2003)
2. Jordan, M. I.: Graphical Models. Stat. Sci. **19**, 140–155 (2004)
3. Minka, T.: Divergence measures and message passing. Microsoft Res. Tech. Rep. Ser. **173**, 1–17 (2005)
4. Ormerod, J. T., Wand, M. P.: Explaining variational approximations. Am. Stat. **64**, 140–153 (2010)
5. Ormerod, J. T., Wand, M. P.: Gaussian variational approximate inference for generalized linear mixed models. J. Comput. Graph. Stat. **21**, 2–17 (2012)
6. O'Sullivan, F.: Nonparametric estimation of relative risk using splines and cross-validation. J. Sci. Stat. Comput. **9**, 531–542 (1988)
7. Stan Development Team: rStan: the R interface to Stan. R package version 2.17.3. http://mc-stan.org/ (2018)
8. Wand, M. P.: Fast approximate inference for arbitrarily large semiparametric regression models via message passing. J. Am. Stat. Assoc. **112**, 137–168 (2017)

Node-specific effects in latent space modelling of multidimensional networks

Effetti nodo-specifici nell'analisi a spazi latenti di reti multidimensionali

Silvia D'Angelo and Marco Alfò and Thomas Brendan Murphy

Abstract Social network analysis is a growing and popular field in statistics since the second half of the last century. Although single networks have been largely studied and a variety of models has been developed for their analysis, multidimensional networks are still a young and quite unexplored subject. In most cases, the attention has been focused on the specific case of dynamic networks. The aim of the present work is to provide an extension of latent space models for network analysis to the case of multidimensional networks. We also introduce node-specific effects (sender and receiver effects), typical to the single network literature, to allow for a more flexible representation of the multidimensional network. Finally, a real data application will be presented.
Abstract A partire dalla metà dello scorso secolo, l'analisi delle reti sociali è un campo popolare e in pieno sviluppo in statistica. Sebbene le reti singole siano state ampiamente studiate e siano disponibili una larga varietà di modelli per la loro analisi, l'estensione al caso di reti multidimensionali è ad oggi poco esplorata. Nella maggior parte dei casi, l'attenzione è stata rivolta al caso particolare delle reti dinamiche. In questo lavoro presenteremo un'estensione dei modelli a spazi latenti alle reti multidimensionali. Inoltre, introdurremo effetti nodo-specifici (sender e receiver), tipici in letteratura nel caso di reti singole, per descrivere un modello flessibile di reti multidimensionalli. Infine, verrà presentata un'applicazione a dati reali.

Key words: Multidimensional networks, Latent space models

Silvia D'Angelo
Sapienza, Università di Roma, Piazzale Aldo Moro 5, Roma. e-mail: silvia.dangelo@uniroma1.it

Marco Alfò
Sapienza, Università di Roma, Piazzale Aldo Moro 5, Roma. e-mail: marco.alfo@uniroma1.it

Thomas Brendan Murphy
University College Dublin, Belfield, Dublin 4, Ireland. e-mail: brendan.murphy@ucd.ie

1 Introduction

Social network analysis is a well known and vibrant branch of statistics. As network structures arise in many different contests, network analysis has seen application in a broad variety of fields.

In general, a network is defined by a set of units (the nodes) among which a relation can be established. In the most simple case, the relation to be recorded between a pair of nodes (a dyad), is either present or not present. If it is present, the dyad is said to be linked by an edge. A large variety of models for the analysis of single networks (where a single relation is recorded) has been proposed in the literature (see [7] for a review) and has paved the way for the analysis of more complex network structures, such as multidimensional networks. Recent works on multidimensional networks are those of Gollini and Murphy [1], D'Angelo et al. [6], Salter-Townshend and McCormick [8], Durante et al. [9] and Sewell and Chen [10]. In the present work, we illustrate a model to describe the ties observed in a multidimensional network by means of underlying similarities among the nodes and node-specific characteristics. Section 2 describes the proposed model, while section 3 presents an application of the model to the well known Vickers data [3].

2 The model

Let us define a multidimensional network (or multiplex) as a collection of K networks (or views), defined on the same node set N. Each network defines a different relation among the nodes in N. That is, the edge set $E^{(k)}$, $k = 1, \dots, K$ may be different in each view. A multidimensional network can be represented by means of graph theory as a collection of graphs (G), defined on a constant node set N:

$$G = \left(N, \left\{E^{(k)}; k = 1, \dots, K\right\}\right),$$

where a single network is $G^{(k)} = \left(N, E^{(k)}\right)$, $k = 1, \dots, K$. If the multiplex collects binary relations, that is whether something is verified or it is not, the realization of G will be denoted by a collection of adjacency matrices

$$\mathbf{Y} = \left\{y^{(1)}, \dots, y^{(k)}, \dots, y^{(K)}\right\}, \quad k = 1, \dots, K,$$

with n the number of observed nodes. The general entry of the k^{th} matrix will be $y_{ij}^{(k)} = 1$ if the relation k is present between nodes i and j and $y_{ij}^{(k)} = 0$ otherwise. Our aim is at describing the association structure underlying the multidimensional network and at estimating edge probabilities.

A first approach in the context of latent space models for networks, has been introduced by Hoff and Raftery [2]. This class of model assumes that the probability of an edge between two nodes depends on how similar they are;

this similarity refers to their distance in a so called latent (or social) space. The coordinates of the nodes in this space (and therefore the distances) are unknown and the aim is at recovering them to depict the similarities among the nodes and reconstruct the edge probabilities. This class of model was partially extended to the context of multidimensional networks by Gollini and Murphy [1] and D'Angelo et al. [6].

As we are modelling binary adjacency matrices, it is reasonable to model edge probabilities with a logistic regression (as in [6]):

$$Pr\left[y_{ij}^{(k)} = 1 \mid d(z_i,z_j), \alpha^{(k)}, \beta^{(k)}\right] = \frac{\exp\left\{f\left(\alpha^{(k)}, \beta^{(k)}, d(z_i,z_j)\right)\right\}}{1+\exp\left\{f\left(\alpha^{(k)}, \beta^{(k)}, d(z_i,z_j)\right)\right\}} \tag{1}$$

where z_i and z_j are the latent coordinates of nodes i and j and $d(\cdot)$ is the squared Euclidean distance. As the aim is at recovering the similarities among the nodes, the latent space is supposed to be common to all the networks. The sets of parameters $\boldsymbol{\alpha} = \left(\alpha^{(1)}, \dots, \alpha^{(K)}\right)$ and $\boldsymbol{\beta} = \left(\beta^{(1)}, \dots, \beta^{(K)}\right)$ help to distinguish network connectivity and, in general, link probabilities in the different networks.

We may define a more flexible specification of the edge probabilities by introducing node-specific parameters, to account for the direction of the edge in direct multiplex:

$$Pr\left[y_{ij}^{(k)} = 1 \mid d(z_i,z_j), \alpha^{(k)}, \beta^{(k)}, \theta_i^{(k)}, \gamma_j^{(k)}\right] = \frac{\exp\left\{f\left(\alpha^{(k)}, \beta^{(k)}, d(z_i,z_j), \theta_i^{(k)}, \gamma_j^{(k)}\right)\right\}}{1+\exp\left\{f\left(\alpha^{(k)}, \beta^{(k)}, d(z_i,z_j), \theta_i^{(k)}, \gamma_j^{(k)}\right)\right\}} \tag{2}$$

The set of parameters $\boldsymbol{\Gamma} = \left(\boldsymbol{\Gamma}^{(1)}, \dots, \boldsymbol{\Gamma}^{(K)}\right)$, where $\boldsymbol{\Gamma}^{(k)} = \left(\gamma_1^{(k)}, \dots, \gamma_j^{(k)}, \dots, \gamma_n^{(k)}\right)$ and $\boldsymbol{\Theta} = \left(\boldsymbol{\Theta}^{(1)}, \dots, \boldsymbol{\Theta}^{(K)}\right)$, where $\boldsymbol{\Theta}^{(k)} = \left(\theta_1^{(k)}, \dots, \theta_i^{(k)}, \dots, \theta_n^{(k)}\right)$ describe, respectively, receiver and sender effects (see for example Krivitsky et al. [5]). Within each view, sender and receiver parameters are assumed to have a multiplicative effect on the intercept parameter $\alpha^{(k)}$ and to be bounded in $[-1,1]$.

Estimation of the proposed model parameters is carried out by employing a hierarchical Bayesian framework, using a Metropolis within Gibbs algorithm.

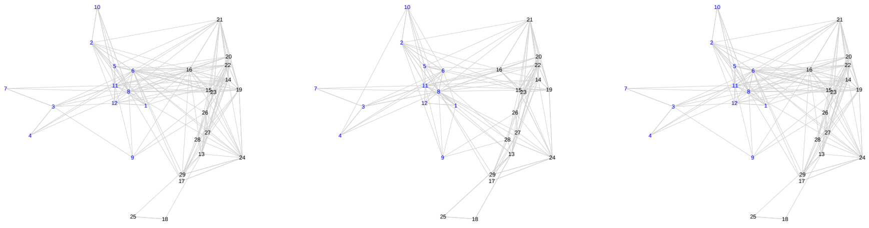
3 Application

The model proposed in section 2 has been applied to a benchmark dataset, the Vickers data [3], see for example [4]. In this case three different kind of relations among 29 students have been collected. In particular, the three corresponding networks describe:

1. if student i gets on with student j,
2. if student i is best friend with student j,
3. if student i works with student j.

None of the networks is sparse and the observed densities are, respectively, 0.445, 0.223 and 0.244. The observed out-degree (the number of ties a node sends) and in-degree (the amount of ties a node attracts) distributions in the three networks show that only a few nodes interact with most of the others, while the majority of them is less active.

Figure 1 shows the estimated latent positions of the students. The nodes have been coloured according to students gender, with blue and black numbers representing, respectively, males and females. A group structure seems to emerge in the latent space, with reference to the gender. This is confirmed by the heatmap in figure 2, which represents the estimated distances among the nodes and exhibits a clear block structure. Indeed, it seems that male students interact more with other male students, while females prefer the company of other females, exception made for a small group of nodes. These are three boys (5, 8, 11) and two girls (16, 21), all lying in the centre of the latent space. These five serve as a bridge between the two groups in the class. Figure 3 shows the estimated sender (and receiver) effects, together with the observed out-degrees (and in-degrees), for each network. The estimated coefficients are in line with what observed in the data, showing that the proposed model is a good candidate in representing the Vickers multiplex.



(a) (b) (c)

Fig. 1: Estimated latent coordinates for the nodes and observed edges. The relations represented are: get on well (figure *a*), best friend (figure *b*) and work with (figure *c*). Black numbers correspond to female students and blue numbers to males.

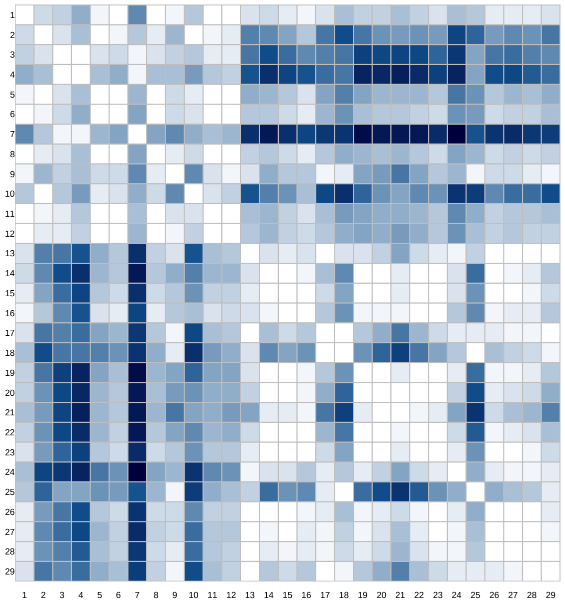
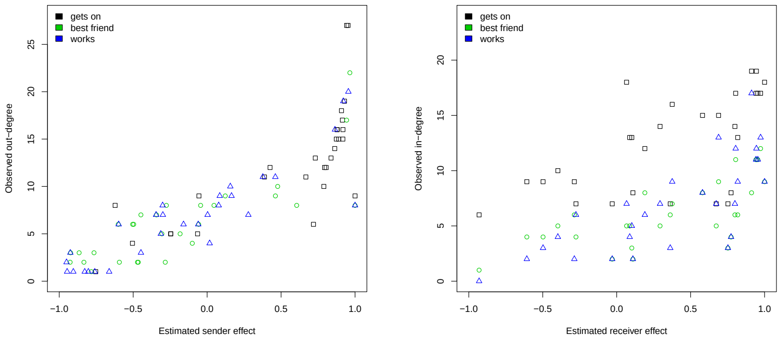


Fig. 2: Estimated distances between the nodes in the latent space. Nodes $1-12$ are male students, while nodes $13-29$ are females.



(a) (b)

Fig. 3: Estimated sender effects vs. observed out-degrees (figure *a*) and estimated receiver effects vs observed in-degrees (figure *b*) in the Vickers seventh grade networks.

4 Discussion

In this work we have proposed a novel approach to model multidimensional networks that builds both on latent space model for networks and on two typical instruments of social networks analysis: sender and receiver effects. A single latent space is employed to model the similarities between the nodes and node-specific parameters are introduced to model the possible presence of heterogeneity across the multiple networks. This approach allows a flexible reconstruction of the edge probabilities and could serve as an alternative to modelling each network of the multiplex via its own latent space [1]. Indeed, the use of sender and receiver effects prevents from choosing the dimensions of the latent spaces, which could be a relevant issue in some contexts. Unidimensional node-specific parameters would capture network-specific behaviours while the common latent space would represent the overall relationships between the actors in the multiplex. Also, a single latent space allows a straightforward visualization of the multiplex, a useful feature, especially when the number of views is high.

References

1. Gollini, I., Murphy, T.B.: Joint Modeling of Multiple Network Views. Journal of Computational and Graphical Statistics. 25, 246–265 (2016)
2. Hoff, P.D., Raftery, A.E.: Latent space approaches to social network analysis. Journal of the American Statistical Association. 97, 1090–1098 (2002)
3. Vickers, M., Chan, S.: Representing Classroom Social Structure. Melbourne: Victoria Institute of Secondary Education. (1981)
4. Wasserman, S., Pattinson, P.: Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. Psychometrika. 61, 401–425 (1996)
5. Krivitsky, P.N., Handcock, M.S., Raftery, A.E., Hoff, P.D.: Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. Social Networks. 31, 204–213 (2009)
6. D'Angelo, S., Murphy, T.B., Alfò, M.: Latent space modeling of multidimensional networks with application to the exchange of votes in Eurovision Song Contest. arXiv:1803.07166. (2018)
7. Salter-Townshend, M., White, A., Gollini, I., Murphy, T. B.: Review of statistical network analysis: models, algorithms, and software. Statistical Analysis and Data Mining. 5, 243-–264 (2012).
8. Salter-Townshend, M., McCormick, T. H.: Latent space models for multiview network data. Annals of Applied Statistics. 11, 1217-–1244 (2017).
9. Durante, D., Dunson, D.B., Vogelstein, J.T.: Nonparametric Bayes modeling of populations of networks. Journal of the American Statistical Association. 112, 1516–1530 (2017).
10. Sewell, D.K, Chen, Y.: Latent space models for dynamic networks. Journal of the American Statistical Association. 110, 1646-–1657 (2015).

Statistics for Consumer Research

A panel data analysis of Italian hotels

Un'analisi di dati panel di hotel italiani

Antonio Giusti, Laura Grassini, Alessandro Viviani[1]

Abstract The present paper aims at presenting a study on the performance of the hotel industry in Italy, by analysing a panel of firms (ATECO 55.1) in the years 2008-2015. After an analysis of price and operating cost changes at the aggregate level (paragraph 2), a quantile regression is applied on microdata for modelling total production (sales). The model tries to distinguish the contribution of the quantity and quality of labor.

Abstract *Il lavoro propone un'analisi della performance dell'industria alberghiera in Italia, analizzando un panel di hotel (ATECO 55.1) negli anni 2008-2015. Dopo un confronto fra costi e prezzi a livello aggregato (paragrafo 2), il lavoro procede con l'applicazione della regressione quantile sui microdati, per modellare l'andamento della produzione in funzione dei fattori produttivi. Particolare attenzione è rivolta al fattore lavoro*.

Key words: hotel industry, operating costs, production, panel data.

1. Introduction

Hotels have a variety of internet distribution channels in selling rooms but the cost of using those intermediaries is considerable. Online Travel Agencies (OTA) and opinion aggregator websites have deeply changed the structure of the hospitality industry with consequences in the mechanisms of economic value creation (Toh, Raven, DeKay, (2011); Santoro, (2015)). It follows that an interesting issue can be the analysis of hotel operating costs and productivity. In this respect, even though the most used metrics in hotel industry refer to the number of rooms (room occupancy

[1] Antonio Giusti, Dipartimento di Statistica, informatica, applicazioni, Università degli Studi di di Firenze, email: giusti@disia.unifi.it.

Laura Grassini, Dipartimento di Statistica, informatica, applicazioni, Università degli Studi di di Firenze, email: grassini@disia.unifi.it.

Alessandro Viviani, Dipartimento di Statistica, informatica, applicazioni, Università degli Studi di Firenze, email: viviani@disia.unifi.it.

rate, revenue per available room, etc.), the presence of an intensive rivalry among firms generates high pressure to increase efficiency and productivity, so that traditional accounting measures are used, as well (Sainaghi, (2011); Sainaghi, Phillips, Zavarrone, (2017)).

As far as productivity and efficiency are concerned, current research at micro level encounters difficulties in selecting input and output, in their measurement and modelling and, namely, many empirical analyses are carried out on primary survey data of limited size (Sainaghi, Phillips, Zavarrone, (2017)). However, there are also applications of traditional growth accounting for hotel and restaurant industries (Smeral, (2009)).

The present work develops an analysis of accounting data of a panel of Italian hotels in the years 2008-2015. Two main issues are addressed: 1) whether the cost changes of inputs are recovered by output price changes; 2) an evaluation of the contribution of production factors, by the estimation of a function explaining total production (sales) through the regression quantile approach. A special attention is given to labor, as we have tried to distinguish between labor quantity and quality. In fact, human resources are a key factor in the service sector and in the accommodation industry in particular. Moreover, the increasing number of services offered by hotels (suites, dining and banquet facilities, etc.) often requires skills that are far away from the hotel core competencies (Hemmington, and King, 2000; Gonzalez-Rivera, 2005). In many cases, outsourcing is an effective way to overcome those problems.

Data are derived from Aida database, and refer to more than 3000 Italian hotels. Unfortunately, Aida database does not provide information about hotel category (number of stars) or number of rooms.

The paper is structured as follows. Section 2 presents the analysis of process and costs at the aggregate level. Section 3 describes the results of the model estimation.

2. Operating costs, operating margin and productivity

The operating-revenue to operating-cost ratio (*RVC*) and global productivity (*GP*) at time *t* are defined respectively as (Bosch-Badia, (2010)):

$$RVC_t = \frac{RV_t}{OC_t} = \frac{Operating\ revenue\ at\ current\ prices}{Operating\ costs\ at\ current\ prices}$$

$$GP_{t,0} = \frac{Operating\ revenue\ at\ constant\ prices}{Operating\ costs\ at\ constant\ prices}$$

Let be $opc_{t,0}$ the synthetic price index of outputs and $ipc_{t,0}$ as the synthetic price index of inputs, at base 0. The ratio between these two price indexes is:

$$pch_{t,0} = \frac{opc_{t,0}}{ipc_{t,0}} \quad \text{so that} \quad RV_t = GP_t\ pch_{t,0}$$

Finally, considering m_t as the operating margin:

$$m_t = \frac{RV_t - OC_t}{RV_t} = 1 - \frac{1}{GP_{t,0}\ pch_{t,0}}$$

and $T_t = RV_t/A_t$ as the asset turnover (A_t is total assets), we derive ROA (Return

On Assets) as:

$$ROA_t = T_t\left(1 - \frac{1}{GP_{t,0}\ \ pch_{t,0}}\right) = T_t\ k_t$$

where k_t represents the conversion coefficient of turnover into ROA.

Operating revenue is total sales; operating costs are: intermediate costs (materials, services etc.), labour costs and other costs (including capital depreciation). All current figures are also expressed at 2008 constant prices. Details of the deflation operations are given. The price index of output (*opc*) is computed by the ratio between current and constant price values of total production for the sector 55 (ATECO 2007). The labour cost index is provided by National Statistical Institute (ISTAT) that also releases price indices for capital depreciation. The implicit price index of intermediate costs is derived from column totals of the use-matrix (ATECO 55), by comparing values at constant and current purchase prices (*use table method*). However, as use-supply tables are available until 2013, we have provided an alternative price index (*method 2*), by applying the weights from the use-table (column values at current prices) to the price indexes of sectorial total (instead of intermediate) production (for 2014 and 2015, weights from the 2013 use-table are employed). As can be argued from Table 1 and Figure 1, the time pattern 2008-2013 of the two alternative index numbers is nearly similar but not their level. As the price index derived from the use-table is more consistent with our analysis, we have estimated 2014 and 2015 values assuming the same growth rate of the price indexes from *method 2* (Figure 1).

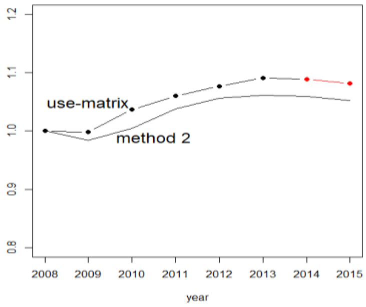


Figure 1 Price indexes for intermediate costs (base year 2008; estimates: red)

Operating costs at 2008 prices are computed by adding up the three cost components expressed at 2008 prices. Finally, the implicit price index of operating costs (*ipc*) is derived by the ratio between operating costs at current prices and operating costs at 2008 prices. Table 1 shows all price indexes and *pch*, which is the ratio between *opc* and *ipc.* The values of *pch*, after a weak increase in 2009, are systematically lower than one, showing that the change of input costs is not compensated by an adequate change in the price of services, although a weak recovery seems to occur in 2015. Figure 2 shows the time pattern of the actual conversion coefficient k_t and the value corresponding to $pch_{t,0}$=1. In 2015, the gap is not recovered yet, despite a positive trend of hotel arrivals (+15%) and nights spent (+4.5%) between 2008 and 2015.

Table 1 Index numbers of costs and prices (base year 2008)

Year	Intermediate costs (*use table*)	Intermediate costs (*method 2*)	Labour costs	Capital depreciation	Operating costs (*ipc*)	Production (*opc*)	*opc/ipc* (*pch*)
2008	1	1	1	1	1	1	1
2009	0.998	0.984	1.025	1.016	1.009	1.014	1.005
2010	1.037	1.004	1.057	1.034	1.043	1.023	0.981
2011	1.060	1.038	1.083	1.045	1.066	1.045	0.981
2012	1.077	1.056	1.103	1.074	1.085	1.061	0.978
2013	1.091	1.061	1.140	1.073	1.104	1.076	0.975
2014	*1.089*	1.059	1.155	1.080	1.109	1.082	*0.976*
2015	*1.082*	1.052	1.149	1.087	1.103	1.092	*0.990*

Source: our elaboration of ISTAT data. *Italics: estimated values*

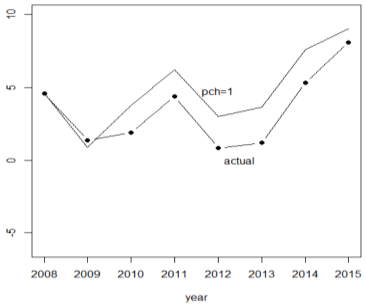


Figure 2 Conversion coefficient of turnover (×100)

3. Results of the quantile regression

The analysis at the micro level, proceeds with the estimation of a function explaining total production, where: total sales *Y* is output and the stock of fixed assets (*K*), labor costs (*L*) and intermediate costs (*C*) are inputs. Data are expressed in thousand Euros at constant 2008 prices. Also the "annual number of nights spent in hotels and similar accommodation facilities" (*N*) is included as it reflects the trend-cycle of the sector in Italy. Furthermore, we have decomposed labor costs into two terms: (1) number of workers, as a measure of labor quantity (*W*); (2) average cost per worker, as a proxy of labor quality (*Cw*), where $L=W\times Cw$. The model (Model 1) with fixed effects is:

$$lnY_{it} = \beta_0 + \sum_{j=1}^{n-1} I_j\, \alpha_j + \beta_1 lnK_{it} + \beta_2 lnW_{it} + \beta_3 lnCw_{it} + \beta_4 lnC_{it} + \beta_5 lnN_t + u_{it}$$

where *i* is the single unit (*i*=1,…, 3058), *t*=2008,…,2015 is time, α_j is the individual fixed effect, I_i is a 0-1 dummy variable assuming 1 for $j=i$, and u_{it} is the error component.

The model is estimated through the quantile regression (QR). Each QR parameter β expresses the change in a specific quantile of the response variable produced by one unit change in the regressor, with the other model covariates taken constant. With QR, we can observe how some quantiles of $logY$ may be more affected by certain predictors, than other quantiles. The presence of a large number of individual fixed effects can significantly inflate the variability of parameter estimators. Regularization and shrinkage of such effects are applied, by using the R *rqpd* library, with standard errors estimated by bootstrapping methods (Koenker, 2004).

Figure 3 illustrates the parameter estimates for the 9 quantiles, and related 95% confidence intervals, while Table 2 shows the estimated values of parameter for some quantiles. Figure 3 and Table 2 are quite revealing in several ways. Almost all parameter values are highly significant (excluding the intercept for the 0.1 quantile) and vary across quantiles. Only the *lnC* coefficient (β_4, not plotted) shows a stable pattern around 0.65.

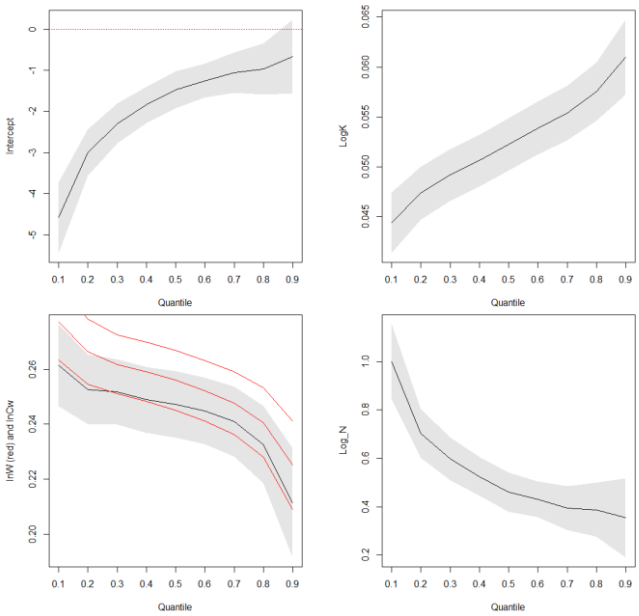


Figure 3 Estimated parameters and 95% confidence intervals

The effect of the variable expressing cycle-trend (*lnN*) is significant with a decreasing pattern over quantiles. It means that aggregate tourist demand affects to a greater extent smaller hotels (lower quantiles of sales), that reveal to be more sensitive to market dynamics.

As the size of the firm grows, the role of capital is increasingly greater whilst the opposite occurs for labor, because both related parameters (β_2 and β_3) exhibit the same decreasing pattern across quantiles. The alternative model (Model 2), with the sole variable *lnL* (in place of *lnW* and *lnCw*) produces similar results (Table 2). Larger firms result less sensitive to labor costs, probably because high skill services are outsourced, or shared within the hotel chain. Finally, if we consider the coefficients attached to the production factors, quasi-constant returns to scale emerge at each quantile.

The most important limitation of the work lies in the fact that information about the quality level of the hotel is not considered. A natural progression of this work is to provide data about hotel category and, possibly, number of rooms, even though it probably will determine a reduction of the dataset size.

Table 2 Results of the quantile regression for 0.1, 0.5, 0.9 quantiles

		Model 1			Model 2	
Covariate	Value	Std. err.	t-value	Value	Std. err.	t-value
Intercept[0.1]	-4.581	0.438	-10.464	-4.655	0.483	-9.633
lnK[0.1])	0.044	0.002	28.687	0.044	0.001	35.957
lnW[0.1]	0.277	0.007	38.969	-	-	-
lnCw[0.1]	0.261	0.008	34.524	-	-	-
lnL[0.1]	-	-	-	0.271	0.007	39.992
lnC[0.1]	0.658	0.006	103.069	0.660	0.006	103.382
lnN[0.1]	0.992	0.079	12.584	1.000	0.087	11.427
Intercept[0.5]	-1.476	0.227	-6.503	-1.461	0.239	-6.105
lnK[0.5]	0.052	0.001	39.223	0.052	0.001	48.405
lnW[0.5]	0.256	0.006	46.099	-	-	-
lnCw[0.5]	0.247	0.006	40.088	-	-	-
lnL[0.5]	-	-	-	0.252	0.005	45.992
lnC[0.5]	0.651	0.005	127.818	0.652	0.005	127.154
lnN[0.5]	0.466	0.041	11.361	0.461	0.043	10.635
Intercept[0.9]	-0.668	0.457	-1.463	-0.661	0.435	-1.520
lnK[0.9]	0.061	0.002	31.998	0.061	0.002	35.049
lnW[0.9]	0.225	0.008	27.571	-	-	-
lnCw[0.9]	0.211	0.010	20.900	-	-	-
lnL[0.9]	-	-	-	0.220	0.008	28.316
lnC[0.9]	0.654	0.007	90.088	0.655	0.007	95.963
lnN[0.9]	0.360	0.083	4.324	0.354	0.079	4.485

References

1. Bosch-Badia M. T.: Connecting productivity to return on assets through financial statements. Int. J. of Accounting and Information. Management, 18(2), 92-104 (2010)
2. Gonzalez-Rivera, G.: Outsourcing: three long run predictions. Global Business and Economics Review. 7(2/3), 226–233 (2005)
3. Hemmington, N., King, C.: Key dimensions of outsourcing hotel food and beverage services, International Journal of Contemporary Hospitality Management. 12(4), 256–261 (2000)
4. Koenker, R.: Quantile regression for longitudinal data. Journal of Multivariate Analysis. 91(1), 74-89. (2004)
5. Sainaghi, R.: RevPar determinants of individual hotels: evidence from Milan. International Journal of contemporary hospitality management. 23(3), 297-311 (2011)
6. Sainaghi, R.: Phillips, P., Zavarrone, E.: Performance measurement in tourism firms: A content analytical meta-approach, Tourism Management. 59, 36-56 (2017)
7. Santoro, G.: Evaluating performance in the hotel industry: An empirical analysis of Piedmont, Journal of Investment and Management. 4(1-1), 17-22 (2015)
8. Smeral, E.: Growth accounting for hotel and restaurant industries, Journal of travel research, 47(4), 413-424 (2009)
9. Toh, R. S., Raven, P., DeKay, F.: Selling rooms: hotels vs third-party websites. Cornell Hospitality Quarterly. 52 (2), 181–189. (2011)

A Bayesian Mixed Multinomial Logit Model for Partially Microsimulated Data on Labor Supply

Un Modello Bayesiano Logit Multinomiale Misto per Dati Parzialmente Microsimulati sull'Offerta di Lavoro

Cinzia Carota and Consuelo R. Nava

Abstract We focus on the determinants of labor choices in the presence of partially microsimulated data and discrete choice sets not identical for all agents under examination. The independence of irrelevant alternative assumption is thus discussed and the variability of the available choice set is taken into account. By comparing a Bayesian mixed multinomial logit model to a model without random effects, we show how the above described scenario affects labor choices made by single females and females within couples when the same discrete choice set is assigned to both individuals in each couple and the partner's choice is known.

Abstract *Si studiano le determinanti delle scelte di lavoro in presenza di dati parzialmente microsimulati e di insiemi discreti di scelte non identici per tutti gli agenti in esame. Viene pertanto discussa l'assunzione di indipendenza dalle alternative irrilevanti e viene tenuta in conto la variabilità dell'insieme delle scelte disponibili. Attraverso un modello bayesiano logit multinomiale a effetti misti comparato a uno privo di effetti aleatori, si mostra come questo scenario impatta sull'analisi delle scelte lavorative delle donne single oppure inserite in una coppia in cui ambedue gli individui sono esposti allo stesso insieme di scelte ed è nota la scelta del partner.*

Key words: Bayesian mixed multinomial logit model, independence of the irrelevant alternatives assumption, labour supply, random discrete choice set.

Cinzia Carota
Università degli Studi di Torino, Via Verdi 8, Torino, e-mail: cinzia.carota@unito.it

Consuelo R. Nava
Università della Valle d'Aosta, Strada Cappuccini 2, Aosta, e-mail: c.nava@univda.it

1 Introduction

Investigating the determinants of individual choices via random utility models (RUMs) [13, 15] is nowadays a common practice. RUMs describe the agent preference scheme in terms of the utility assigned to each discrete choice option, in a set of mutually exclusive alternatives (choice set). RUMs define a mapping from observed individual and/or choice characteristics into preferences with challenging theoretical and empirical statistical implications. The latter are investigated in various research fields, among which psychology and economics are by far the most important.

Recently, also policy evaluations take advantage of RUMs [3]. Instead of mere comparisons between events before and after a policy implementation, suitable RUMs are combined with microsimulation methods to anticipate, simulate and estimate the effects of socio-economic interventions. Such methods can simulate changes caused not only by hypothetical policies, but also by individual behaviours. In [4, 7] these tools are jointly used to conduct a "controlled experiment" in order to predict effects of tax and benefit reform interventions by using micro-data from national household surveys.

In this context, as a result of the microsimulation of certain fiscal variables and/or a sampling procedure applied to the available alternative options [1], it is quite common that the choice set does not exhibit the required homogeneity across decision makers (households). Even if the latter face the same number of job types (defined on the basis of a discretization of the weekly working hours in intervals, hereafter referred to as classes), microsimulation needs for each decision maker a random selection of a specific amount of working hours within each class, in order to simulate net household incomes, taxes and benefits. This implies that the i^{th} choice option refers to the i^{th} class of weekly working hours, but each household makes his decision by comparing his punctual amounts of working hours and other characteristics of jobs included in his own choice set. In this study, gross and net wage rates, given the amount of working hours, are computed via EUROMOD, a static microsimulation model [10] for tax and benefits, while the remaining variables are based on the Italian Survey on Household Income and Wealth (SHIW)[1]. The resulting partially microsimulated database contains information on households (e.g. singles or couples), while the required sampling procedure for microsimulation produces eight distinct choice sets $\{\mathscr{C}_h\}_{h=1}^{8}$, each one defined by 10+1 jobs with a specific amount of working hours[2]. Formally, household j, with $j = 1,\ldots,J$, is assigned the h^{th} choice set when $\mathscr{C}_j = \mathscr{C}_h$. In what follows, we distinguish the variables available in such database in two groups: variables directly introduced in the analysis as explanatory variables, and variables used to create groups of individuals as homogeneous as possible by means of a preliminary cluster analysis. We focus only on

[1] Also tax and benefits are, therefore, simulated according to the current fiscal system. See [7] for further details.

[2] Each choice set is composed by jobs with a different and increasing amount of working hours. For instance, the first choice set, $\mathscr{C}_1$, proposes jobs with 0, 1, 9, 17, 25, 33, 41, 49, 57, 65, 73 working hours, while the last one, $\mathscr{C}_8$, proposes with 0, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80 working hours.

labour choices made by single and non-single females, labelled female-single and female-couple.

Due to the particular structure of the available data, this article discusses the validity of the independence of irrelevant alternatives (IIA) assumption[3] associated with widely used RUMs, such as logit, conditional logit and multinomial logit models, and tries to overcome it by incorporating the decision maker heterogeneity. The main contribution is a statistical model, specifically a Bayesian mixed multinomial logit model (MMLM) [5], able to address both the violation of the IIA assumption and the just described choice set variability. In recent labour supply studies (see [6, 7] and references therein) the bias induced by the discretization of weekly working hours, and the random selection of the choice set are unaddressed problems yet[4].

2 Methods, data and results

RUMs assume a utility maximization process in order to select one of the available alternative options. For each agent j, with $j = 1,\ldots,n$, with choice set $\mathscr{C}_j = \mathscr{C}_h$, we define a random variable describing his utility U_{ij} for each alternative c_i^h in $\mathscr{C}_h = \{c_1^h, c_2^h, \ldots, c_I^h\}$. We assume, for every $j, i = 1,\ldots,I$ and $h = 1,\ldots,H$, the conditional distribution $U_{ij}|\mathscr{C}_h \sim f_{ih}(\cdot|V_{ij})$, where V_{ij} is the ground truth utility or the score assigned to each c_i^h in $\mathscr{C}_h$ [2]. In particular, $U_{ij}|\mathscr{C}_h : c_i^h \to \mathbb{R}$ and we assume $\mathbb{E}[U_{ij}] = V_{ij}$.

In our case, agents are provided with choice sets with the same cardinality, I, but made up of different alternatives across decision makers. The j^{th} agent decision process defines a permutation τ^j of $\{c_1^h, c_2^h, \ldots, c_I^h\}$ such that a linear order can be defined $[c_{\tau^j(1)} \succ c_{\tau^j(2)} \succ \ldots \succ c_{\tau^j(I)}]$. The latter manifests individual preferences to which correspond an equivalent order of the random utilities $\boldsymbol{U}_j = \{U_{1j}, U_{2j}, \ldots, U_{Ij}\}$ such that

$$\mathsf{Pr}(c_{\tau^j(1)} \succ \ldots \succ c_{\tau^j(I)} | \boldsymbol{V}_j = \{V_{1j}, \ldots, V_{Ij}\}) = \mathsf{Pr}(U_{\tau^j(1)} > \cdots > U_{\tau^j(I)}). \qquad (1)$$

Under RUMs, conditionally on h, every U_{ij} is given by the sum of V_{ij} and a stochastic (unobserved) component, ε_{ij}, i.e. $U_{ij} = V_{ij} + \varepsilon_{ij} \; \forall \; i = 1,\ldots,I;\; j = 1,\ldots,n$. Therefore, the j^{th} agent selects the alternative c_i^h in the choice set $\mathscr{C}_j$, if and only if $U_{ij} > U_{kj} \, \forall \, k = 1,\ldots,I; k \neq i$ where $\varepsilon_{ij} = U_{ij} - V_{ij}$ is a random variable (r.v.) whose mean is 0. In turn, $V_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}$ where $\mathbf{x}_{ij}$ is a $r \times 1$ vector of observed explanatory variables (for individual j and choice i), and $\boldsymbol{\beta}$ denotes a $r \times 1$ vector of fixed effects.

[3] The IIA assumption has been first used in the Luce's Axiom of Choice [11] and postulates that, when estimating the probability to select a job across a particular slate of alternatives in $\mathscr{C}$, the likelihood of choosing job a over job b will not change based on whether a third job $c \notin \mathscr{C}$ is present.

[4] Only few contributions on voters, as for instance [8], address these issues with $|\mathscr{C}_j| \neq |\mathscr{C}_i|$ for some $j \neq i$, i.e. when the cardinality of choice sets is different across decision makers.

The stochastic component ε_{ij}, instead, represents subjective noises and accommodates different sources of uncertainty like unobservable characteristics, unobservable variations in individual utilities, measurement errors and functional misspecification [12].

Assuming i.i.d. standard Gumbel or Extreme Value Type I errors, we obtain the well known multinomial logit model (MLM) [14]. To avoid the IIA irrealistic assumption and to model household heterogeneity we introduce in the expected utility V_{ij} a random component $\mathbf{z}'_{ij}\gamma_{ij}$, where z_{ij} is a $s \times 1$ design vector (assumed to be known) and γ_{ij} is a vector of s individual-specific and/or choice-specific random effects. In particular, here we exploit the individual-specific information stored in the variables not represented in $\mathbf{x}'_{ij}$ by performing a suitable hierarchical cluster analysis to assign each household j to a cluster k_j. Then, a random component $\alpha_j k_j$ is included in V_{ij}, with α_j denoting the random effect associated to cluster id k_j. Similarly, we introduce a second individual-specific random effect, δ_j, for female-couple, to make the choice made by individual j dependent on the choice made by her partner in the couple, labelled p_j. Notice that both individuals in the couple are assigned the same choice set. Finally, we take into account the above described heterogeneity of choice sets by including a random effect, η_j, of the choice set $\mathscr{C}_j$, to which decision maker j is assigned. Hence, in the more general case, the probability π_{ij} to select alternative i by agent j, for $\mathbf{z}_{ij} = \{k_j, p_j, \mathscr{C}_j\}$ and $\gamma_{ij} = \{\alpha_j, \delta_j, \eta_j\}$, can be rewritten as

$$\pi_{ij} = \mathsf{Pr}(Y_j = i|\mathscr{C}_j = \mathscr{C}_h) = \frac{\exp\{\mathbf{x}'_{ij}\boldsymbol{\beta} + \alpha_j k_j + \delta_j p_j + \eta_j \mathscr{C}_j\}}{\sum_{i=1}^{I} \exp\{\mathbf{x}'_{ij}\boldsymbol{\beta} + \alpha_j k_j + \delta_j p_j + \eta_j \mathscr{C}_j\}} \tag{2}$$

where Y_j is a random variable that takes values between 1 and I, the cardinality of the choice set $\mathscr{C}_h$, $h = 1,\ldots,8$. The probability in eq. (2) can be embedded in the following hierarchy:

$$Y_j \sim \mathsf{Multinom}(1, \boldsymbol{\pi}_j)\ \forall j = 1,\ldots,n \tag{3}$$

$$\mathsf{Logit}(\pi_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\gamma_{ij} \tag{4}$$

$$\boldsymbol{\beta} \sim \mathsf{N}(\mu_{\boldsymbol{\beta}}, V_{\boldsymbol{\beta}}), \quad \gamma_{ij} \sim \mathsf{N}(\mathbf{0}, V_\gamma), \quad V_\gamma \sim \mathsf{IW}(\Psi, \nu), \tag{5}$$

where $\boldsymbol{\pi}_j = \{\pi_{1j},\ldots,\pi_{Ij}\}$ is a vector including the "success" probabilities for each alternative in the choice set $\mathscr{C}_j$. We assign an Inverse Wishart (IW) prior distribution to the random effect (co)variance matrix V_γ while fixed effects are assigned a normal distribution. Such a hierarchical model is now applied to study whether or not the probability of choosing job i, given its features and household characteristics, is the same in the two groups of female-single and female-couple. Estimates are based on simulated (tax, benefit, gross income) and real data at a micro level, respectively from EUROMOD and SHIW[5].

[5] Here variables microsimulated by EUROMOD are based on real data provided by the Bank of Italy from the SHIW-1998 and on the 1998 Italian fiscal policy. The microsimulation model enables to calculate, in a comparable way, the effects of taxes and benefits on household incomes and work incentives.

Available data consist of 291 observed female-single and 2955 female-couple. Agents are aged between 20 and 55, neither retired nor students. They can choose among 10 different types of jobs (i.e. $i = 1, \dots, 10$) and the non labour-market participation (indexed as job 0)[6]. Type of jobs are defined on the basis of a discretization of the weekly working hours 0, 1–8, 9–16, 17–24, 25–32, 33–40, 41–48, 49–56, 57–64, 65–72, 73–80. Eight distinct choice sets $\{\mathscr{C}_h\}_{h=1}^{8}$, given the sampled weakly working hours, group singles and couples, and a new variable stores the choice set id h when $\mathscr{C}_j = \mathscr{C}_h$[7]. A hierarchical cluster analysis for each sub-population defines the cluster id for female-single and female-couple[8]. This implies that we ignore uncertainty about clusters resulting from this analysis, but we are currently working to implement more sophisticated grouping techniques, as fuzzy clustering. Variables used as predictors are the ones typically used in the literature: weekly hours of work, gross wages, age, son, choice set, taxes and benefits. Only the last two variables and gross wages were simulated with EUROMOD[9].

A Markov Chain Monte Carlo method with a block Gibbs sampling algorithm is implemented to estimate model coefficients using the `R` package `MCMCglmm` [9]. Algorithms are run for 30000 iterations, with a burn-in phase of 6000 and a thinning interval equal to 10. Hyperprior parameters were set to be: $\mu_\beta = 0$, $V_\beta = \mathbf{I}_4 \cdot 10^{10}$, with $\mathbf{I}_4$ denoting a 4×4 identity matrix. The residual covariance matrix was $\frac{1}{I \cdot 10^2} \cdot (\mathbf{I}_{I-1} + \mathbf{U})$, where $\mathbf{I}_{I-1}$ is a $(I-1) \times (I-1)$ identity matrix and $\mathbf{U}$ is a $(I-1) \times (I-1)$ unit matrix [9], given I as the number of possible choices. For the inverse-Wishart prior, Ψ was set to be equal to $\mathbf{I}_{I-1}$. Standard diagnostic tools confirmed the convergence of runs. Models did not include a global intercept, hence the first 10 estimated coefficients represented actual job type specific intercepts compared to the non-working alternative. Point estimates, under the proposed Bayesian MMLM compared to a MLM, are set out in Table 1.

The main improvements in the results under the MMLM can be appreciated both in the sign of the 10 choice coefficients, counterintuitive under the MLM, and in the large number of HPD intervals bounded away from zero.

References

1. Aaberge, R., Colombino, U., Strøm, S.: Labor Supply in Italy An Empirical Analysis of Joint Household Decisions, with Taxes and Quantity Constraints, Journal of Applied Econometrics, **14**, pp. 403–422 (1999)
2. Azari, H., Parks, D., and Xia, L., Random utility theory for social choice. In Advances in Neural Information Processing Systems pp. 126–134 (2012)

[6] In such a way, couples have 121 mixed alternatives among male and female.

[7] The number of agents from each sub-population for each choice set $h = 1, .., 8$ is 43, 43, 43, 42, 46, 52, 54, 43 for female-single and 388, 355, 356, 359, 353, 390, 382, 372 for female-couple.

[8] The hierarchical cluster analysis, based on the Ward's method and the Euclidean distance, identifies 8 groups both for female-single (with cardinality 44, 45, 69, 71, 20, 95, 19, 3) and female-couple (with cardinality 582, 557, 169, 646, 181, 367, 72, 381).

[9] Other details on data description or partial simulation can be found in [6, 7].

Table 1 Bayesian estimates for female-single and female-couple under MLM and MMLM; "***", "**", "*" and "." indicate respectively that the corresponding 99.9%, 99%, 95% and 90% HPD intervals are bounded away from zero

	Female-single		Female-couple	
	MLM	MMLM	MLM	MMLM
c_1	0,236	0,479 ***	-0,076	0,134 ***
c_2	0,153	0,427 **	-0,092	0,083 *
c_3	0,086	0,398 ***	-0,125	0,104 *
c_4	0,192	0,411 **	-0,044	0,166 ***
c_5	0,167	0,483 ***	-0,107	0,089 *
c_6	0,164	0,523 ***	-0,087	-0,001
c_7	0,134	0,482 ***	-0,106	0,034
c_8	0,292	0,383 **	-0,055	0,078 .
c_9	0,330	0,535 ***	-0,142	0,067
c_{10}	0,225	0,311 **	-0,122	0,141 **
hours	0,002	0,003 ***	0,010 ***	0,003 ***
wage	0,013 ***	0,009 ***	0,008 ***	-0,001 ***
tax	0,0001 .	-0,0001 ***	-0,00021 ***	-0,00001 *
benefit	0,0003 ***	0,0003 ***		
age	-0,011 ***	-0,020 ***	-0,007 ***	-0,002 ***
son	-0,031	0,163 ***	-0,029 ***	-0,027 ***

3. Blundell, R., MaCurdy, T.: Labour supply: a review of alternative approaches. In: Ascenfelter, O., Card, D. (eds.) Handbook of Labour Economics, pp. 1559–1695. North Holland, Amsterdam (1999)
4. Bourguignon, F., Spadaro, A.: Microsimulation as a tool for evaluating redistribution policies. Journal of Economic Inequality, **4**, pp. 77–106 (2006)
5. Cardell, N. and Dunbar, F.: Measuring the societal impacts of automobile downsizing. Transportation Research A, 14 423–434 (1980)
6. Colombino, U.: A new equilibrium simulation procedure with discrete choice models. International Journal of Microsimulation, **6**(3), pp. 25–49 (2013)
7. Colombino, U.: Five Crossroads on the Way to Basic Income. An Italian Tour. Italian Economic Journal, **1**, 353–389 (2015)
8. Gallego, M., Schofield, N., McAlister, K., Jeon, J. S.: The variable choice set logit model applied to the 2004 Canadian election. Public Choice, 158(3-4), 427-463 (2014)
9. Hadfield, J. D.: MCMC methods for Multi-Response Generalized Linear Mixed Models. The MCMCglmm R Package. Journal of Statistical Software, **33**(2), pp. 1–22 (2009)
10. Immervoll, H., O'donoghue, C., Sutherland, H.: An introduction to EUROMOD. Microsimulation Unit, Department of Applied Economics, University of Cambridge (1999)
11. Luce R. D.: On the possible psychophysical laws. Psychological Review, pp.66–81, (1959)
12. Manski, C. F.: The Structure of Random Utility Models, Theory and Decision, 8, 229–54 (1977)
13. Marschak, J.: Binary choice constraints on random utility indications. In K. Arrow, ed. Stanford Symposium on Mathematical Methods in the Social Sciences, pp. 312–329. Stanford University Press, Stanford (1960)
14. McFadden, D.: Conditional logit analysis of qualitative choice behavior. In P. Zarembka, ed., Frontiers in Econometrics, pp. 105–142. Academic Press, New York (1974)
15. Train, K. E.: Discrete Choice Methods with Simulation. Cambridge University Press (2003)

Comparison between Experience-based Food Insecurity scales

Confronto tra scale di insicurezza alimentare basate sull'esperienza

Federica Onori, Sara Viviani and Pierpaolo Brutti

Abstract In order to face food insecurity as a global phenomenon, it is essential to rely on measurement tools that guarantee comparability across countries. Although the official indicator adopted by the United Nations in the context of the Sustainable Development Goals (SDGs) and based on the Food Insecurity Experience Scale (FIES) already embeds cross-country comparability, other experience-based scales currently employ national thresholds. In this paper we address the issue of comparability by presenting two different studies. The first one between FIES and three national scales (ELCSA, EMSA and EBIA) included in national surveys in Guatemala, Ecuador, Mexico and Brazil. The second one between the adult and children versions of these national scales. Different methods from the equating practice of educational testing are explored: parametric, nonparametric, classical and based on the Item Response Theory (IRT).

Abstract *Al fine di affrontare il problema dell'insicurezza alimentare come un fenomeno globale, è essenziale poter contare su strumenti di misurazione che garantiscano comparabilità tra Paesi. Nonostante l'indicatore ufficialmente adottato dalle Nazioni Unite nel contesto dei Sustainable Development Goals e basato sulla scala FIES, formalmente assicuri questa possibilità, altre scale di insicurezza alimentare utilizzano soglie nazionali. Questo lavoro propone due studi di comparabilità. Il primo riguarda la scala FIES e le scale nazionali ELCSA, EMSA ed EBIA, mentre il secondo confronta le scale nazionali riferite a famiglie con e senza minori. Vengono implementati diversi metodi di confronto utilizzati nell'educational testing: parametrici, nonparametrici, classici e basati sull'Item Response Theory (IRT).*

Federica Onori
Sapienza University of Rome,e-mail: `federica.onori@uniroma1.it`

Sara Viviani
Food and Agriculture Organization of the United Nations, e-mail: `sara.viviani@fao.org`

Pierpaolo Brutti
Sapienza University of Rome, e-mail: `pierpaolo.brutti@uniroma1.it`

Key words: Food insecurity, Test equating, Item Response Theory, SDGs

1 Introduction

Food insecurity is formally defined as *the state of being without reliable access to a sufficient quantity of affordable and nutritious food* and food security is one of the target of the Sustainable Development Goals (SDGs) of the 2030 Agenda adopted by the United Nations. A number of indicators have been proposed to measure food insecurity and among all, *experience-based food insecurity scales* have proved to be valid and reliable tools to this aim [1]. These scales address the *access dimension* of food insecurity from the point of view of individual behaviours by directly asking people about the four aspects of psychological concern, food quality, food quantity and hunger. The first experience-based food insecurity scale was formulated in the United States where since 1995, the Household Food Security Survey Module (HFSSM) has been applied annually to monitor the phenomenon. Countries in Latin America, inspired by the HFSSM, developed their own national scales and, in this study we consider the *Brazilian Scale of Food Insecurity* (EBIA), the *Latin American and Caribbean Food Security Scale* (ELCSA) and the *Mexican Food Security Scale* (EMSA) [8], all included in national surveys for periodical monitoring. In order to provide a *global* measurement tool for food insecurity, in 2013 FAO launched the Voices of the Hungry project and introduced the Food Insecurity Experience Scale (FIES) [3]. Conceived as a global adaptation of the previous experience-based scales, the FIES was designed to produce formally comparable prevalences of food insecurity across countries and an index based on FIES was adopted as one of the official indicators for tracking progresses toward target 2.1 of the SDGs. Despite a common evolution, each national scale use specific thresholds to measure prevalences of food insecurity for *nominally* the same level of severity, and the problem of comparability arises. This paper aims at addressing this issue by proposing two comparability studies employing both classical and IRT-based methods from the educational testing field. A first study is proposed that compares FIES and the national scales ELCSA, EMSA and EBIA, while the second study computes, for each national scale, the corresponding raw scores between the adult and the children-referenced versions of the scale. In section 2 we present our data and the methods applied. In section 3 we focus on the main results and conclusions.

2 Materials and Methods

Equating is a statistical method that is used to adjust for differences in difficulty between tests' forms built to be similar in content and difficulty, so that scores can be used interchangeably [5]. The common evolution of EBIA, EMSA, ELCSA and FIES makes their Survey Modules similar to each other, allowing for statistical equating procedure. Nonetheless, the way they ”build” measures of food insecurity from observations differ under three main aspects: methodology, reference period and reporting unit. In fact, survey Modules used for ELCSA, EMSA and EBIA measure food insecurity at the *household* level and with a reference period of 3 months, while in this paper we consider for FIES a recalling period of 12 months referring to the *adult individual* (people age 15 or more). As far as for the methodology, national statistical offices using either ELCSA, EMSA or EBIA adopt a **deterministic** approach: a *raw score* is computed for each household by counting the number of items affirmed by that household. Prevalences of food insecurity at different levels of severity are then calculated as percentages of households scoring within a certain range expressed in terms of raw scores (Table 1). Conversely, FIES methodology is **probabilistic** in nature relying on the Item Response Theory (IRT) and, more specifically, on the Rasch model as the main tool for data validation and scale building [4]. Following this methodology a common metric called Global Standard is used as the reference metric to adjust model parameters estimates at each application and each respondent is assigned a distribution of his/her food insecurity along the latent trait used to compute percentages of the population whose severity is beyond global thresholds. Two indicators are then computed: the Prevalence of Food Insecurity at moderate or severe levels ($FI_{Mod+Sev}$, threshold -0.25 on the Global Standard) and the Prevalence of Food Insecurity at severe level (FI_{Sev}, threshold 1.81 on the Global Standard) [3]. Finally, FIES consists of one single scale based on 8 items referred to the adults, while EBIA, EMSA and ELCSA consist of two distinct scales, one for households without children and one for households with children, each one with specific thresholds for the different levels of food insecurity (Table 1).

Scale	Food insecurity Level	Households without children	Households with children
ELCSA	mild	1 to 3	1 to 5
	moderate	4 to 6	6 to 10
	severe	7 to 8	11 to 15
EMSA	mild	1 to 2	1 to 3
	moderate	3 to 4	4 to 7
	severe	5 to 6	8 to 12
EBIA	mild	1 to 3	1 to 5
	moderate	4 to 6	6 to 10
	severe	7 to 8	11 to 15

Table 1 National classification of food insecurity using ELCSA, EMSA and EBIA.

2.1 First Study: Equating FIES and National Scales

The aim of this comparability study is to find raw scores on the national scales EBIA, EMSA and ELCSA that are *equivalent* to the continuous FIES global thresholds used to compute the two indicators $FI_{Mod+Sev}$ and FI_{Sev} (i.e. -0.25 and 1.81). Data come from the administration of EBIA in Brazil in 2013, EMSA in Mexico in 2014 and ELCSA in Ecuador and Guatemala in 2016 and 2014 respectively and, in order to perform equating with FIES, only adult questions of the national scales have been considered. Equating between FIES and national scales was carried out by means of three equating methods for investigation purposes:

1. **IRT True Score** (IRT-TS) equating
2. **Linking** via a linear transformation applied to ability parameters
3. **Minimization** of the difference between prevalences of food insecurity

The *IRT-True Score equating* (IRT-TS) method is an IRT-based technique that consists of three steps [2]: at first, an IRT-model is fitted to the data (Estimation), then the parameters' estimates are put on a common metric through a linear transformation based on a set A of common items (Linking) and finally, equivalent expected Raw Scores are computed through the Test Characteristic Curves (TCC) of the two tests (Equating). In this study, two IRT models have been fitted to the data, namely the Rasch model and the nonparametric Mokken Scale [7] with a Kernel smoothing estimation of the Item Characteristic Curves (ICCs) [6]. When the nonparametric IRT model is considered, a Kernel smoothing estimation of the ICCs is computed and the points on the latent trait for which the corresponding estimated ICCs equals 0.5 are taken as the item severities used to estimate the linear transformation. The Standard Error of Equating (SEE) [5] for the IRT-TS method was estimated using 1000 bootstrap replications and, due to the computational costs of the procedure, is only provided together with the fitting of the Rasch model.

The second method (Linking) consists in considering the linear transformation obtained at the second step of the IRT-TS method and applying it to the estimated ability parameters of the Rasch model. Once ability parameters are adjusted to the Global standard metric, raw scores corresponding to the ability parameters that are the closest to the two global thresholds are considered as the *equivalent* raw score.

Finally, the third method (Minimizing) consists in computing prevalences of food insecurity at the household level applying the FIES methodology to the data used for the national scales and comparing the prevalences so obtained with the percentages of population scoring from a certain raw score on. The two raw scores that realize the minimum distance with the two global thresholds are considered the *corresponding* raw scores in accordance to this method.

2.2 Second Study: Comparing Adult- and Children-referenced item scales

This second analysis aims at comparing the Adult and the Children scales within each national scale. To this aim, we consider the scores obtained by the households with children on the two Module Surveys following the approach also known as the Single Group (SG) data collection design [5]. Equating of the Adult and Children scales in the four countries was carried out through implementation of five equating methods for investigation purpose: IRT True Score equating with the Rasch model, Mean, Linear, Equipercentile and Kernel Equipercentile equating methods [5].

3 Results

Outcomes from the first comparability study are summarized in Table 2, reporting the raw scores equivalent to the global thresholds used for $FI_{Mod+Sev}$ and FI_{Sev}. Results show that the global threshold used for $FI_{Mod+Sev}$ sometimes reflects a *less severe* condition of food insecurity compared to the one measured by national scales for the moderate category of food insecurity, all equated raw scores being either equal to or around one point less than the thresholds currently used by ELCSA, EMSA and EBIA (compare Table 1). On the contrary, the global threshold used for FI_{Sev} generally reflects a *more severe* condition of the food insecurity than the one captured by the national scales for the severe level of food insecurity, equated raw scores being either equal to or one point higher than the national thresholds currently in use.

FIES	Food Insecurity Scales	Internal Monitoring	IRT-TS Rasch (SEE)	IRT-TS NP	Linking	Min. Diff.
$FI_{Mod+Sev}$	ELCSA (Guatemala)	4	3.3 (0.19)	3.4	3	4
	ELCSA (Ecuador)	4	4.2 (0.14)	4.1	4	4
	EMSA (Mexico)	3	2.0 (0.23)	2.0	2	2
	EBIA (Brazil)	4	4.0 (0.09)	4.0	4	5
FI_{Sev}	ELCSA (Guatemala)	7	7.8 (0.18)	8.0	8	8
	ELCSA (Ecuador)	7	7.1 (0.18)	7.7	7	8
	EMSA (Mexico)	5	6.0 (0.26)	6.0	6	6
	EBIA (Brazil)	6	7.9 (0.07)	8.0	8	8

Table 2 Raw Scores on the national scales equivalent to the thresholds for $FI_{Mod+Sev}$ and FI_{Sev}.

Regarding the second comparability study, the equated raw scores on the Children scale shown in Table 3 seem to suggest that sometimes the current thresholds reflect a different severity of the condition measured by the two scales referred to households with and without children, respectively. This is mainly evident for the most severe category of food insecurity, for which the corresponding raw scores on the Children scale for ELCSA in Guatemala and EMSA in Mexico are generally around one point *higher* than the thresholds currently in use and between one and two points *lower* for EBIA (column ”Severe”). On the other hand, we see that the corresponding raw scores for the moderate food insecurity substantially align with the thresholds currently in use for this category (column ”Moderate”). Moreover, minor differences emerge between the behaviour of ELCSA in Guatemala and Ecuador (analysis not shown), possibly due to the specific features of the phenomenon in the two countries, confirming the importance of an equating analysis even between applications of the same scale. Finally, among all methods implemented, the Equipercentile equating method is the one whose results generally resemble the current thresholds the most.

We believe that these comparability studies can contribute in creating a more homogeneous and consistent picture of the phenomenon of food insecurity by allowing the utilization of results from application of different scales. As a consequence, this would enable a more reliable monitoring of the progress toward the goal of a global food security, as expressed in target 2.1 of the Sustainable Development Goals.

Scale	Equating	Moderate		Severe	
	Method	Raw score	(SEE)	Raw score	(SEE)
ELCSA Guatemala	IRT-TS	6.2	(0.09)	12.1	(0.1)
	Mean	6.6	(0.07)	12.2	(0.07)
	Linear	6.5	(0.07)	11.7	(0.11)
	Equip	6.3	(0.09)	11.3	(0.15)
	Kernel Equip	6.1	(0.02)	12.0	(0.04)
EMSA Mexico	IRT-TS	4.8	(0.12)	8.7	(0.13)
	Mean	5.5	(0.05)	9.5	(0.05)
	Linear	5.1	(0.07)	8.6	(0.10)
	Equip	4.8	(0.13)	8.1	(0.14)
	Kernel Equip	4.5	(0.03)	8.8	(0.04)
EBIA Brazil	IRT-TS	4.8	(0.12)	8.7	(0.13)
	Mean	5.5	(0.05)	9.5	(0.05)
	Linear	5.1	(0.07)	8.6	(0.10)
	Equip	4.8	(0.13)	8.1	(0.14)
	Kernel Equip	4.5	(0.03)	8.8	(0.04)

Table 3 Raw scores on the Children scale respectively corresponding to 4 and 7 on the Adult scale (Guatemala), 3 and 5 on the Adult scale (Mexico) and 4 and 7 on the Adult scale (Brazil).

References

1. Cafiero, C., Melgar-Quiñonez, H.R., Ballard, T.J., Kepple, A.W.: Validity and reliability of food security measures. Annals of the New York Academy of Sciences **1331**(1)
2. Cook, L.L., Eignor, D.R.: Irt equating methods. Educational measurement: Issues and practice
3. FAO.2016: Methods for estimating comparable rates of food insecurity experienced by adults throughout the world. Rome, Italy. FAO
4. Fischer, G.H., Molenaar, I.W.: Rasch models: Foundations, recent developments, and applications. Springer Science & Business Media (2012)
5. Kolen, M.J., Brennan, R.L.: Test equating, scaling, and linking: Methods and practices. Springer Science & Business Media (2014)
6. Ramsay, J.O.: Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika **56**(4), 611–630 (1991)
7. Sijtsma, K., Molenaar, I.W.: Introduction to nonparametric item response theory. Sage (2002)
8. Urquía-Fernández, N.: La seguridad alimentaria en méxico. Salud Pública de México **56** (2014)

Sovereign co-risk measures in the Euro Area
Dipendenza del rischio sovrano tra paesi dell'area Euro

Giuseppe Arbia, Riccardo Bramante, Silvia Facchinetti, Diego Zappa

Abstract We propose a method to extract significant risk interactions between Countries adopting the Graphical Lasso algorithm, used in graph theory to sort out the spurious effect of common components. In this context, the major issue is the definition of the penalization parameter. We propose a search algorithm aimed at the best separation of the variables (expressed in terms of conditional dependence) given an a priori desired partition. The case study focuses on Sovereign Bond Yields over the period 2009–2017. The proposed algorithm is used in systemic risk estimation of the Euro area sovereigns.

Abstract *L'algoritmo Glasso è noto nella letteratura associata alla teoria dei grafi per filtrare correlazioni spurie, se presenti. Il principale problema di questo metodo è la calibrazione del parametro di penalizzazione. In questo lavoro viene proposto un algoritmo volto a trovare la miglior separazione delle variabili in esame, in termini di indipendenza condizionale, assegnata una partizione ottimale definita a priori. L'applicazione esamina la relazione tra rendimenti dei bond sovrani nell'area Euro nel periodo 2009–2017 al fine di verificare l'assunto della separazione o della interdipendenza tra paesi Core e Periferici.*

Key words: Graphical Lasso algorithm, systemic risk, network dependence

Giuseppe Arbia, Department of Statistical science, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milano, giuseppe.arbia@unicatt.it

Riccardo Bramante, Department of Statistical science, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milano, riccardo.bramante@unicatt.it

Silvia Facchinetti, Department of Statistical science, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milano, silvia.facchinetti@unicatt.it

Diego Zappa, Department of Statistical science, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Milano, diego.zappa@unicatt.it

1 Introduction

We propose an original approach based on Graphical Lasso (GLasso; see Friedman et al. 2008) to investigate government bond yield data interactions from a systemic risk perspective.

The recent debt crisis in the Euro Area has turned researchers' attention to sovereign default risk measurement of. Specifically, the degree of co–movements of Sovereign bond spreads among countries can help us to understand how correlations of default probabilities – as measures of perceived country risk – evolve over time and are diffused in space. For both descriptive purposes and quality picture representation, Figure 1 depicts the corresponding bond trajectories from January 2009 to October 2017, focusing only on the 4 major developed countries in the EU (Italy, Spain, Germany and France).

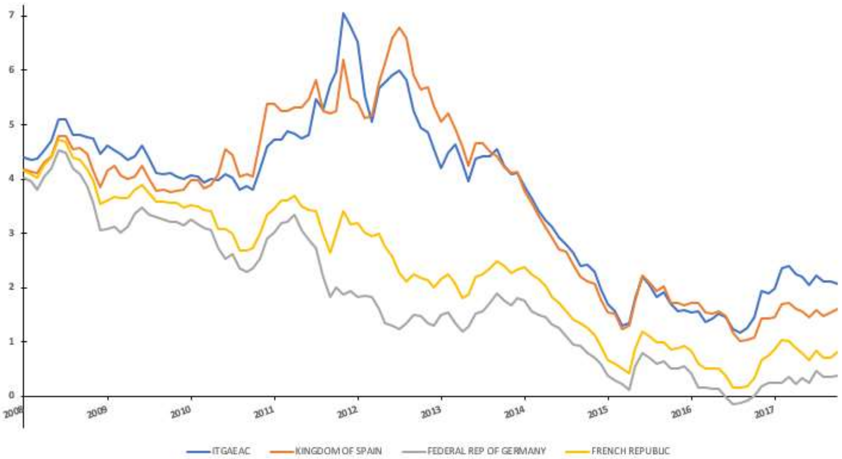


Fig. 1. Ten–year Sovereign Bond Yields

The pattern of the series in the graph unambiguously shows the effect of the crisis on Italy and Spain. Bond yields have dramatically risen from 2011 to 2013, a period affected by wide changes in global risk aversion. In particular, from 2010 yields began moving upwards, continuing to widen sharply in 2011. The sharp decline started in the second half of 2012 due to the European Central Bank policies, with a subsequent stabilization around a roughly flat trend.

In this paper, we propose the use of GLasso to focus the study only on relevant sovereign risk co–movements within the Euro Area assuming the conditional independence between Core and Peripheral Countries (ECB, 2016)[1], following the approach described in Arbia et al. (2018).

[1] For an application of GLasso in the financial field see, for example, Goto and Xu, 2015.

More specifically, the aim of this paper is twofold. First of all, our contribution focuses on the choice of the penalty parameter within the GLasso framework. As this choice is in most cases subjective and expertise-driven, we propose to modify the calibration search algorithm in Friedman et. al (2008) by searching for the minimum of the absolute difference between the Government Bond yield returns expected precision matrix and its estimate after penalization. Secondly, we analyse cross-country contagion effects by investigating – in a rolling framework – the characteristics of the degree of connectivity of the examined countries, in the graph obtained after penalization.

The rest of the paper is organized as follows. Section 2 reports some details about the GLasso algorithm and the procedure developed to calibrate the penalization parameter. In light of the Sovereign debt crisis, the application proposed in Section 3 refers to Euro Zone systemic risk analysis. Conclusions follow.

2 Methodology

Let $X=\{X_1, X_2,..., X_p\}$ be a p–multivariate random variable. Let Σ and $\Theta=\Sigma^{-1}$ be its covariance and precision matrix, respectively. It may be shown that Θ is proportional to the partial correlation matrix and so Θ can be effectively used to characterize the interrelationship of the variable of interest through the associated graph (Edwards, 2000).

The key role played by Θ justifies the many approaches proposed to estimate it efficiently and robustly with respect to abnormal deviations (see e.g. Ledoit & Wolf, 2012). A recent approach is represented by the GLasso algorithm which uses a regularization framework to estimate the covariance matrix under the assumption that its inverse is sparse. The aim of GLasso is to estimate Σ or Θ by removing the elements that likely denote spurious correlations. The way to achieve this is by introducing a penalization into the maximum likelihood estimation of the precision matrix using an L_1 penalty function over nonnegative definite matrices Θ:

$$\arg\max_{\Theta\succ 0}\{\log\det\Theta - \mathrm{tr}(S\Theta) - \lambda||\Theta||_1\} \qquad (1)$$

where $||\Theta||_1$ is the L_1 norm of Θ, S is the empirical covariance matrix and λ a scalar parameter that controls the size of the penalty. The smaller the value of λ is, the higher will be the degree of dependence between the variables and the density of the graph.

According to the work of Banerjee et al. (2008), it is possible to define the dual of sparse maximum likelihood problem in (1) as

$$\arg\max_{W}\{\log\det W : ||W - S||_\infty \le \lambda\} \qquad (2)$$

where $W=S+U$ and U is a symmetric matrix that allow to represent $||\Theta||_1$ as

$$||\Theta||_1 = \max_{||U||_\infty \le 1} \mathrm{tr}[\Theta U]$$

The dual problem (2) estimates the covariance matrix while the primal problem (1) its inverse. Moreover, log function is a monotone increasing function, thus we can also use the equivalent problem removing the log.

For further details on GLasso algorithm see Friedman et al. (2008), Banerjee et al. (2008) and Witten et al. (2011), among other.

The penalized maximum likelihood estimation for Σ can be computed $\forall\lambda \geq 0$.

Let Θ_H and $\widehat{\Theta}_\lambda$ represent the expected/desired precision matrix and the estimate of Θ after penalization, respectively. We define a flexible search algorithm to solve:

$$\min_{\lambda} \left\| \Theta_H - \widehat{\Theta}_\lambda \right\|_1 \tag{3}$$

It allows us to define a conditional dependence structure without the constraint to get a solution that exactly matches that structure. Differently from standard GLasso constraints procedures, the solution to Equation (3) allows some elements of Θ to be different from zero.

3 Empirical evidence from Euro Area Sovereign bond markets

Graphical Lasso algorithm has been applied to 10-year Government bond yields monthly data covering 17 countries in the Eurozone. This includes five Peripheral Countries, five Core Countries and seven "Other" remaining countries, which subsequently (except for Finland) adopted the Euro[1]. Data spans the period from January 2009 to October 2017; covariance and precision matrices, along with the corresponding λ parameters, are estimated over three rolling windows of size 150, 200 and 300 days.

To reflect the evolution of systemic risk and assess the degree of connection among the Core, Peripheral and "Other" countries we dynamically apply the GLasso algorithm to estimate at time *t* the precision block matrix Θ_t as follows:

$$\Theta_t = \begin{bmatrix} {}_{PP}\Theta_t & {}_{PC}\Theta_t & {}_{PO}\Theta_t \\ {}_{CP}\Theta_t & {}_{CC}\Theta_t & {}_{CO}\Theta_t \\ {}_{OP}\Theta_t & {}_{OC}\Theta_t & {}_{OO}\Theta_t \end{bmatrix}$$

where *t* is the time index and subscripts *P*, *C* and *O* indicate Peripheral, Core and remaining (Other) countries, respectively.

Our aim is to apply Equation (3) to extract graphs with a good balance between sparsity and density, avoiding the selection of large λ, which may artificially inflate sparsity in the precision matrix. Specifically, the goal is a zero *PC* block in Θ_t e.g., the Peripheral and Core Countries are conditionally independent and no co–risk between the two groups is considered. The dynamics of λ_t according to the results

[1] Luxembourg is excluded from the analysis since no data are available.

obtained by the application of Equation (3) is reported in Figure 2, by comparing the MLE and L_1 estimates.

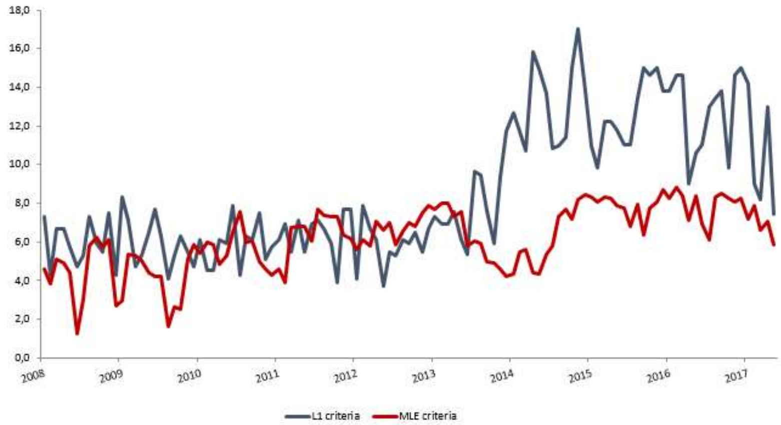
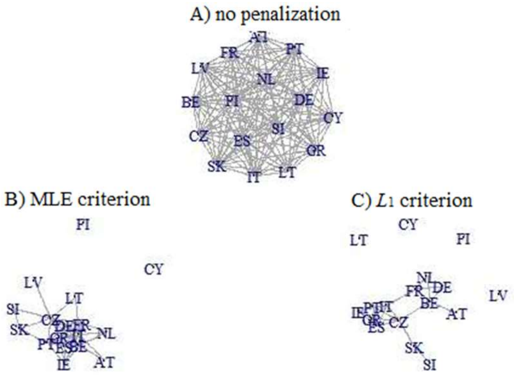


Fig. 2. λ_t comparison: *MLE* versus L_1 criterion

This series give evidence of a sharp increase in λ_t values – when the L_1 criterion is considered – during the mounting sovereign debt crisis. Moreover, while 2014 seems to represent a lower boundary of the crisis, the aftermath of a continued erratic pattern indicates that the sovereign debt crisis seems far from a final resolution.

As an example, Figure 3 shows the graphs obtained by using the best solution to Equation (3): that is, the application of MLE (Figure 3B) and L_1 criterion (Figure 3C), referring to the last rolling time window (October 2017).


A) no penalization

B) MLE criterion

C) L_1 criterion

Fig. 3. Network with A) no penalization, optimal B) MLE and C) L1 values

The graphs map out the active connections obtained in the two frameworks, giving evidence of the capabilities of the L_1 criterion to filter out the spurious effects of common components. In particular, the reinforced interconnectedness within the *PP* countries after the crisis are shown and the evidence of only an active contagion transmission channel between *PP* and *CC* groups referred to the country pair Italy-France is given. MLE solution, on the other hand, lacks of a clear pattern.

4 Conclusions

This paper exploited the Graphical Lasso algorithm to extract significant correlations in Government bond yields returns series. A calibration criterion to identify the best regularization parameter along time and cross–sectionally has been proposed.

Empirical evidences from the Euro area show that the proposed method allows to extract the relevant systemic risk contributions and identify the most interconnected nodes (countries), thus lowering the network dimension and isolating spurious relationships. We also show that the penalization parameter can be used as an indicator of the intensity of the crisis and that a better description of the relationships between Peripheral and Core Countries is obtained using the L_1 criterion instead of the *MLE* one.

References

1. Arbia, G., Bramante, R., Facchinetti, S., Zappa, D.: Modeling Inter-Country Spatial Financial Interactions with Graphical Lasso: an Application to Sovereign Co-Risk Evaluation. Regional Science and Urban Economics **70**, 72-79 (2018).
2. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. Journal of machine learning research **9**, 485-516 (2008).
3. European Central Bank: *Financial Stability Review* (2016).
4. Edwards, D.: Introduction to graphical modelling, Springer, New York (2000).
5. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**, 432-441 (2008).
6. Goto, S., Xu, Y.: Improving mean variance optimization through sparse hedging restrictions. Journal of Financial and Quantitative Analysis **50**, 1415-1441 (2015).
7. Ledoit, O., Wolf, M.: Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics **40**, 1024-1060 (2012).
8. Witten, D.M., Friedman, J.H., Simon, N.: New insights and faster computations for the Graphical Lasso. Journal of Computational and Graphical Statistics **20**, 892-900 (2011).

Simultaneous unsupervised and supervised classification modeling for clustering, model selection and dimensionality reduction

Modellizzazione simultanea di metodi di classificazione non-supervisionata e supervisionata per classificare, validare il modello e ridurre la dimensione dei dati

Mario Fordellone and Maurizio Vichi

Abstract In the unsupervised classification field, the choice of the number of clusters and the lack of assessment and interpretability of the final partition by means of inferential tools denotes an important limitation that could negatively influence the reliability of the final results. In this work, we propose to combine unsupervised classification with supervised methods in order to enhance the assessment and interpretation of the obtained partition, to identify the correct number of clusters and to select the variables that better contribute to define the groups structure in the data. An application on real data is presented in order to better clarify the utility of the proposed approach.

Abstract Nella classificazione non supervisionata, la scelta *a priori* del numero ottimale di gruppi da considerare e la mancanza di interpretazioni inferenziali, rappresenta un grosso limite per questi modelli. In questo lavoro, proponiamo la combinazione di modelli di classificazione non-supervisionata e supervisionata per identificare il numero ottimale di gruppi da considerare, selezionando le variabili che incidono in modo significativo sulla partizione trovata. E' prevista un'applicazione su dati reali.

Key words: Supervised Classification, Unsupervised Classification, Assessing Clustering, Model Selection, Dimensionality Reduction

Mario Fordellone
Sapienza, University of Rome e-mail: mario.fordellone@uniroma1.it

Maurizio Vichi
Sapienza, University of Rome e-mail: maurizio.vichi@uniroma1.it

1 Introduction

In the unsupervised classification techniques, clusters of homogeneous objects are detected by means of a set of features measured (observed) on a set of objects without knowing the membership of objects to clusters. In these applications the aim is to discover the heterogeneity structure of the data. Often, techniques based on separability and homogeneity criteria of the groups are used, giving *a priori* the number of groups [9].

Conversely, supervised classification is based on the idea to forecast the membership of new objects (output) based on a set of features (inputs) measured on a training set of objects for which the membership to clusters is known. Therefore, in these applications the aim is to generalize a function or mapping from inputs to outputs which can then be used speculatively to generate an output for previously unseen inputs [4, 6].

In this work, we propose a clustering algorithm based on the use of supervised classification modeling. In particular, the approach consists in the combination of K-Means (KM) and Logistic Regression (LR) modeling in order to find the correct number of clusters, select the most important variables and have an assessment on the partition identified through KM. An application on real data is finally proposed.

2 K-Means and Logistic Regression modeling into a clustering algorithm

In unsupervised classification modeling we are not interested in prediction, because we do not have an associated response variable y [5] like in a supervised classification model. The proposal of this paper consists in the combination of the unsupervised (i.e., K-Means (KM)) and supervised classification (i.e., Logistic Regression (LR)) approaches, where the latter, aiming to evaluate and to improve the former with adding data structure information. For simplify, we will call this approach K-Means - Logistic Regression (KM-LR). In particular, KM-LR is composed by the following principal steps:
Given the $n \times J$ data matrix $\mathbf{X}$, for $K = 2, \ldots, Kmax$, where $Kmax$ is the maximum number of clusters the researcher thinks the data might have, the algorithm works as follows:

1. let g_K be the unknown categorical membership variable which is estimated by using KM on the n-dimensional multivariate variables in $\mathbf{X}$ thus minimizing the objective function $\left\|\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\right\|^2$ [7];
2. g_K is used as response variable of the LR model with explanatory variables $\mathbf{x}$; LR is applied on g_K for estimating the probabilities for its $K-1$ response categories $\pi_k(\mathbf{x})$, and to estimate the probabilities for its *baseline* category $\pi_0(\mathbf{x})$, here fixed $k = 1$ [1];

3. if in second step some LR coefficient are not statistically significant, then we exclude the corresponding variables and we repeat from the step 1, otherwise $K = K + 1$.

Stopping rule: the algorithm continues until when the *Kmax* is reached. At the end, the optimal K is identified, together with a reduced set of statistically significant variables and a set of inferential tools to assess the quality of the partition.

In this way, through the analysis of the LR results (e.g., explained variance, parameters significance, residual variance, etc.) we have an evaluation of the partition obtained by KM. In fact, a good performance of the LR model on the response variable derived by the KM outcome, means that the variables included in the model well-explain the groups structure in the data. Moreover, through the LR coefficients analysis we can see which variables contribute most to identify the groups structure and to what extent they do it (then, analyzing statistical significance, estimates value, and sign of the coefficients).

In the next section, an application on real data is presented.

3 Application on real data

In this section a real data application of K-Means - Logistic Regression (KM-LR) is presented. The data set named *Wine Recognition Data*, is available at the UCI repository website (http://archive.ics.uci.edu/ml/). It is the result of the chemical analysis of wines grown in an Italian region, derived from three different cultivars. The 13 constituents were measured on 178 types of wine from the three cultivars: 59, 71 and 48 instances are in class one, two and three, respectively.

In the analysis we have tried to select the optimal number of clusters without considering the *a priori* information that $K = 3$, and using the KM-LR algorithm, i.e. through the maximization of *chi-squared* test computed on the partitions obtained by K-Means (KM) and Logistic Regression (LR). For comparison purpose, other two approaches have been used. The procedure has been random repeated 50 times from 2 to 10 clusters. In Table 1 have been reported the results obtained by *chi-squared* (first column), *Gap-method* proposed by Tibshirani [10] (second column), and Calinski and Harabasz [3] criterion (third column).

Thence, the best performance has been obtained by KM-LR approach, where the optimal number of clusters has been captured 36 times on 50 (72%). Whereas, KM-*Gap-method* has been obtained worst performance, since the optimal number of clusters has been captured 5 times only (10%). Then, the KM-LR approach seems to reduce the effect of the local minima problem of the KM algorithm [2], which is more relevant in the case no modification of the KM partition is proposed as in the KM-*Gap-method*, and KM-*Calinski-Harabasz*.

In Table 2 the estimation results of LR applied on the groups labels identified through KM model as response variable and including only variables with significant coefficient as predictors are shown.

Table 1 Optimal K selection from 2 to 10 clusters on the 50 random starts

	Chi-squared		Gap-method		Calinski-Harabasz	
K	Count	Percent	Count	Percent	Count	Percent
2	0	0.00	0	0.00	0	0.00
3	36	72.00	5	10.00	22	44.00
4	10	20.00	0	0.00	5	10.00
5	2	4.00	0	0.00	3	6.00
6	2	4.00	0	0.00	3	6.00
7	0	0.00	2	4.00	0	0.00
8	0	0.00	1	2.00	0	0.00
9	0	0.00	15	30.00	6	12.00
10	0	0.00	27	54.00	11	22.00
Total	50	100.00	50	100.00	50	100.00

Table 2 Estimation results obtained by Logistic Regression applied on the K-Means partition including only predictors with significant coefficient

	Estimate	SE	t-Stat	p-Value
Const.	2.0169	0.0296	68.2200	2.66E-122
Alc	-0.2306	0.0465	-4.9579	1.76E-06
Mal	-0.0865	0.0382	-2.2674	2.47E-02
Ash	-0.1261	0.0438	-2.8778	4.54E-03
AAsh	0.1022	0.0444	2.3041	2.25E-02
Mg	-0.1264	0.0353	-3.5808	4.51E-04
Phe	0.0740	0.0617	1.1993	2.32E-01
Fla	-0.2012	0.0786	-2.5597	1.14E-02
NPhe	-0.0331	0.0397	-0.8326	4.06E-01
Pro	0.0885	0.0417	2.1243	3.51E-02
Col	-0.0806	0.0516	-1.5634	1.20E-01
Hue	0.0970	0.0474	2.0492	4.20E-02
ROD	-0.0832	0.0577	-1.4418	1.51E-01
Pro	-0.3627	0.0498	-7.2806	1.31E-11

178 observations, 164 error degrees of freedom
Dispersion: 0.138, AICc=160.34, BIC=185.95
R-Squared Adj.=0.8135
F-statistic: 93.70, p-value=5.19E-55

From Table 2 we can note that the model shows a good performance, with about 80% of the total variance explained. In the model the variables *Ash*, *Alcalinty of Ash*, *Total phenols*, *Nonflavanoid phenols*, *Proanthocyanins*, *OD280-OD315 of diluted wines*, have been excluded because are not statistically significant at the 1% level.

Tables 3 show (i) the confusion matrix between real data partition and KM partition (i.e., KM applied on the complete data) and (ii) the confusion matrix between real data partition and KM-LR partition.

The misclassification Rate and the Adjusted Rand Index [8] applied on the left table (i.e., real partition versus KM partition) are equal to 0.3708 and 0.2977, respectively; whereas, the same indices applied on the right table (i.e., real partition versus KM-LR) are equal to 0.1818 and 0.5465, respectively.

Table 3 Confusion matrix between: (*i*) real data partition and *K*-Means partition; (*ii*) real data partition and *K*-Means - Logistic Regression partition

	K-Means					*K*-Means - LR			
Real	C_1	C_2	C_3	Total	Real	C_1	C_2	C_3	Total
C_1	32	5	22	59	C_1	51	3	5	59
C_2	9	61	1	71	C_2	3	66	2	71
C_3	2	27	19	48	C_3	0	12	36	48
Total	43	93	42	178	Total	54	81	43	178

Moreover, applying LR on the real data partition we obtain the following confusion matrix between the real partition and that one fitted by LR (Table 4).

Table 4 Confusion matrix between real data partition and Logistic Regression partition

	Logistic Regression			
Real	C_1	C_2	C_3	Total
C_1	15	44	0	59
C_2	6	62	3	71
C_3	2	38	8	48
Total	23	144	11	178

Also in this case the performance of KM-LR is better. In fact, the misclassification Rate and the Adjusted Rand Index applied on Table 4 are equal to 0.5225 and 0.0247, respectively. In Table 5 the performances obtained both LR applied on real partition and KM-LR are shown.

Table 5 Comparison between LR and KM-LR

	Logistic Regression	*K*-Means - LR
F-Statistic	14.5000	93.7000
p-value	0.0002	5.19E-55
R-Squared Adj.	0.0710	0.8135
AICc	403.3673	160.3400
BIC	409.6623	185.9500

We can note that the diagnostics indices obtained by KM-LR are very better with respect to those obtained by the LR application on the real data partition. Furthermore, note that in the application of LR on the real data partition, only the variable *Color intensity* has obtained a statistically significant coefficient and then, only this variable has been included in the model.
In Figure 1 the distributions of the three KM-LR clusters on the reduced set of variables are shown.

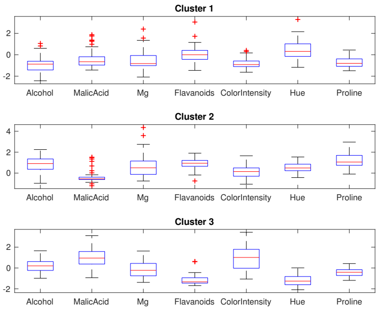


Fig. 1 Boxplots of the three KM-LR clusters distributions represented on the variables included in the model

4 Concluding remarks

In the unsupervised classification approaches, the choice of the number of clusters and the lack of assessment of the final partition are crucial issues that could negatively affect the reliability of the results. In this work we propose an algorithm that combines K-Means (KM) and the Logistic Regression (LR) modeling in order to have an evaluation of the partition identified through KM, assess the correct number of clusters (clustering) and verify the selection of the most important variables (model selection), removing in the model the non-significant variables (dimensionality reduction). In this way, we have a parsimonious set of variables that defines the best partition of data. Thus, the methodology seems promising, however, in a following work, we wish to better discover and assess, by an extensive simulation study, the performances of the proposed methodology.

References

1. Agresti, A., Kateri, M. Categorical data analysis. In International encyclopedia of statistical science, pp. 206-208 (2011)
2. Aloise, D., Deshpande, A., Hansen, P., Popat, P. NP-hardness of Euclidean sum-of-squares clustering. Machine learning, 75(2), pp. 245-248 (2009)
3. Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), pp. 1-27.
4. Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7), pp. 1895-1923.
5. Filipovych, R., Resnick, S. M., & Davatzikos, C. (2011). Semi-supervised cluster analysis of imaging data. NeuroImage, 54(3), pp. 2185-2197.
6. Hepner, G., Logan, T., Ritter, N., & Bryant, N. (1990). Artificial neural network classification using a minimal training set- Comparison to conventional supervised classification. Photogrammetric Engineering and Remote Sensing, 56(4), pp. 469-473.
7. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1(14) pp. 281-297.

8. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), pp. 846-850.
9. Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining, 400(1), pp. 525-526.
10. Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), pp. 411-423.

CONSUMERS' PREFERENCES FOR COFFEE CONSUMPTION: A CHOICE EXPERIMENT INCLUDING ORGANOLEPTIC CHARACTERISTICS AND CHEMICAL ANALYSES

Preferenze del consumatore e consumo di caffè: un esperimento di scelta comprendente caratteristiche organolettiche e analisi chimiche

Rossella Berni, Nedka D. Nikiforova and Patrizia Pinelli

Abstract In this work we propose an innovative approach for the analysis of consumers' preferences for coffee consumption by integrating a choice experiment with a consumer sensory test and chemical analyses (caffeine and antioxidants by HPLC). The same choice experiment has been administered in two consecutive time occasions, e.g. before and after the sensory test (including a descriptive illustration by an expert), in order to analyze the role of tasting in guiding the consumers' preferences. All these elements, e.g. the attributes involved in the choice experiment, the scores obtained for each coffee through the sensory tests and the HPLC analyses, are analyzed through Random Utility Models.
Abstract *Nel presente lavoro si propone un approccio innovativo per l'analisi delle preferenze del consumatore relativo al consumo del caffè. Tale approccio permette di integrare un esperimento di scelta con i risultati derivanti dal test d'assaggio e analisi chimiche (caffeina e composti antiossidanti valutati tramite il metodo HPLC). Lo stesso esperimento di scelta viene somministrato sia prima che dopo il test d'assaggio con l'obiettivo di capire se l'assaggio, unito alla descrizione operata da un esperto, può costituire una "guida" nella definizione delle preferenze del consumatore. Gli attributi dell'esperimento di scelta, i punteggi ottenuti dal test d'assaggio e i risultati HPLC sono stati analizzati con Modelli di Utilità Casuale.*

Key words: choice experiment, consumer sensory test, Random Utility Models-RUM, HPLC analysis

Rossella Berni
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy e-mail: berni@disia.unifi.it

Nedka D. Nikiforova, Patrizia Pinelli
Department of Statistics, Computer Science, Applications "G. Parenti",University of Florence, Italy e-mail: nikiforova@disia.unifi.it; patrizia.pinelli@unifi.it

1 Introduction

In this work, consumers' preferences for the coffee consumption are analyzed with an innovative approach, which integrates a choice experiment with consumer sensory tests and chemical analyses of two different types of coffee ground for moka (e.g. 100% Arabica, and a blending of Arabica and Robusta). More specifically, at the beginning a choice experiment based on optimal design theory is planned. In addition, a coffee tasting is planned in order to better analyze the consumers' preferences. To this end, a scoring card is developed, where tasters have to give a score for each organoleptic descriptor of coffee; this scoring card is administered jointly with the choice experiment. Moreover data relating to the coffee, e.g. quali-quantitative composition (caffeine, antioxidants), were acquired by using a High Performance Liquid Chromatography (HPLC) method and specific calibration curves, and then involved in the modeling step. More precisely, the same choice experiment is administered in two consecutive time occasions (Lombardi et al., 2017), e.g. before and after the sensory test, in order to analyze the role of tasting in the determination of consumers' preferences. All these elements, e.g. the attributes involved in the choice experiment, the scores obtained for each coffee from the sensory test and the HPLC analyses, are jointly analyzed in the modeling step to better evaluate the consumers behavior relating to the coffee consumption and to verify if the tasting session produced a modification in consumer's attitude.

2 Choice experiment and Random Utility Models

In the following Subsection we briefly describe the choice experiment and the collection of the data. Following, the theory of Random Utility Models (RUM) is briefly explained.

2.1 Choice experiment and data collection

At the beginning, a choice experiment based on optimal design theory is planned for building the choice-sets with the following aims: i) an efficient estimation of the attributes for the choice experiment, and ii) the detection of the effect of the sensory assessment's scores obtained through the sensory test. To this end, a compound design criterion (Atkinson et al., 2007) is applied in order to address the issues described above.

In order to collect the data, a background questionnaire about the respondent (age, gender, education) was administered at the beginning. Furthermore the choice experiment was supplied before and after the coffee tasting. For the tasting, a scorecard was developed, in which the consumer/taster is asked to assign a vote (ranging from 1 to 7, evaluated as discrete votes) to each organoleptic characteristic (smell, taste and tactile sensations). Moreover, the two types of coffee ground for moka (100% Arabica, and a blending of Arabica and Robusta varieties) were previously analyzed for their content in polyphenolic antioxidants (chlorogenic acid and other caffeoyl-quinic derivatives) and caffeine by a HPLC/DAD method.

2.2 Choice Modeling

Once data were collected, the preferences expressed by the consumers are analyzed through Random Utility Models. Therefore, as an initial step, the class of Random Utility Models (RUM) is defined. In general, every alternative is indicated by j, so that the choice-set is formed by J alternatives $(1,...,j,...,J)$, while i denotes the respondent $(i=1,...,I)$. The respondent is asked to give his/her preference within each choice-set, formed by two or more alternatives. In the Random Utility class of models, the individual i who chooses the alternative j has a random utility U_{ij} that may be generally expressed as in formula (1). Furthermore, it is assumed that the respondent i maximises his/her utility by choosing the alternative j, belonging to the choice-set C_i so that U_{ij} is the highest of all the utilities U_{ik}, $k=1,...,J$.

Thus, the following expression is characterized by a stochastic utility index U_{ij}, which may be expressed, for each unit i, as a linear function of the attributes for the alternative j, as:

$$\begin{aligned} U_{ij} &= V_{ij} + \varepsilon_{ij} \\ V_{ij} &= x'_{ij}\beta \end{aligned} \tag{1}$$

where V_{ij} is the deterministic part of the utility and is defined here in relation to a vector x_{ij}, containing the characteristics of respondent i and alternative j, β is the vector of unknown coefficients and ε_{ij}, $j=1,...,J$ is the random component. The random component is generally supposed to be independent and also Gumbel or type I extreme value distributed.

It must be noted that each alternative will be characterized by a vector of characteristics (attributes), while the response (dependent) variable is the binary variable related to the expressed preference for each choice-set.

In this study, we start by applying the Multinomial (conditional) Logit model, after we apply the Heteroscedastic Extreme Value-HEV model, described in the following Subsection.

2.2.1 The Heteroscedastic extreme value model

The Heteroscedastic Extreme Value (HEV) model (Bhat, 1995; Hensher, 1999) also belongs to the RUM class. The main feature of this model concerns the modified assumptions on the random component, which is supposedly distributed as a type I extreme value distribution, independently but not identically distributed. It must be noted that this different hypothesis on the random component makes it possible to treat the relaxation on the Independence of the Irrelevant Alternatives (IIA) property differently with respect to the Multinomial Logit model. This relaxation is fundamental and strengthens the improvement with respect to the basic logit model. Furthermore, in the HEV model, different scale parameters between alternatives are estimated. Moreover, the presence of large variances for the error terms influences the effects of changing the systematic utility for the generic alternative j. The main evident advantage is that the scale parameters may be defined as the weights in order to measure the uncertainty relating to the alternatives and the attributes involved. Therefore, the probability that a respondent i chooses the alternative j from a choice-set C_i is:

$$P_{ij} = \int_{\varepsilon} \prod_{k \in C_i; k \neq j} \Lambda \left(\frac{x'_{ij}\beta - x'_{ik}\beta + \varepsilon_{ij}}{\theta_k} \right) \frac{1}{\theta_j} \lambda \left(\frac{\varepsilon_{ij}}{\theta_j} \right) d\varepsilon_{ij} \tag{2}$$

with the error term distributed as follows:

$$f(\varepsilon_{ij};\theta_j) = \lambda \left(\frac{\varepsilon_{ij}}{\theta_j} \right) = \exp \left(-\frac{\varepsilon_{ij}}{\theta_j} \right) \exp \left\{ - \left[\exp \left(-\frac{\varepsilon_{ij}}{\theta_j} \right) \right] \right\} \tag{3}$$

In formula (2), θ_j is the scale parameter for the j alternative and $\lambda(.)$ is the probability density function of the Gumbel distribution, as detailed in formula (3), while $\Lambda(.)$ in formula (2) is the corresponding cumulative distribution function evaluated by considering two distinct choices for the i respondent. In fact, the term $x'_{ij}\beta$ denotes the deterministic part of utility of formula (1) related to alternative j and alternative k, respectively. Note that the integral function is defined on the domain [-∞,+∞] of the random component ε, related to the unit i and the alternative j. In this case, preferences of respondent i are evaluated by considering a scaling term (scale parameter) θ_j for the alternative j in the choice-set C_i i.e., the heteroscedasticity of the error term. In the case-study, two alternatives are included in each choice-set and therefore only one scale-parameter is estimated.

3 Results and Discussion

In what follows, we describe the results we have obtained for the HEV model by considering i) Choice1, that is the choice experiment administered before the tasting (Table 1), and ii) Choice2, that is the same choice experiment administered after the tasting (Table 2). More precisely, the following attributes of the choice experiment has been analyzed through the HEV model: price for a quantity of 250 grams at 3 levels: €4.50 considered as a reference level, €6.00 (labeled as ”Price 2”) and €7.50 (labeled as ”Price 3”); coffee type with two levels: ”-1” for blending Arabica and Robusta, and ”+1” for 100% Arabica; packaging with two levels: ”-1” for soft bag with modified atmosphere, and ”+1” for jar with modified atmosphere; Label

Indication with two levels: ”-1” for the presence of an indication of origin, and ”+1” for a certification about the product sustainability; soft and velvety taste with two levels: ”-1” for fairly present and ”1” for highly present (labeled as Soft Velvety Taste), and typical of a 100% Arabica coffee; intense and aromatic taste with two levels: ”-1” for fairly present and ”1” for highly present (labeled as Intense Aromatic Taste) typical of a blending Arabica and Robusta. It must be noted that in both estimated models we also included the constant (”Constant”) term settled with all the attributes at lower level.

Table 1 HEV model results before tasting (Choice1)

Variable	Estimate	Std.Error	t-value	p-value
Constant	0.6086	0.5546	1.10	0.2725
Price 2	-0.0251	0.0268	-0.94	0.3486
Price 3	-0.0914	0.0268	-3.41	0.0006
Coffee Type	-0.2855	0.1034	-2.76	0.0058
Packaging	0.0995	0.0718	1.39	0.1659
Soft Velvety Taste	-0.2002	0.2197	-0.91	0.3620
Intense Aromatic Taste	0.2304	0.1128	2.04	0.0411
Label Indication	0.1795	0.0693	2.59	0.0097
Caffeine	-0.5326	0.5572	-0.96	0.3392
Soft Velvety Taste*Caffeine	1.3756	0.8725	1.58	0.1149
Scale	1.3546	0.4660	2.91	0.0037

Table 2 HEV model results after tasting (Choice2)

Variable	Estimate	Std.Error	t-value	p-value
Constant	-0.1240	0.4963	-0.25	0.8027
Price 2	0.0272	0.0227	1.20	0.2308
Price 3	-0.0224	0.0201	-1.11	0.2650
Coffee Type	0.0866	0.0744	1.16	0.2448
Packaging	-0.1176	0.0649	-1.81	0.0700
Soft Velvety Taste	0.1871	0.0823	2.27	0.0230
Intense Aromatic Taste	-0.0583	0.0711	-0.82	0.4126
Label Indication	0.1570	0.0822	1.91	0.0560
Caffeine	0.4443	0.2395	1.85	0.0636
Taste Score	0.5643	0.3527	1.60	0.1097
Taste Score*Soft Velvety Taste	-0.6174	0.3882	-1.59	0.1118
Taste Score*Label Indication	-0.4264	0.3967	-1.07	0.2824
Scale	1.4265	0.5049	2.83	0.0047

When observing the results for Choice1 (1) and Choice2 (Table 2), we can note that the tasting session, together with the information provided on each type of coffee, have a relevant role in unequivocally guiding the consumers’ preferences. More precisely, in Choice1 the negative signs of the estimated coefficients related to the type of coffee and caffeine indicate a preference toward the blending 100% Arabica. This result is also confirmed when considering the estimated coefficients related to the attributes of soft velvety and intense aromatic taste. In fact, the consumers' preferences go towards a highly present level of soft and velvety taste that is typical for the blend 100% Arabica, and towards a fairly present level of the intense aromatic taste. However, when considering the interaction between caffeine and soft velvety taste, when the level of caffeine increases the presence of soft and velvety taste also increases; therefore we note that this result is not in line with the preference expressed towards the blend 100% Arabica. In fact, the blend 100% Arabica is characterized by a lower level of caffeine with respect to the blend Arabica and Robusta, and therefore the level of caffeine should decrease when the soft and velvety taste increases. Therefore, the result related to the interaction term is probably due to the fact that during the session ”Choice1” the respondents do not have enough knowledge about the two types of coffee, and consequently their preferences are not perfectly defined.

Instead, when considering the session ”Choice2” (results shown in Table 2), we can see a notable change in consumers' preferences with respect to Choice1, e.g., apart from the constant term that is negligible, a clear preference towards the blend Arabica and Robusta is outlined. Moreover, the positive signs of the estimated coefficients related to the type of coffee and caffeine indicates a preference towards the blend Arabica and Robusta. In line with this result, the respondents choose a highly present intense and aromatic taste (typical for the blend Arabica and Robusta) and fairly present soft and velvety taste. The estimated coefficient of the tasting scores (labeled ”Taste Score” in Table 2) obtained through the sensory tests, is positive and slightly significant, by indicating a preference towards the blend Arabica and Robusta. Furthermore, the interaction term between ”Taste Score” and ”Soft Velvety Taste” confirms that, after the intermediate session with the sensory test and information step, the consumers are more capable of differentiating between the two types of coffee. In fact, this interaction indicates that when the ”Taste Score” goes towards the blend 100% Arabica, then the ”Soft Velvety Taste” increases. This result confirms the role of guiding performed by the intermediate session, which helps the respondents for giving a coherent evaluation during the 2^{nd} choice session.

A change in the consumer preferences also concerns the packaging: in Choice1 the estimated coefficients related to the packaging goes towards the soft bag with modified atmosphere, even though this coefficient is not significant. In Choice2 instead the packaging coefficients becomes almost statistically significant with a clear preference towards the jar in a modified atmosphere. When considering the label indication, there is no change in the consumers' preferences between Choice1 and Choice2: in both occasions the respondents choose the indication of geographical origin with respect to the certification of product sustainability. The interaction between the taste score and the label indication, even though not significant, indicates that more the consumers' preferences go towards the blend 100%Arabica, more the importance of the certification of product sustainability decreases. This result could be in accordance with a more perceived quality of 100% Arabica coffees by the consumers, hence, less concerned in this case to sustainability issues.

In Choice1 (Table 1) we can observe that both price coefficients (”Price 2” and ”Price 3”) are negative; this means that the willingness-to-pay decreases when price increases, as expected. Nevertheless, after the tasting session (Choice2) a light increment of a willingness-to-pay is highlighted (the intermediate level of price shows a positive coefficient), even though in Choice2 both price coefficients are always not significant.

Moreover, a highly significant scale coefficient relating to the measurement of the heteroscedasticity effect is obtained for both Choice1 and Choice2 HEV models, e.g. the alternatives are not so irrelevant for the respondents, when doing their choices. In this direction, and with respect to the conditional logit model, the HEV model and the Mixed MNL logit model could be considered as competitive models for identifying and measuring the presence of an over-dispersion when modelling the respondent preferences. Nevertheless, the results for the Mixed MNL logit model are not presented in this paper because they require further investigations.

References

1. Atkinson, A.C., Donev, A.N., Tobias, R.D.: Optimal experimental designs, with SAS. Oxford University Press, Oxford (2007)
2. Bhat, C.R.: A heteroscedastic extreme value model of intercity travel mode choice. Transportation Research Part B-Methodological. **29**, 471–483 (1995)
3. Hensher, D.A: HEV choice models as a search engine for the specification of nested logit tree structures. Marketing Letters. **10**, 339–349 (1999)
4. Lombardi, G.V., Berni R., Rocchi B.: Environmental friendly food. Choice Experiment to assess consumers attitude toward climate neutral milk: the role of communication. Journal of Cleaner Production. **142**, 257–262 (2017)

Statistics for Earthquakes

How robust is the skill score of probabilistic earthquake forecasts?

Sulla robustezza della stima delle performance predittive di modelli sismologici probabilistici

Alessia Caponera and Maximilian J. Werner

Abstract Earthquake scientists continue to improve models of the spatio–temporal evolution of seismicity, including complex aftershock sequences. The Collaboratory for the Study of Earthquake Predictability (CSEP) prospectively evaluates the predictive skill of probabilistic forecasts by such models. Here, we assess the robustness of one popular skill score, the information gain per earthquake, with respect to temporal fluctuations of the seismicity rate. We conduct a numerical experiment with a widely-used temporal stochastic seismicity model, a special case of Hawkes process. Our simulations reveal that the information gain fluctuates substantially with time, because a central limit theorem does not hold in a realistic parameter regime. Our results may eventually contribute to more robust inferences.

Abstract *Gli scienziati della terra propongono modelli probabilistici sempre più sofisticati per descrivere l'evoluzione spazio–tempo dei terremoti. Il CSEP (Collaboratory for the Study of Earthquake Predictability) stima prospettivamente le performance predittive di tali modelli. Qui, viene valutata l'incertezza relativa a uno stimatore utilizzato da CSEP, l'information gain per earthquake. Viene condotto un esperimento numerico con un noto modello temporale di sismicità, caso particolare di processo di Hawkes. Le simulazioni effettuate rivelano che, per valori realistici dei parametri, l'information gain mantiene una variabilità elevata nel tempo. I risultati possono contribuire a rendere le conclusioni inferenziali più robuste.*

Key words: Earthquake forecasting, temporal Hawkes processes, information gain per earthquake

Alessia Caponera
Department of Statistical Sciences, Sapienza University, Rome, Italy
London Mathematical Laboratory, London, UK
e-mail: alessia.caponera@uniroma1.it

Maximilian J. Werner
School of Earth Sciences and Cabot Institute, University of Bristol, Bristol, UK
London Mathematical Laboratory, London, UK
e-mail: max.werner@bristol.ac.uk

1 Introduction

Over the last decade, stochastic and physics-based models of seismicity have matured to sophisticated system-specific forecast models that can reliably forecast the evolution of seismicity, including complex aftershock sequences. The international Collaboratory for the Study of Earthquake Predictability (CSEP) provides independent and prospective evaluations of model forecasts, and thereby aims to support robust inferences about the performances of models and guide model development [8]. One challenge is the lack of data: large earthquakes are rare, especially at the regional scales. In addition, seismicity fluctuates over orders of magnitude because earthquakes cluster in space on faults and in time during aftershock sequences.

Because short-term earthquake forecasts are now starting to inform decision-making of societal relevance, there is an urgent need to understand quantitatively the robustness of performance metrics. Here, we focus on a popular measure of the relative predictive skill of two models: the information gain per earthquake.

Despite the information gain's growing importance, its robustness has not been studied in detail. Using simulations from a popular model of clustered seismicity, we show that the information gain suffers from substantial fluctuations, because a central limit theorem does not hold under realistic parameters. Our ultimate goal is to make CSEP model inferences more robust by providing guidelines for the uncertainty in the information gain.

2 The Information Gain per Earthquake

The information gain per earthquake is defined as the log likelihood ratio between two models, say A and B, divided by the total number of earthquakes N, observed in a given time window, i.e.

$$\mathscr{I}_N(A,B) = \frac{\log L_A/L_B}{N} = \frac{\log L_A - \log L_B}{N}. \tag{1}$$

Popular statistical models for earthquake occurrences are based on marked point processes [1, Chap. 6], with magnitudes and locations as marks. We will refer here only to the time–magnitude analysis, discarding the spatial component.

Consider a marked point process on $\mathbb{R}^+ \times \mathbb{R}^+$, adapted to the filtration $\{\mathscr{F}_t\}_{t\geq 0}$, with conditional intensity function $\lambda(t,m|\mathscr{F}_{t-})$, given the history up to but not including time t. Suppose, in addition, that the process is such that the likelihood, given a realization $\{(t_i,m_i)\}_{i=1}^n$ over the interval $[0,T]$, for some positive finite T, is well defined (for further details, see [1, Chap. 7]). Then, the likelihood L of such a realization is expressible in the form

$$L = \left[\prod_{i=1}^n \lambda(t_i,m_i|\mathscr{F}_{t_i-})\right] \exp\left(-\int_0^T \int_{\mathbb{R}^+} \lambda(t,m|\mathscr{F}_{t-}) \mathrm{d}m\mathrm{d}t\right). \tag{2}$$

Its log likelihood ratio on $[0,T]$ relative to a compound Poisson process [1, Chap. 6] with constant intensity ν and mark distribution $\pi(m)$, that is independent of time t, is given by

$$\log\frac{L}{L_0} = \sum_{i=1}^{n} \log\frac{\lambda(t_i, m_i|\mathscr{F}_{t_i-})}{\nu\pi(m_i)} - \int_0^T \int_{\mathbb{R}^+} \lambda(t,m|\mathscr{F}_{t-})\mathrm{d}m\mathrm{d}t + \nu T. \qquad (3)$$

The basic idea behind the information gain is that a forecast with a higher joint log likelihood is "better". However, before observing an earthquake sequence, $\mathscr{I}_N(A,B)$ is a random variable and has its own uncertainty. Additionally, it depends on the duration of the time interval in which we observe the events. We use simulations from a popular model of seismicity to estimate the uncertainty of this statistic and to explore its robustness to the addition of new sequences.

3 The Epidemic Type Aftershock Sequence (ETAS) Model

The Epidemic Type Aftershok Sequence (ETAS) model was introduced by Ogata [6]. Belonging to the class of marked Hawkes processes [4, 1], the model approximates seismicity by an epidemic process: any earthquake increases the rate of future events for some period of time (Hawkes' self-exciting property), and large quakes induce more aftershocks (higher infection rate).

Formally, the ETAS model corresponds to a marked point process [1, Chap. 6] on $\mathbb{R}^+ \times \mathbb{R}^+$, adapted to the filtration $\{\mathscr{F}_t\}_{t\geq 0}$, with conditional intensity function

$$\lambda(t,m|\mathscr{F}_{t-}) = \left[\mu + \sum_{i:t_i<t} k(m_i)g(t-t_i)\right] p(m), \qquad (4)$$

where $\mu > 0$ represents the background seismicity rate; the term $k(m_i)g(t-t_i)$ is the contribution to seismicity rate by the i-th event (t_i, m_i), specifically

$$k(m) = Ae^{\alpha(m-m_0)}, \qquad m \geq m_0, \qquad (5)$$

is the mean number of direct offspring from an event sized m, m_0 being the magnitude threshold, and

$$g(t) = \frac{c^{p-1}(p-1)}{(t+c)^p}, \qquad t \geq 0, \qquad (6)$$

is the modified Omori law [7] for the occurrence times of direct offspring. Magnitudes are distributed independently according to the Gutenberg–Richter law [3]

$$p(m) = \beta\, e^{-\beta(m-m_0)}, \qquad m \geq m_0, \qquad (7)$$

which is the probability density function of a translated exponential distribution with rate parameter $\beta = b\log 10,\ b > 0$. Magnitudes are independent of past seismicity. $A,\ \alpha,\ c,\ p$ are constant positive parameters. For the sake of simplicity, we fix $m_0 = 0$.

The evolution of the information gain is tied to the evolution of the underlying process, as we will show in Section 4 (see also [1, 2]). Thus, we are interested in studying the behaviour of the process itself over time.

Stability properties of the ETAS model are simpler to derive in its branching process representation; we can interpret the mark m_i as the "type" of an individual in a multi-type Galton–Watson process with a modified time dimension. In this context, stability is closely related to the concepts of criticality and a branching ratio. The branching ratio is defined as the number of descendants for one immigrant over the size of their entire family (all descendants plus the original immigrant); that is

$$\rho = \frac{A\beta}{\beta - \alpha}. \tag{8}$$

Sufficient conditions for the existence of a stationary version [1, 2] are

$$p > 1, \qquad \beta > \alpha, \qquad \rho < 1, \tag{9}$$

which implies a *subcritical* process. When $\rho > 1$, the process is *supercritical*: there is a finite probability of an infinite number of events in a unit time interval.

Now, let $N(T)$ denote the number of events in the interval $(0,T]$[1]. If in addition $\beta > 2\alpha$, it can be shown that, for every sequence $\{T_n\}_{n\in\mathbb{N}}$ such that $T_n \to \infty$, a central limit theorem holds [5], namely

$$\sqrt{T_n}\left(\frac{N(T_n)}{T_n} - \frac{\mu}{1-\rho}\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{\mu(1+\sigma^2)}{(1-\rho)^3}\right), \qquad T_n \to \infty, \tag{10}$$

where

$$\sigma^2 = \frac{A^2\beta}{\beta - 2\alpha} - \rho^2.$$

The previous condition $\beta > 2\alpha$ is necessary and sufficient for the existence of σ^2, and hence the asymptotic variance in (10). These results suggest that

$$\mathbb{E}\left[\frac{N(T)}{T}\right] \approx \frac{\mu}{1-\rho}, \qquad \mathbb{V}\left[\frac{N(T)}{T}\right] \approx \frac{1}{T} \times \mathrm{const} \tag{11}$$

for sufficiently large T. However, the condition $\beta > 2\alpha$ does not hold for quakes.

4 Simulation Study

For each model in Table 1, we simulate ten thousand catalogs within the time window $[0, T_{\max}]$, with $T_{\max} = 10\,000$ days, and compute the information gain per earthquake over a finite grid of time $T_1 < T_2 < \cdots < T_k = T_{\max}$, based on the log likelihood

[1] $N(0) = 0$ almost surely.

ratio in (3). The compound Poisson process used in (3) provides a benchmark. We set $\nu = 5$ and $\pi(\cdot) = p(\cdot)$ as the mark distribution.

Experiment		μ	β	α	n	p	c
1	no clustering	1	2.3	0	0	–	–
2	short memory cl.	1	2.3	0	.5	5	.1
3	long memory cl.	1	2.3	$\beta/3$	.5	1.2	.1
4	no CLT	1	2.3	2.2	.5	1.2	.1

Table 1 Simulation scheme.

Remarkable results are displayed in Fig. 4. Experiment 3 shows well behaved trajectories of the number of events per unit time $N(T)/T$. The sample variance is bounded from above by $1/T$ times the asymptotic variance $\mu(1+\sigma^2)/(1-\rho)^3$ from (10). As a result, the information gain stabilizes around a single value. On the other hand, in experiment 4, for which the central limit theorem does not hold, trajectories and sample variance have a completely different behaviour. The information gain that does not converge to a stable value but continues to fluctuate. This is a result of large seismicity variations caused by the high aftershock rates of great earthquakes.

5 Discussion and Conclusions

The lack of an obvious convergence of the information gain per earthquake to a stable value is a warning flag: a gain measured at a moment in time, even if supported by a large data set, may change substantially in the future. The fluctuations result from the empirically-supported near-equality between the Gutenberg–Richter exponent β and the productivity exponent α. Under these conditions, a central limit theorem, which otherwise ensures convergence, does not hold.

A next step is to investigate the importance of a physically required maximum magnitude that truncates the Gutenberg–Richter law. This will theoretically restore the central limit theorem. However, observed magnitudes near the upper limit are extremely rare, and therefore the finite variance may not reign in the fluctuations for decades. Our ultimate goal is to provide CSEP with guidelines for inferring relative model performance on the basis of the information gain.

References

1. Daley, D.J., Vere-Jones, D.: An Introduction to the Theory of Point Processes, vol. I, 2nd edn. Springer-Verlag New York (2003)
2. Daley, D.J., Vere-Jones, D.: An Introduction to the Theory of Point Processes, vol. II, 2nd edn. Springer-Verlag New York (2008)
3. Gutenberg, B., Richter, C.F.: Frequency of earthquakes in California. Bulletin of the Seismological Society of America **34**(4), 185–188 (1944)

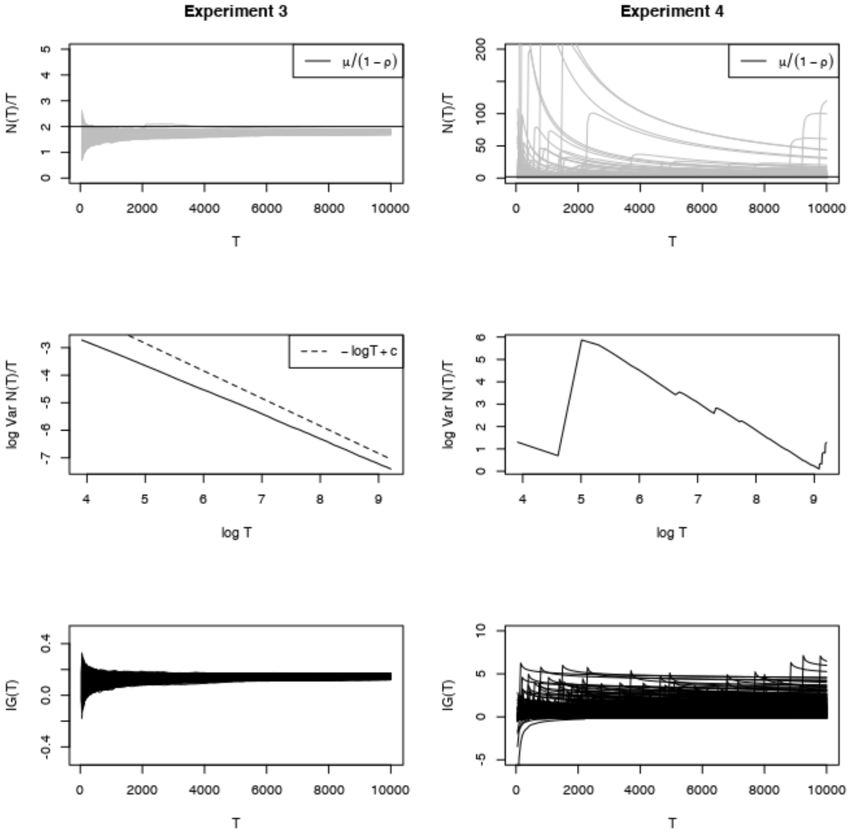


Fig. 1 From top to bottom: ETAS trajectories (number of events per unit time); sample variance (solid line) and asymptotic variance (dashed line) with time, c $= \log\left(\mu(1+\sigma^2)/(1-\rho)^3\right)$; and information gain per earthquake. Note the different y-axis scales between the left and right panels.

4. Hawkes, A.G., Oakes, D.: A cluster process representation of a self-exciting process. Journal of Applied Probability **11**(3), 493–503 (1974)
5. Karabash, D., Zhu, L.: Limit theorems for marked hawkes processes with application to a risk model. Stochastic Models **31**(3), 433–451 (2015)
6. Ogata, Y.: Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical association **83**(401), 9–27 (1988)
7. Utsu, T.: A statistical study on the occurrence of aftershocks. Geophysical Magazine **30**(4) (1961)
8. Zechar, J.D., Schorlemmer, D., Liukis, M., Yu, J., Euchner, F., Maechling, P.J., Jordan, T.H.: The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science. Concurrency and Computation: Practice and Experience **22**(12), 1836–1847 (2010)

Functional linear models for the analysis of similarity of waveforms

Modelli lineari funzionali per l'analisi della similaritá di forme d'onda

Francesca Di Salvo, Renata Rotondi and Giovanni Lanzano

Abstract In seismology methods based on waveform similarity analysis are adopted to identify sequences of events characterized by similar fault mechanism and propagation pattern. Seismic waves can be considered as spatially interdependent three dimensional curves depending on time and the waveform similarity analysis can be configured as a functional clustering approach, on the basis of which the membership is assessed by the shape of the temporal patterns. For providing qualitative extraction of the most important information from the recorded signals we propose an integration of the metadata, related to the waves, as explicative variables of a functional linear models. The temporal patterns of this effects, as well as the residual component, are investigated in order to detect a cluster structure. The implemented clustering techniques are based on functional data depth.

Abstract *In sismologia i metodi basati sull'analisi della similaritá tra onde sismiche vengono impiegati per l'identificazione di eventi caratterizzati dallo stesso meccanismo di frattura o di propagazione. Le onde sismiche possono essere considerate come curve tridimensionali, funzioni del tempo e correlate nello spazio, e l'analisi della similaritá delle forme d'onda si puó configurare come un'analisi di clustering funzionale, secondo cui l'appartenenza di un'onda ad un cluster si stabilisce in relazione al pattern temporale. Al fine di estrarre una corretta informazione dai sismogrammi, viene proposto un approccio che integra nell'analisi anche i metadati, riferiti alle onde; questi vengono considerati come esplicative di un modello lineare funzionale. Il pattern temporale degli effetti e della parte residuale, viene analizzato per l'individuazione di strutture di cluster. Le tecniche di clustering implementate sono basate su misure di 'data depth' funzionale.*

Francesca Di Salvo
Department of Agriculture, Food and Forest Sciences, University of Palermo e-mail: francesca.disalvo@unipa.it

Renata Rotondi
CNR IMATI, Milan e-mail: reni@mi.imati.cnr.it

Giovanni Lanzano
INGV, Milan e-mail: giovanni.lanzano@ingv.it

Key words: structured functional principal component, waveforms clustering, functional data depth

1 Introduction

The problem of investigating the seismotectonic structures of an area involves several methods based on waveform similarity analysis. In particular, seismic networks often record signals characterized by similar shapes and methods studying their similarity are adopted to identify sequences of foreshock, main shock and aftershock; in this field the goal is the definition of group of events characterized by similar fault mechanism and propagation pattern, under the hypothesis that a group of dependent events (multiplets) represents a chain led by seismogenetics background of a common earthquake.
The detection of earthquake families or multiplets is finalized to the identification of sources related to the same fault [6] or to obtain instrumental catalogues of independent earthquakes cleaned of dependent ones [3].

Statistical approaches are powerfull tools for detecting dependent events in a seismic data set; waveform similarity analysis is considered to join seismic episodes into a single multiplet [3]; clustering has been demonstrated as a useful method for identifying members of the same group that possess similar waveform. Different techniques for assessing the cluster membership of a earthquake are also reported in literature [8], [1].

Seismic waveform contains information of multiple attributes, that make up the set of integrating metadata concerning the source, the localization of the recording, the time of the event, the dynamics of the registration. A common opinion is that this class of data are also noisy and most techniques, including clustering, can be optimized by using appropriate data preprocessing. Signal filtering as singular value decomposition, as well as short-time Fourier transforms (STFT) are recognized as proper techniques for extracting the key features [9], [5]. In section 2, this complex functional structure is faced by models with explicit functional effect components; this step allow to integrate information from the metadata and to implement cluster analysis on the residual part of the model. Approaches relied on depth measures are considered in order to construct robust tools for the clustering of the curves. In the application, the approach is applied to a set of recordings of the seismic sequence Amatrice - Norcia - Visso, from August 2016 to January 2017, provided by the Engineering Strong Motion database (ESM).

2 The methodology

Seismic waves, that are three dimensional spatially interdependent curves, can be considered as realizations of a multivariate functional random field:

$$f^p(t) = Y^p(t) + \varepsilon^p(t)$$

The couple $(t, f^p(t))$ denotes the time and the oserved value of the function $Y^p(t)$at time t and $t \in [0,T]$. Standardizing the time interval in $[0,1]$:

$Y^p = \{Y^p : [0,1] \to R\}, p = 1,2,3$ is the set of real-valued functions on the closed interval $[0,1]$:

Each $Y^p(t)$ has a set of attributes, among which here we consider the event (metadata concerning the origin identification) and the distance from epicentre of the site where the signal is recorded; we indicate with I the number of events and with J the number of discretized distances. The curves are then indicized by a pair of indexes (i,j) in order to indicate the recordings $f^p_{ij}(t)$ of the i^{th} event at a distance d_j:

$$Y^p_{11}(t), \dots Y^p_{ij}(t) \dots Y^p_{IJ}(t) \quad | \quad Y^p_{ij} \in Y^p,\ t \in [0,1]$$

.

The presence of these effects on their dynamics is modeled in a functional two-way crossed design [10] [11]:

$$Y^p_{ij}(t) = \mu^p(t) + X_i(t) + Z_j(t) \tag{1}$$

Each curve is decomposed into an event-effect $X_i(t)$ and epicentral distance effect $Z_j(t)$, both affecting its shape. Using the Karhunen-Loéve expansion for the effects $X_i(t),\ Z_j(t)$ the model (1) becomes:

$$Y^p_{ij}(t) = \mu^p(t) + \sum_{k=1}^{\infty} \phi^X_k(t)\xi^X_{ik} + \sum_{l=1}^{\infty} \phi^Z_l(t)\xi^Z_{jl} \tag{2}$$

The proposed approach is described for $p = 1$ but its generalization for $p > 1$ can follow from the framework of functional principal components for multidimensional curves. The functional processes $X_i(t), Z_j(t)$ characterize the observed variability and their covariance operator can be estimated using method of moments estimators from the observed curves, [10]. Principal component decomposition of covariance operators results in structured principal components aims to derive the effects of the two attributes on the temporal pattern of the waves: other structures and correlation on the data are detected on the basis of the shape of the temporal patterns using a functional data depht measure. Based on the concept of data depth, this robust nonparametric tool is applied for clustering purposes in the functional data setting, providing an order within a sample of curves. Data depth notion measures the centrality of an observation within a sample and allows the definition of a natural ordering from center outwards; several depth notions generalize unidimensional concept, here we focus on Modified Band Depth [7]. Given the set of n continuous functions $f(t)$ in $[0,1]$ (the double indicization i,j is not necessary here) and λ, a Lebesgue measure in $[0,1]$, for any f of the sample the *Modified Band Depth* is the portion of time that $f(t)$ is inside the regions, made up of $2,3,\dots,k,\dots,K$ of the n curves:

$$MBD_n(f) = \sum_{k=2}^{K} n_{(k)}^{-1} \sum_{1<\ldots r_1,r_2\ldots<n} \frac{\lambda(A(f;f_{r1},f_{r2}))}{\lambda(T)} \tag{3}$$

where $A(f;f_{r1},f_{r2})$ the region delimited by f_{r1} and f_{r2}, is as follows:

$$A(f;f_{r1},f_{r2}) = \{f(t) : min_{r=r_1,r_2} f_r(t) \leq f(t) \leq max_{r=r_1,r_2} f_r(t)\} \tag{4}$$

Two previous paper [2], [4] describe the algorithm based on the notion of Modified Band Depth and adopted here for clustering the curves.

3 The application

A set of recordings of the seismic sequence Amatrice - Norcia - Visso, from August 2016 to January 2017 is considered; data are preprocessed through alignment techniques [12] dealing with different lengths of the sequences, temporal aspects, and signal filtering. The resulting sequences are aligned signals of the same length, sampled at 10 Hz; the curves are represented in figure 1 (a) .After having estimated the model (2), the estimated effects $\hat{X}_i(t)$ and $\hat{Z}_i(t)$, are represented in figure 1, respectively in (b) and in (c). The figure 2 (left) shows the functional residuals from the functional model (2), that are the input of the Modified Band depth algorithm. The algorithm finds an intrinsic order within the set of the curves, and the similarity between consecutive curves can be analyzed. In the figure 2 (right) the kernel of 50% inner curves is represented (on the top) from the 50% of the the most external curves (bottom). It is evident that the two groups differ not for shape but for the amplitude variability. As intermediate result we report in figure 3 a structure of three groups, the (25%) of the inner curves (a), the (25%) of intermediate curves (b) and the (50%) of the external curves (c): the separation is obtained on the basis of MBD and the clusters are well separated. The deepest curves of the three clusters are compared in (d).

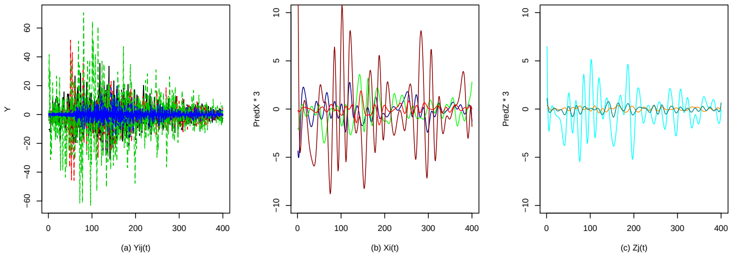


Fig. 1 Waves (a) and estimated components (b) and (c)

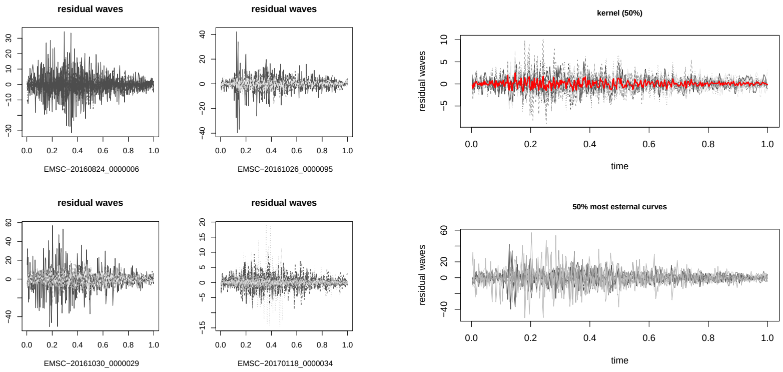


Fig. 2 left: Residuals for the main 4 events; right: the kernel (top) and the external curves (down) of the whole set of residuals from the functional linear model.

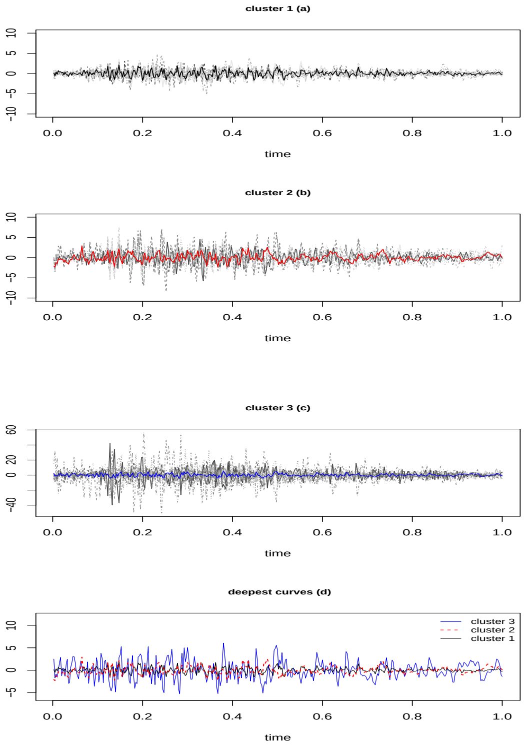


Fig. 3 Three final clusters (a) , (b), (c) obtained from ordered residuals waves. The deepest curves of the three clusters (d).

4 Results and discussion

The proposed approach is an adaptive data-driven method based on the integration of information from metadata in the analysis of the temporal pattern of the waves. Seismic waveform and attributes are used as inputs in a clustering process. Since the seismic waveform contains also some unnecessary information such as noise, they are preprocessed through alignment technique. The functional linear model derives the variability due to the events and the epicentral distance: this results improves the final identification of the clusters of curves with similar amplitude and high correlations. The interest is motivated by further analysis finalized to the study of other seismic features or geological features of the sites. The methodology can be easily extended from two-way to a m-way model [11] and to the case of three-dimensional waves, as the functional principal component analysis is well known technique for $p-$dimensional curves, $p > 1$.

References

1. Adelfio, G., Chiodi, M., D'Alessandro, A. and Luzio, D., D'Anna, G., Mangano, G. Simultaneous seismic wave clustering and registration. Computers Geosciences, 8(44), 6069 (2012)
2. Adelfio G., Di Salvo F., Sottile G. Depth-based methods for clustering of functional data TIES 2017 Conference, Bergamo, Italy, July 24th 26th, (2017).
3. Barani, S. Ferretti, G., Massa, M., Spallarossa D. : The waveform similarity approach to identify dependent events in instrumental seismic catalogues , Geophys. J. Int. doi: 10.1111/j.1365-246X.2006.03207.x (2006)
4. Di Salvo, F., Rotondi, R., Lanzano, G.: Detecting clusters in spatially correlated waveforms, GNGTS conference, Trieste, November 13th - 16th (2017)
5. Hao-kun, D., Jun-xing, C.,Ya-juan, X., Xing-jian, W., Seismic facies analysis based on self-organizing map and empirical mode decomposition, Journal of Applied Geophysics, 112, 5261 (2015)
6. Jagla, E. A., Kolton A. B.: A mechanism for spatial and temporal earthquake clustering, J. Geophys. Res. doi:10.1029/2009JB006974 (2010)
7. Lopez-Pintado, S., Romo, J., : Depth-based inference for functional data, Computational Statistics and Data Analysis 51 (10), 4957-4968, (2007).
8. Reasenberg, P. : Second-order moment of Central California seismicity, 1969 - 1982, J. geophys. Res., **90**, 54785495 (1985)
9. Silvestrov, I., Tcheverda V.:SVD analysis in application to full waveform inversion of multicomponent seismic data,Journal of Physics: Conference Series 290 doi:10.1088/1742-6596/290/1/012014 (2011)
10. Shou, H., Zipunnikov, V., Crainiceanu, C. M., Greven, S. : Structured Functional Principal Component Analysis, Biometrics, 71(1), 247257 doi.org/10.1111/biom.12236, (2015)
11. Suk, H. W., Hwang, H. : Functional Generalized Structured Component Analysis. Psychometrika, 81(4), 940968. doi.org/10.1007/s11336-016-9521-1, (2016)
12. Tucker, D.J., Wu, W. , Srivastava, A., Generative models for functional data using phase and amplitude separation, Computational Statistics and Data Analysis, 61, 5066, (2013)

Detection of damage in civil engineering structure by PCA on environmental vibration data
Valutazione del danno in opere infrastrutturali mediante PCA su dati di vibrazione ambientale

G. Agrò, V. Carlisi, R. Mantione

Abstract The dynamic behavior of civil engineering structures is usually studied by means of ambient vibration observations and their performance is analyzed by Peak Picking and/or Operational Modal Analysis methods. This paper reports the first results of a statistical multivariate approach, specifically Principal Component Analysis, to detect a suspected structural damage on a Sicilian highway bridge.
Furthermore, the damage simulated in a simple structural model made it possible to understand the characteristics of the method consisting in comparing the observed data on an undamaged structure with those coming from a damaged one.

Riassunto *Il comportamento dinamico nelle costruzioni civili viene usualmente studiato attraverso prove di caratterizzazione dinamica utilizzando il metodo del Peak Picking e/o quello dell'Analisi Modale Operazionale sul dominio delle frequenze. Il presente lavoro riporta i primi risultati di un approccio statistico multivariato, in particolare l'Analisi in Componenti Principali, al fine di determinare un possibile danneggiamento strutturale di un ponte autostradale siciliano.*
Infine la simulazione del danno in un modello strutturale semplice ha permesso l'individuazione delle variazioni di parametri significativi del metodo utilizzato confrontando i risultati del modello danneggiato con l'omologo integro.

Key words: Principal Component Analysis, Damage Detection, Subspace Angles, Operational Modal Analysis, Fast Fourier Transform, Peak Picking technique.

[1] Gianna Agrò, DSEAS, University of Palermo; email: gianna.agro@unipa.it

Valentina Carlisi, Laboratorio DISMAT; email: valecarlisi@libero.it

Roberta Mantione, Laboratorio DISMAT, email: robertamantione.rm@gmail.com

1. Introduction

One of the most used approaches for damage detection in engineering structures is environmental vibration test that allows the gathering of natural frequencies and obtaining the so-called mode shapes and structural dampings. The test is performed using a predetermined number of uniaxial piezometric accelerometers connected to a control unit for the acquisition of environmental acceleration times. By means of Fast Fourier Transform (FFT), the time series are transferred to the frequency domain and the Peak Picking and the Frequency Domain Decomposition (FDD) techniques were used to extract the dynamic parameters from the spectral densities matrices.

The Peak Picking (PP) method leads to reliable results provided that the basic assumptions of low damping and well-separeted modes are satisfied. In fact this method allows to identify the operational deflection shape that, in the case of closely modes, represent the overlap of numerous modes. The Frequency Domain Decomposition (FDD) technique, which represents a significant improvement of the PP, through the Singular Value Decomposition (SVD) of the spectral densities matrices, is able to detect closely spaced modes: the singular value will have a maximum in the resonant frequencies [3,4].

A recent approach in the context of damage identification on engineering structures is the analysis of the principal components (PCA) applied to investigate the existence of any change between a suspected damaged structure and an undamaged similar one adopted as a reference model. The results of the n experimental tests, obtained by means of p sensors, constitute the X matrix, of dimension nxp, which is the starting point for the PCA. The aim of PCA is to reduce the space of p correlated variables in such a way do not lose the bulk of information contained in the data. In synthesis from the data collected in X, the correlation matrix R is calculated such as eigenvectors and eigenvalues of R which are used to identify the subspaces among which choosing the reduced dimension k<p corresponding to a fixed amount of the system variance [1]. This procedure is adopted for the matrix X deriving from the healthy structure and for the matrix Y from the damaged structure. The selected subspaces, one for each systems, are compared by means of the maximum angle θ between the subspaces [6].

Section 2 shows the simulation study, while section 3 shows the results of the analysis on a Sicilian motorway bridge; finally, in section 4, some conclusions are drawn.

2. Numerical application

The system, consisting of two equal masses connected in series through linear springs and adopted for the simulation study (Figure 1), is named a two-degree-of-freedom system (2DOF) and is subject to a free harmonic movement with a natural frequency ω.

The system responses are the vectors u_1 and u_2 that representing respectively the displacement of mass m_1 and mass m_2..

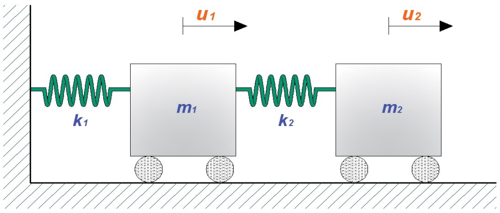


Figure 1: Two degree of freedom mass-spring model undamped and unforced vibration system

The homogeneous linear differential equations of motion, for the 2DOF system, can be written as

$$\mathbf{M}\,\ddot{\mathbf{u}}(t) + \mathbf{K}\,\mathbf{u}(t) = \mathbf{0}$$

or in extended matrix form

$$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}\begin{bmatrix} \ddot{u}_1 \\ \ddot{u}_2 \end{bmatrix} + \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where **M** is a diagonal mass matrix**,** **K** is a stiffness matrix, while **u** and **ü** are vectors of time varying displacements and accelerations respectively. The trial solution **u**=**U**cos(ωt-φ) with ω *natural frequency*, φ *phase* and **U** *time independent amplitude vector*, lead to solve the classical eigenvalue problem

$$(\mathbf{K} - \lambda\,\mathbf{M})\mathbf{U} = \mathbf{0}$$

where λ_i, i=1,2 are the *eigenvalues*. The *eigenvalues* so obtained depend only from physical parameters of the system. The replacement of $\lambda_i=\omega_i^2$ in the classical eigenvalue problem allows to found the *eigenvectors* U_i corresponding to the natural frequencies, the so-called *mode shapes* [4]. The response of the system is calculated with a sampling frequency of 0.03 sec and the damage is simulated by varying the stiffness of a single spring [5], ($k_1=k_2$=213'330 N/cm, k_2=1, 5, 10, 15, 20, 30 and 40% of k_1), leaving the masses unchanged, $m_1=m_2$=9700 N, and giving an initial displacement. Finally, the simulated response data were perturbed by adding, to each of the time series, a white Gaussian noise with S_d = 15% of the Root Mean Square value of the respective series [2,7]. Natural frequencies were obtained by using the FFT technique that allows to transform the responses from the time domain to the frequency one where we can read the abscissa value of the peaks showed in the graphic of the power spectral density (Figure 2).

Observing Figure 2, where the spectra of the different system conditions are presented (level of damage from 0% to 40%), it is very difficult to detect the presence of damage since the abscissa values of the first peak (frequency) are coincident while for the second one (frequency) the abscissa range variation is very small.

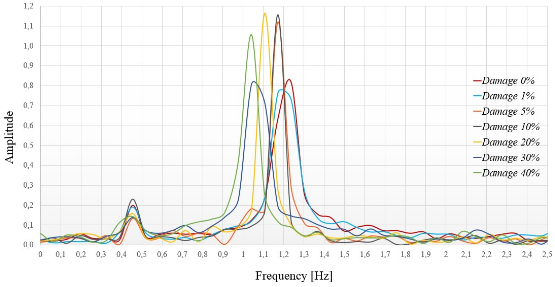


Figure 2: Comparison of frequency

The results of the Principal Component Analysis, applied to the described model, are shown in Figure 3 where the scatterplot of ($u_1(t)$,$u_2(t)$), of the undamaged system is presented and the orthogonal axes (PC1, PC2), a couple for each level of damage, are superimposed.

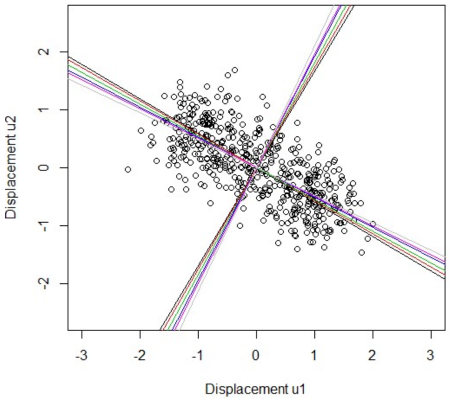


Figure 3: Scatterplot of u_1,u_2 and rotation of the PCA axes for damaged systems

The black axes are related to the undamaged system (dam=0%) and the others are for the systems damaged in increasing way. It can be noted that there is a rotation (in counter-clockwise) of the PC_i axes. The angles θ_i between the black and colored axes could be considered as a measure of the extent of the damage as we will see in the next section.

3. Damage detection: the case of a sicilian highway bridge

The case study concerns a Sicilian highway bridge [8], built in the 70s and 1533 m long; it includes 35 reinforced concrete spans. The bridge consists of isostatic spans, each of which has a deck consisting of four pre-stressing R.C. beams with double T section; each span is 45 m long and 9.8 m wide. Ambient vibration tests were conducted on two adjacent spans 8, considered undamaged, and 9, suspected damaged, using ten uniaxial piezoelectric accelerometers, Figure 4 shows the layout of the sensors. The ambient acceleration time histories were recorded for 2400 sec at interval of 0.01 sec. The present study was conducted using only six vertical sensors, in Figure 4 the red ones, and data were collected in matrices of dimension (240000x6).

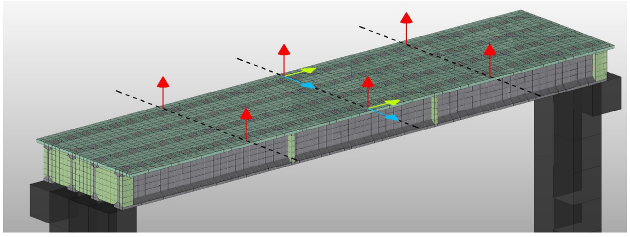


Figure 4: Location of the sensors on the bridge deck

Operational Modal Analysis, according to FDD technique, was used to identify natural frequencies and mode shapes. By comparing the frequency of the spectra of the two spans (Figure 5), the difference in abscissa values of the peaks is very low (less than 10%) and consequently, by means of the PP technique, it is difficult to detect a damage.

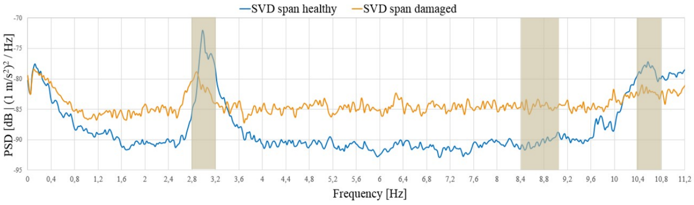


Figure 5: Singular Values of Spectral Densities of Test Setup

The application of the PCA to study the behaviour of healthy span (n.8 matrix X) and damaged span (n.9 matrix Y) led to the identification of optimal four-dimensional subspaces since the percentages of the total variance explained were 75% for healthy span and 80% for damaged span.

In order to measure the difference among the spans, the principal angle between the four-dimensional subspaces was calculated [6]:

$$\cos\theta_k = \max_{d\in D}\ \max_{h\in H} d^T h = \max D^T H \qquad (1)$$

subject to $\|h\| = \|d\| = 1 \quad h^T h = 0\ and\ d^T d = 0$

where $D_{(6x4)}$ and $H_{(6x4)}$ are the matrices of eigenvectors for the damaged and healthy span respectively.

In our case the difference between the two spans exists and it is showed by the angle θ=arcos(-0.7429298)=42°; on this result it is evident that the span 9 has a damage.

Table 3: Cos(θi) calculated by formula (1): maximum value red coloured

H vs D	***PC1***	***PC2***	***PC3***	***PC4***
PC1	**-0.7429298**	0.2386534	0.3945228	-0.3908226
PC2	-0.5139214	0.2406920	-0.2453405	0.3750808
PC3	-0.3328730	-0.1497690	-0.6219700	0.3301877
PC4	-0.1773170	-0.6197478	-0.3634516	-0.6058055

4. Conclusion

The paper presents an investigation of the damage detection capability of the Principal Component Analysis applied to the response structural vibration tests in time domain. The numerical simulation here reported is a starting point of a complete simulation study on different structural models.

On the case study, the analysis of the vibration tests by Peck Peaking technique did not produce easily interpretable evidence while the analysis of the subspaces resulting from the reduction by means of PCA gave evidence of the damage present in the structure.

References

1. Anderson T. W.: An introduction to multivariate statistical analysis, Wiley (2003)
2. Brandt A.: Noise and Vibration Analysis – Signal Analysis and Experimental Procedures. Wiley (2011)
3. Brincker R.,Ventura C.: Introduction to Operational Modal Analysis. Wiley. (2015).
4. Ewins D.J.: Modal Testing – Theory, practice and application. Research Studies Press Ltd., England (2000)
5. Friswell M.I: Damage Identification using Inverse Methods In Morassi A., Vestroni F. Editors: Dynamic Methods for Damage Detection in Structures, pp. 13-66, Springer Wien New York (2008)
6. Golub G., Van loan C.: Matrix computations. The Johns Hopkins University Press, Baltimore (1996)
7. Poncelet F., Kerschen G., Golinval JC., Experimental modal analysis using blind source separation techniques - International Conference on Noise and Vibration Engineering, Leuven (2006)
8. Technical Report n° 51216-2, DISMAT (2017)

Statistics for Financial Risks

Conditional Value-at-Risk: a comparison between quantile regression and copula functions

Conditional Value-at-Risk: un confronto tra la regressione quantile e le funzioni copula

Giovanni De Luca and Giorgia Rivieccio

Abstract The popular Value-at-Risk of an institution provides a measure of its own risk. However, in many cases it is of interest to know the measure of the contribution of an institution to the systemic risk, based on the Conditional Value-at-Risk. In this paper we compare the estimation of such measure according to two different methods. The former is based on the quantile regression, the latter on copula functions. In both cases, the heteroskedasticity of the time series is explicity taken into account through a GARCH structure. Moreover, the comparison is also made across different distributional assumptions.

Abstract *Il Value-at-Risk è un'importante misura di rischio di un'istituzione. Tuttavia, in molti casi, è di maggiore interesse la misura del contributo di un'istituzione al rischio sistemico. Tale misura è basata sul Conditional Value-at-Risk. In questo lavoro si confrontano due metodi per la stima del contributo, il primo basato sulla regressione quantile, il secondo sulle funzioni copula. In entrambi casi la struttura eteroschedastica è inclusa attraverso una struttura GARCH. Il confronto è svolto anche considerando diverse ipotesi distributive.*

Key words: systemic risk, quantile regression, copula functions.

1 Introduction

Literature on systemic risk has been recently enriched by the introduction of a new measure, the Conditional Value-at-Risk (*CoVaR*), measuring the effect of an individual financial institution to the systemic risk. *CoVaR* has been proposed in Adrian and Brunnermeier (2011). In its general formulation, it is defined as the Value-at-

Giovanni De Luca
University of Naples Parthenope, via G. Parisi, 13, Napoli, Italy e-mail: gdeluca@uniparthenope.it

GIorgia Rivieccio
University of Naples Parthenope, via G. Parisi, 13, Napoli, Italy e-mail: rivieccio@uniparthenope.it

Risk (*VaR*) of the financial system given that an institution is under distress. The importance of this measure is given by the possibility of quantifying the contribution of an institution to the financial system distress.

2 Methodology

The Value-at-Risk at level q of the financial system is defined as the q-th quantile of the distribution of the returns Y_f of an index representing the financial system, i.e.

$$Pr\left(Y_f \leq VaR_q^f\right) = q.$$

We denote by $CoVaR_q^{f|Y_i=VaR_q^i}$ the VaR of the index conditional on financial distress of institution i, that is conditional on $Y_i = VaR_q^i$, where Y_i is the return of institution i. As a result, $CoVaR_q^{f|i}$ is implicitly defined by the q-th quantile of the conditional probability distribution $Y_f|Y_i$, that is

$$P\left(Y_f \leq CoVaR_q^{f|i} \left| Y_i = VaR_q^i \right.\right) = q$$

This is the original definition that will be considered in this paper.[1]

The *CoVaR* can be intepreted as the estimate of the effect on the system of a critical situation of institution i. Moreover, it is defined the so-called $\Delta CoVaR_q^{f|i}$, given by

$$\Delta CoVaR_q^{f|i} = CoVaR_q^{f|Y_i=VaR_q^i} - CoVaR_q^{f|Y_i=Median^i}$$

where $CoVaR_q^{f|X_i=Median^i}$ is the q-th quantile of Y_f conditional on a normal situation of institution i, that is the return of institution i is equal to its median. The $\Delta CoVaR_q^{f|i}$ is the increase of $CoVaR_q^{f|i}$ when the institution i goes into distress.

The estimation of the $CoVaR_q^{f|i}$ is usually based on the linear quantile regression technique (Koenker and Bassett, 1978). In this case, the q-th quantile of Y_f, denoted as Y_q^f, is estimated as function of Y_i

$$\hat{Y}_q^f = \hat{\alpha}_q + \hat{\beta}_q Y_i$$

So, the estimate of $CoVaR_q^{f|i}$ is given by

$$\widehat{CoVaR}_q^{f|Y_i=VaR_q^i} = \hat{\alpha}_q + \hat{\beta}_q VaR_q^i$$

while the estimate of $CoVaR_q^{f|Y_i=Median}$ is given by

[1] Girardi and Ergun (2013) have modified the original definition of *CoVaR*, enlarging the conditional event (the institution i is at maximum at VaR_q^i, that is $Y_i \leq VaR_q^i$).

$$\widehat{CoVaR}_q^{f|Y_i=Median^i} = \hat{\alpha}_q + \hat{\beta}_q Median^i.$$

Finally,

$$\Delta\widehat{CoVaR}_q^{f|i} = \hat{\beta}_q(VaR_q^i - Median^i).$$

This procedure does not take into account the heteroskedasticity that typically characterizes financial returns. In this work we modify the quantile regression to take into account this feature of returns. The modified quantile regression approach is based on the following steps:
(a) estimate of a GARCH-type model for returns of both the index of the financial system and institution i and derivation of the standardized residuals η_f and η_i;
(b) estimate of the parameters α and β using a quantile regression such that the q-th quantile of η_f is given by

$$\hat{\eta}_q^f = \hat{\alpha}_q + \hat{\beta}_q \eta^i$$

(c) transformation of the estimated conditional quantiles $\hat{\eta}_q^f$ to obtain $\Delta CoVaR_q^{f|i}$.

An alternative approach is the estimate of $\Delta CoVaR_q^{f|i}$ using the Copula-GARCH approach (Jondeau and Rockinger, 2006). The steps of this procedure are:
(a) estimate of a GARCH-type model for returns of both the index of the financial system and institution i and derivation of the standardized residuals η_f and η_i;
(b) selection of a copula function describing the bivariate relationship between η_f and η_i, $C(F_{\eta_f}, F_{\eta_i})$, where F_{η_f} and F_{η_i} are the distribution functions of the variables η_f and η_i, respectively;
(c) derivation of the conditional copula, $C(F_{\eta_f}|F_{\eta_i})$;
(d) transformation of the estimated quantiles $\hat{\eta}_q^f$ from the conditional copula to obtain $\Delta CoVaR_q^{f|i}$.

Finally, the two procedures are compared through the comparison of the estimates of $\Delta CoVaR_q^{f|i}$.

3 Application to data

We have considered the daily log-returns of eigth assets: A2A, BPER, Enel, FCA, Generali, Intesa San Paolo, STM, Unicredit. Three of them belong to the banking sector (BPER, Intesa San Paolo and Unicredit). The returns have been observed in the period from February 27th, 2013 to January 30th, 2018 (in total we have 1249 observations). The financial system is represented by the FTSEMIB index.

In the quantile regression approach, the estimates have been carried out considering different distributional assumptions in the GARCH specification (in particular we have considered three conditional distribution: Normal, Student's t and Skew Student's t). The same distributional assumptions have been considered in the Copula-GARCH approach, while the copula function is the Student's t copula, characterized by lower and upper tail dependence.

Figure 1 reports the scatterplots of average $\Delta CoVaR_{0.05}^{f|i}$ against the empirical q-th quantile for the eight assets. The top row contains the scatterplots with $\Delta CoVaR_{0.05}^{f|i}$ computed using the quantile regresion approach, in the bottom the scatterplots consider the $\Delta CoVaR_{0.05}^{f|i}$ estimated based on the copula approach. The columns identify the distributional assumptions: Normal (left), Students's t (middle) and Skew Student's t (right). To facilitate the comparison, we have kept the same scale on the y-axis.

The most relevant findings are the following:

1. The quantile regression approach tend to group the assets in two cluster, regardless the distribution of the returns. Interestingly, the two assets with the lowest $\Delta CoVaR_{0.05}^{f|i}$ are two banking institutions (Intesa SanPaolo and Unicredit, that is the biggest Italian banks) shown in the box.
2. According to the copula approach, no clear clustering of the eight assets seems to be plausible. On the other hand, a common feature is evident: Intesa SanPaolo and Unicredit are still the assets with the lowest $\Delta CoVaR_{0.05}^{f|i}$ (still shown inside the box).
3. The values of $\Delta CoVaR_{0.05}^{f|i}$ provided by the copula approach are generally smaller. The details can be found in Table 1 reporting the percentage of cases the $\Delta CoVaR_{0.05}^{f|i}$ obtained by the procedure in row is smaller than the corrensponding measure obtained using the procedure in column. We can conclude that the use of the Student's t copula function which admits tail dependence ensures a more flexible description of the association in the tails of the distributions allowing smaller quantiles.

Table 1 Percentages of success (smaller value) of $\Delta CoVaR_{0.05}^{f|i}$ provided by the copula procedure (marginal distributions are in brackets).

	QR (Gaussian)	QR (Studens *t*)	QR (Skew Student's *t*)
Copula (Gaussian)	75%	75%	75%
Copula (Student's *t*)	100%	100%	100%
Copula (Skew Studens *t*)	100%	100%	100%

References

1. Adrian T., Brunnermeier M.K.: CoVaR. NBER working paper series, w17454 (2011)
2. Girardi G., Ergun T.: Systemic risk measurement: Multivariate GARCH estimation of CoVaR. J. Bank. Financ., **37**, 3169–3180 (2013)
3. Jondeau E., Rockinger M.: The Copula-GARCH model of conditional dependencies: An international stock market application. J. Int. Money Financ. **25**, 827–853 (2006)
4. Koenker R., Bassett G.: Regression Quantiles. Econometrica **46**, 33-50 (1978)

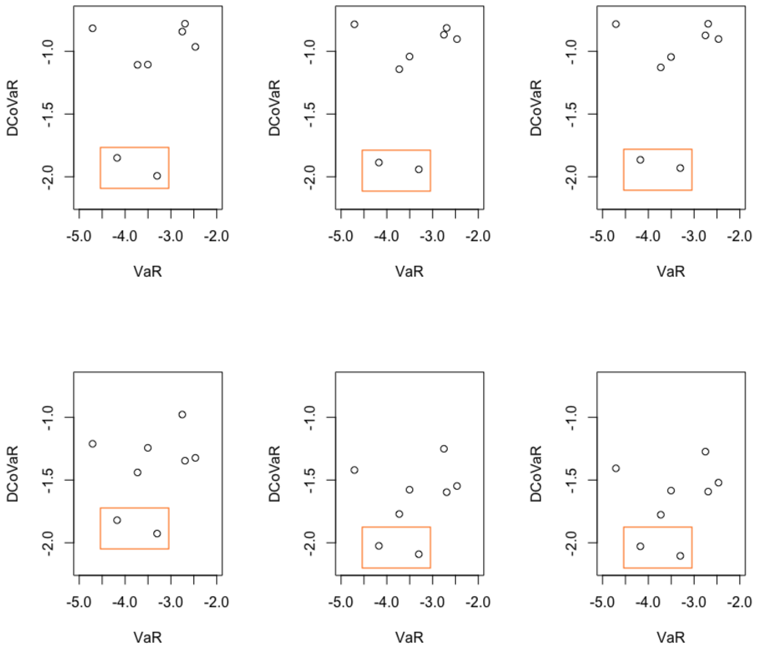


Fig. 1 Scatter plot of VaR-ΔCoVaR for the eight assets ($q = 0.05$). Two methods have been considered: quantile regression (top) and copula functions (below). Moreover, different distributional assumptions have been considered: Normal (left), Student's t (middle) and Skew Students's t (right).

Systemic events and diffusion of jumps

Eventi sistemici e diffusione dei jumps

Giovanni Bonaccolto, Nancy Zambon and Massimiliano Caporin

Abstract We propose two indexes informative of the cross-sectional diffusion of jumps from the analysis of a very large dataset of high-frequency returns that is not common in the literature. The two indexes have important implications in terms of asset pricing, as they capture part of the variability in stock returns that is not explained by the factors of the standard capital asset pricing model.
Abstract *Attraverso l'analisi di un ampio dataset di rendimenti ad alta frequenza, non comune nella letteratura, proponiamo due indici che forniscono informazioni sulla diffusione dei jumps tra le varie societá analizzate. Tali indici sono particolarmente informativi in termini di pricing, dato che incorporano una parte della variabilitá dei rendimenti che non é spiegata dai fattori che caratterizzano il tradizionale capital asset pricing model.*

Key words: multiple co-jumps, systemic jumps, systematic jumps, cross-sectional jump diffusion, systemic risk.

1 Data and jump detection

We analyse co-jumps involving a relatively large number of stocks using a huge dataset of high-frequency returns, which is not common in the literature. The

Giovanni Bonaccolto
University of Enna "Kore", viale delle Olimpiadi, 94100 Enna,
e-mail: giovanni.bonaccolto@unikore.it

Nancy Zambon
Department of Economics and Management, University of Padova, via del Santo 33, 35123 Padova,
e-mail: nancy.zambon@unipd.it

Massimiliano Caporin
Department of Statistical Sciences, University of Padova, via C. Battisti 241, 35121 Padova,
e-mail: massimiliano.caporin@unipd.it

database includes the $N = 3,509$ assets belonging to the basket of the Russell 3000 index for the period January 2, 1998—June 5, 2015.[1] The stock prices are sampled at a frequency of 1 minute from 09:30 a.m. to 04:00 p.m. for each of the 4,344 business days.[2] As a result, we record at the t-th day $M = 390$ 1-minute closing prices for each stock, denoted as $p_{t,i}$, for $t = 1,...,T$ and $i = 1,...,M$. Following a common practice—see [3] and [6], among others—we discard the first 5 minutes of each day to avoid potentially erratic price behaviour resulting from market opening.

[4] note that very high-frequency data are mostly composed of market microstructure noise and suggest to use a 5-minute frequency to mitigate microstructure effects. As a result, our empirical analysis builds on 5-minute returns, that we obtain by aggregating the original 1-minute returns.[3] To cope with market liquidity conditions, we restrict the attention on stocks with a sufficient number of non-null intraday returns to obtain accurate estimates of integrated volatility and jumps. In particular, we implement testing methods, described below, under the condition that, on a given trading day, the percentage of non-zero intraday returns is greater than or equal to 75%. In contrast, we treat the days on which the percentage of non-null returns is lower than 75% as days where no jumps occur. With some abuse of wording, we define assets with more than 25% of intradaily returns equal to zero as illiquid.

We use the *C-Tz* test proposed by [5] to identify the presence of jumps within each trading day in a cross-section of Russell 3000 constituents. Notably, the *C-Tz* test provides greater power than other tests based on multipower variation (see [5]). Following [2], we implement the test after standardising the returns to correct for volatility periodicity. Therefore, we improve the detection of small jumps during low volatility periods and reduce spurious detections of jumps at high volatility times. We detect the presence of jumps at both daily and intradaily levels. Then, we also gain knowledge about the location of jumps during the day. We highlight that only a few works use non-parametric tests to explicitly detect intraday jumps.

2 Common jumps

We analyse the cross-sectional diffusion of jumps by using intradaily returns with a 5-minute frequency. By using the co-exceedance rule of [6], we first implement the *C-Tz* test at the significance level $\alpha = 0.01\%$ to detect intraday jumps. Then, we compute the following variable:

$$C_{t,i} = \sum_{j=1}^{N} \mathbb{I}\{\text{Jump}_{t,i,j} > 0\} \begin{cases} \geq 2 & \text{Co-jump} \\ \leq 1 & \text{Single jump} \end{cases} \tag{1}$$

[1] The dataset is provided by Kibot, and the details are available at http://www.kibot.com. Our dataset includes also dead stocks or stocks that are no longer included in the Russell 3000 index.

[2] The original number of business days in our sample is equal to 4,384. We discard 40 days for which we observe particular tight liquidity conditions.

[3] The results obtained with other frequencies are available upon request.

to verify whether two or more assets record a jump in the same interval, where $\mathbb{I}$ is an indicator function taking the value of 1 when a jump is detected for asset j ($j = 1,\dots,N$, $N = 3509$) at the intraday interval i ($i = 1,\dots,77$ using 5 minutes intervals) on day t ($t = 1,\dots,4344$), and the value of 0 otherwise.

Notably, we checked that the *C-Tz* test is better able to detect common jumps than other common tests in the literature, such as the *s-BNS* test of [1]. It also highlights structural changes from 2001. Interestingly, we observed that around lunchtime co-jumps tend to increase, whereas individual jumps decrease. Several studies report the existence of a U-shaped pattern for the average volume of traded stocks and, in particular, relatively light trading in the middle of the day. The co-jump intraday pattern, with larger detection in the middle of the day, supports such evidence.[4]

Table 1 Asset jumps and market jumps. The table reports the number of days in which we observe at least one RUA jump (N_{RUA}), the amount of days in which we observe at least one intraday jump (N_j) or one co-jump (N_{cj}) in the constituents of the Russell 3000 and the days with both a jump in the index and a jump ($N_{RUA} \cap N_j$) or a co-jump ($N_{RUA} \cap N_{cj}$) in the underlying assets. These statistics are computed for the period January 1998—June 2015. We consider three observation intervals—1, 5 and 11 minutes. The results are obtained by implementing the *C-Tz* test.

Frequency	N_{RUA}	N_j	$N_{RUA} \cap N_j$	N_{cj}	$N_{RUA} \cap N_{cj}$
1 min	1,512.00	4,119.00	1,418.00	3,858.00	1,345.00
5 min	176.00	4,333.00	175.00	4,109.00	172.00
11 min	57.00	4,344.00	57.00	4,269.00	57.00

Jumps of individual stocks may affect the entire market. For instance, market-level news causing co-jumps of individual stocks might also be reflected in jumps of market portfolios. Further, co-jumps of stocks involving the market index can be seen as non-diversifiable events, with important implications for portfolio selection and hedging. Here, we define as systematic the co-jumps that occur simultaneously with a jump in the market index. Likewise, we define as non systematic the co-jump events detected from single-asset co-jumps but not reflected in a jump at the market index level. In short, we label as RUA jumps those that occur on the Russell 3000 index (RUA), our proxy for the market index. Table 1 shows jump days for the market index, jump and co-jump days in the underlying assets and the number of jump and co-jump days that are also RUA jump days. As a robustness check, Table 1 reports the results for three observation intervals, that is, 1, 5 and 11 minutes. Interesting findings emerge from Table 1. First, jump (N_j) and co-jump (N_{cj}) days are positively related to the interval length, whereas the opposite holds for the RUA index (N_{RUA}). Second, the number of days with at least one intraday co-jump is always lower than the corresponding number of days with at least one intraday jump.

[4] Tables and figures displaying such evidences, that we omit here for the sake of space, are available upor request.

Nevertheless, it still takes relevant values: from a minimum of 3,858 days (89% of the sample days) to a maximum of 4,269 days (98% of the sample days). Thus, we observe a co-jump for nearly each day of the sample. We stress that the definition of co-jumps is not particularly restrictive and, thus, we might have co-jumps involving only a small number of assets. Third, columns 4 and 6 of Table 1 present the intersections between jump days in the market index and jump or co-jumps days detected in the underlying assets. $\mathrm{N}_{RUA} \cap \mathrm{N}_j$ and $\mathrm{N}_{RUA} \cap \mathrm{N}_{cj}$ are useful to evaluate the capability of RUA to reflect cross-sectional jump events. Starting from the evidence that the majority of jumps and co-jumps do not occur simultaneously with jumps in the index, it is possible to deduce that the jumps in the index are not really informative of the presence of jumps and co-jumps in the cross-section. Since we are aggregating results for daily frequency, outcomes derived from intraday intervals would show even fewer intersections.

3 Multiple co-jumps, diffusion indexes and pricing

We now define as multivariate jump (or MJ) the subset of co-jumps such that:

$$MJ_{t,i} = \begin{cases} \sum_{j=1}^{N} \mathbb{I}\{\mathrm{Jump}_{t,i,j} > 0\} & \text{if } \sum_{j=1}^{N} \mathbb{I}\{\mathrm{Jump}_{t,i,j} > 0\} \geq K \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $K > 2$.

On the basis of (2), we build two indexes: a daily diffusion index (or DID) and an intraday diffusion index (or DII). The DID, for each day from January 2, 1998 to June 5, 2015, equals the largest number of stocks simultaneously jumping within the day. Note that the index might also take a zero value when no MJ occurs in a given day. The DII, in contrast, has an intradaily frequency of 77 observations per day. Each observation points out the number of stocks involved in a multivariate jump, if present, and 0 otherwise. The aim is to analyse the pricing implications of multivariate jumps by extending the standard capital asset pricing model (CAPM)—see [9], [7] and [8]. For the DID (a daily index), we estimate the CAPM and our two-factor model, respectively defined as follows:

$$R_{t,j} - R_{t,F} = \alpha_j + \beta_j MKT_t + e_{t,j}, \tag{3}$$

$$R_{t,j} - R_{t,F} = \alpha_j + \beta_j MKT_t + \beta_{DID,j} DID_t + e_{t,j}, \tag{4}$$

where $R_{t,j}$ is the daily return of the j-th asset, $R_{t,F}$ is the risk-free return that we record from the Kenneth R. French data library, DID_t is the daily diffusion index computed using the C-Tz test and $e_{t,j}$ is a zero-mean residual; $MKT_t = (R_{t,M} - R_{t,F})$ is the excess return on a capitalisation-weighted stock market portfolio, where $R_{t,M}$ is the daily RUA Index return.

Table 2 shows the statistical significance of estimated β_{DID} (denoted as $\hat{\beta}_{DID}$) along with the variations in the R^2_{adj} values we obtain, including the diffusion index in the CAPM model. The second column of Table 2 reports the percentage of $\hat{\beta}_{DID}$ with absolute t-statistic greater than 1.645 (10% significance level), using different time windows. Considering the full sample, January 2, 1998—June 5, 2015, we observe that 10% of $\hat{\beta}_{DID}$s are statistically significant. This suggests that the diffusion index could be an important risk factor in asset pricing. Table 2 also reports results for different sub-periods, highlighting how the relevance of DID changes over time and, in particular, focusing on economic crises. DID slopes are significantly different from zero in a relevant number of cases for all sub-periods, with values larger than the full-sample regression. Moreover, it appears that DID slopes are more frequently significant during the pre-2008 economic crisis, namely from 2002 until 2006. The analysis of R-squared highlights the ability of DID to capture part of the variation in stock returns not explained by the traditional market factor. Table 2 reports information on the variations in the R^2_{adj} values we obtain, including the diffusion index in the CAPM model. Min and Max correspond, respectively, to the minimum and maximum difference values, while $Q(0.25)$, Median and $Q(0.75)$ show the values for the first, second and third quartiles of the R^2_{adj} variation. Even if the majority of the variations are negative, the third and fourth quartiles suggest that many variations are positive and larger in absolute value with respect to negative variations. Increases are particularly pronounced during the years 2002—2006 and 2012—2015.

We now move our focus to intraday data. Similar to the daily case, we run monthly regressions using 5-minute data to study how DII helps in explaining stock returns. Our two-factor model for intraday data is:

$$R_{t,i,j} - R_{t,i,F} = \alpha_j + \beta_j MKT_{t,i} + \beta_{DII,j} DII_{t,i} + e_{t,i,j}, \tag{5}$$

where $R_{t,i,j}$ is the return on a security j, on day t for the intraday interval i, $R_{t,i,F}$ is the risk-free return that we approximate equal to 0, $DII_{t,i}$ is the C-Tz intraday diffusion index, $MKT_{t,i} = (R_{t,i,M} - R_{t,i,F})$ is the excess return on the Russell 3000 market portfolio and $e_{t,i,j}$ is a zero-mean residual. The use of high-frequency data allows us to obtain long samples of stock returns. Consequently, it is possible to run regressions using data from a reduced number of days and thus track how the significance of β_{DII} changes over time. We estimate the parameters of the model using non-overlapping rolling windows with a size of 22 days, which corresponds to 1,694 5-minute observations, or about one month of data. We observed high fractions of significant betas for almost all intervals from 2004 until 2015. This confirms that multivariate jumps help to explain stock returns by capturing common variations that are missed by the standard market factor and that might have some economic relevance when focusing on high-frequency data. Moreover, we do not observe higher levels of significance during the 2008 pre-crisis months but, instead, high picks clustered in 2007, 2008, 2010 and 2013.[5] Therefore, the DID performs well during calm periods, while the DII is more effective during more turbulent economic phases.

[5] Tables and figures displaying such results, omitted here for the sake of space, are available upon request.

Table 2 $\hat{\beta}_{DID}$ **significance and DID** R^2_{adj} **variation**. Here, we compare the CAPM model defined in (5) with our two-factor model defined in (6). Regressions are subject to the condition that stocks presents at least 251 days (about a year) of non-null returns in the window of interest. Column $\hat{\beta}_{DID}$ reports the percentage of times in which $\hat{\beta}_{DID}$ in (6) is statistically significant at the 0.1 level in the window of interest. From the third to the seventh column we focus on the variation in the coefficient of determination (or CoD) we observe by moving from the CAPM to the two-factor model. Min and Max are respectively the minimum and the maximum difference values whereas $Q(0.25)$, Median and $Q(0.75)$ are the first, second and third quartiles of the same differences.

Window	$\hat{\beta}_{DID}$	$R^2_{adj}(\text{2-factor}) - R^2_{adj}(\text{CAPM})$				
		Min	$Q(0.25)$	Median	$Q(0.75)$	Max
1998-2015	10%	-0.000	-0.000	-0.000	0.000	0.020
2002-2006	35%	-0.001	-0.001	-0.000	0.008	0.072
2007-2011	11%	-0.001	-0.001	-0.000	0.000	0.011
2012-2015	20%	-0.003	-0.001	-0.001	0.000	0.123

Acknowledgements The authors acknowledge financial support from University of Padova project CPDA143827/14 Multi-jumps in financial asset prices: detection of systemic events, relation with news, and implications for pricing.

References

1. Barndorff-Nielsen, O.E., Shephard, N.: Econometrics of testing for jumps in financial economics using bipower variation. J. Financ. Econ. **4**, 1–30 (2006)
2. Boudt, K., Croux, C., Laurent, S.: Robust estimation of intraweek periodicity in volatility and jump detection. J. Empirical Finance. **18**, 353–367 (2011)
3. Caporin, M., Kolokolov, A., Renò, R.: Systemic co-jumps. J. Finan. Econ. **forthcoming** (2017)
4. Christensen, K., Oomen, R.C., Podolskij, M.: Fact or friction: Jumps at ultra high frequency. J. Finan. Econ. **114**, 576–599 (2014)
5. Corsi, F., Pirino, D., Renò, R.: Threshold bipower variation and the impact of jumps on volatility forecasting. J. Econometrics. **159**, 276–288 (2010)
6. Gilder, D., Shackleton, M.B., Taylor, S.J.: Cojumps in stock prices: Empirical evidence. J. Banking Finance. **40**, 443–459 (2014)
7. Lintner, J.: The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Rev. Econ. Statist. **47**, 13–37 (1965)
8. Mossin, J.: Equilibrium in a capital asset market. Econometrica. **34**, 768–783 (1966)
9. Sharpe, W.F.: Capital asset prices: A theory of market equilibrium under conditions of risk. J. Finance. **3**, 425–442 (1964)

Traffic Lights for Systemic Risk Detection

Massimiliano Caporin

Department of Statistical Sciences
University of Padova

Laura Garcia-Jorcano

Instituto Complutense de Análisis Económico (ICAE)
Facultad de Ciencias Jurídicas y Sociales, Toledo
Universidad de Castilla-La Mancha

Juan-Angel Jiménez-Martin

Instituto Complutense de Análisis Económico (ICAE)
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense de Madrid

June 2018

Abstract

Girardi and Ergün (2013) (GE) modify Adrian and Brunnermeier's (2016) Conditional Value at Risk (CoVaR) from the maximum loss of the system conditional on the financial institution being in its VaR, to the financial institution being at most at its VaR. Here we extend GE's CoVaR computation using Filtered Historical Simulation and a DCC multivariate GARCH model of Engle (2002). Filtered Historical Simulation allows us to compute one-day ahead forecasts of CoVaR and ΔCoVaR. Additionally, we propose a new Traffic Light System (TLS) for Systemic Risk Detection that provides a comprehensive color-based classification that groups companies according to both the level of stress reaction of the system when the company is in distress and the level of stress of the company. TLS can be used to enhance the performance and robustness of current systemic risk measures.

Keywords: DDC, conditional correlations, CoVaR, FHS, Systemic Risk

JEL codes: G11, G21, G32, G38

Acknowledgment. The authors wish to thank the Ministerio de Economía, Industria y Competitividad of Spain through the research project ECO2015-67305-P and the Banco de España through the research project PR71/15-20229. JA Jimenez-Martin thanks the Department of Statistical Science for their hospitality during a research visit at the University of Padova and the Spanish Ministry of Education, Culture and Sport through the grant PRX17/00404

1 – Introduction

The recent financial crisis has fueled the search for precise systemic risk measures. Several papers have proposed different systemic risk measures as Billio et al. (2012), Zhou (2010), Huang et al. (2009), Segoviano and Goodhart (2009), Acharya et. al. (2010), Brownless and Engle (2011) and Allen et al. (2010). Girardi and Ergün (2013) (GE) modify Adrian and Brunnermeier's (2011, 2016) Conditional Value at Risk (CoVaR), the VaR of the financial system conditional on an institution being in financial distress. GE propose a change in the definition of CoVaR, from the maximum loss of the system conditional on the financial institution being in its VaR, to the financial institution being at most at its VaR.

CoVaR is an extension of the VaR which only gauges the risk of an institution in isolation. Therefore, computing CoVaR required precise estimation of the relationship between the system and the company. GE report time varying in sample CoVaR estimating a bivariate GARCH model with Engle's (2002) DCC specification for the returns of the institution and the financial system to fit volatility and correlations. Mainik and Shaanning (2012) show that the dependence parameter between institutions and system plays a key role in the accurate systemic risk measurement. In particular, as the institution and the system become more and more correlated, the institution systemic risk increases. This fact makes accurate relationship measures extraordinarily important when analyzing financial distress.

In this paper, we extend GE's CoVaR computation using Filtered Historical Simulation and a DCC (Engle, 2002) multivariate GJR-GARCH model of Glosten, Jagannathan, and Runkle, (1993). Following GE, we start by estimating CoVaR for 58 out of 78 companies used in GE and Acharya et. al. (2010) with an extended sample from January 2000 until January 2018. A novel contribution of this paper is about the distributional assumption on the error. Multivariate normal and skewed-t distributions have already been used in GE. Risk measurement might suffer from the shortcomings involved in assuming a parametric error distribution. Filtered Historical Simulation (FHS) has emerged as one of the best tools for calculating risk measures as VaR and consequently a robust alternative as a procedure to forecast CoVaR. FHS, introduced by Barone-Adesi et al. (1998) and Barone-Adesi et al. (1999) uses a combination of volatility models and past returns to build the probability distribution of the future returns. The daily row returns are first scaled by the volatility that prevail that day and then are multiplied by the current forecast of volatility. The first step, the scaling, is necessary in order to make the past returns stationary and to provide suitable innovations for a simulation process. The second step, the re-scaling, endows the historical returns to reflect the current volatility conditions prevailing in the markets.

Combining the MGARCH specifications and FHS we produce 1-day ahead CoVaR forecast using a multi-step modelling approach: (1) a multivariate GJR-GARCH DCC (Engle, 2002) model is used to fit volatility and correlations; (2) FHS is used to simulate paths of 1-day ahead forecast returns. With FHS the distributional assumptions of the errors are relaxed, although the market condition are taking into account by using the conditional volatility and correlations estimated in step one; (3) from the joint empirical distribution of the simulated returns for the system and company, the empirical CoVaR and $\Delta CoVaR$ are computed. $\Delta CoVaR$ was proposed by Adrian and Brunnermeier (2011, 2016) and is computed as the difference between the CoVaR of the systems and the unconditional financial system VaR. $\Delta CoVaR$ captures the marginal contribution of a particular institution to the overall systemic risk. However, $\Delta CoVaR$ is not sufficient as a precise measurement of company systemic risk.

For instance, what happen to the system in a company's high stress state, when the realized loss exceeded the VaR forecast, might be of particular relevance for an accurate ranking of systemic companies. To see our argument more explicitly, let's assume that two institutions, *j* and *i,* individually contribute to a system's large loss (realized loss exceeds $CoVaR$), they even might have reported the same $\Delta CoVaR$. Nonetheless, when analyzing the company′s losses of *j* and *i*, it might happen that the *i-th company*'s loss is relatively shorter than *j*'s loss (*j'*s companies exceedance is larger than *i*'s company). Under these circumstances, *i* and *j*'s contribution to the systemic risk should not be reported as similar. Institution *i* seems to be riskier than company *j.* Short exceedances in the *i-th company* hurt the system as much as large exceedances in company *j* do.

Our Traffic Light System (TLS) for Systemic Risk Detection proposed in this paper deals with the above issue. *TLS* makes use of loss functions frequently used in backtesting to provide a more accurate systemic risk ranking for financial companies. TLS classifies financial institutions by a color code that will group companies into four categories according to the joint behavior of the company and system exceedances magnitudes.

The remainder of the paper is organized as follows. In Section 2, we describe CoVaR and ΔCoVaR and the way to compute it. Section 3 presents the new TLS. Section 4 describes the results, including a preliminary statistical analysis, and Section 5 presents the conclusions.

2 – Measuring Systemic Risk with ΔCoVaR under Multivariate GARCH

The Conditional Value-at-Risk (CoVaR) is an extension of the Value-at-Risk (VaR) which only gauges the risk of an institution in isolation. The classical formulation of Value-at-Risk implicitly defines it as the *q*-quantile of the return

$$\Pr\left(R_t^j \leq VaR_{q,t}^j\right) = q \tag{1}$$

where $VaR_{q,t}^j$ is the *q%* Value-at-Risk for company j, whose asset returns are R^j. The CoVaR extends the classical VaR definition by adding a conditional event, and points at the estimation of the conditional quantile of the financial system rather than on a single company. In this paper, we use the definition of CoVaR proposed by Girardi and Ergün (2013). They condition the evaluation of the system Value-at-Risk on the event that institution *j* is *at most at* its VaR. Such a choice differs from the most know definition of CoVaR due to Adrian and Brunnermeier (2011). In fact, the latter condition on institution *j* being *exactly* at its VaR. Therefore, Girardi and Ergün (2013) defines $CoVaR_{q,t}^{s|j}$ as the VaR of the system conditional on institution *j* been *at most at* its VaR. This is implicitly defined by the *q%* - quantile of the conditional probability distribution:

$$\Pr\left(R_t^s \leq CoVaR_{q,t}^{s|j} \mid R_t^j \leq VaR_{q,t}^j\right) = q \tag{2}$$

where $CoVaR_{q,t}^{s|j}$ is expressed in terms of R^s, the asset returns of the system and $VaR_{q,t}^j$ is the Value-at-Risk of company *j*. Note that we use the same reference level, *q*, for both the system and the company; this is not a required and we can easily allow for different levels in the VaR of the company and the CoVaR of the system. The CoVaR is a system-based risk measure conditional to the possible presence of a distress in one of the companies included in the financial system.

Adrian and Brunnermeier (2015), followed by Girardi and Ergün (2013), propose not to monitor directly the CoVaR but a different quantity, namely, the marginal contribution to systemic risk of a single institution. They suggest focusing on ΔCoVaR, which is the difference between the Conditional VaR of the system when the institution is in *distress* and the Conditional VaR of the system when the institution is in a *normal state*. In the CoVaR formulation of Girardi and Ergün (2013) this rationalizes as

$$\Delta CoVaR_{q,t}^{s|j} = 100 \times (CoVaR_{q,t}^{s|j} - CoVaR_{q,t}^{s|b^j}) / CoVaR_{q,t}^{s|b^j} \qquad (3)$$

where $CoVaR_{q,t}^{s|b^j}$ is the VaR of the financial system conditional on the *benchmark* state of institution j, b^j, defined as a one-standard deviation about the mean event: $\mu_t^j - \sigma_t^j \le R_t^j \le \mu_t^j + \sigma_t^j$ where μ_t^j and σ_t^j are the conditional mean and standard deviation of institution j, respectively, evaluated with a proper model. For additional details, we refer the reader to Girardi and Ergün (2013).

3.- Traffic Light System for Systemic Risk Detection

This paper presents an approach for reliably assessing financial companies' relative contribution to the systemic market stress as implied by distress states of the companies. Our methodological contribution, summarized into a Traffic Light System (TLS) for Systemic Risk Detection, provides a comprehensive color-based classification that groups companies according to both the level of stress reaction of the system when the company is in distress and the level of stress of the company. In particular, TLS produces a quantitative assessment of relative contribution to systemic stress based on the magnitude of exceedances while both the company and the system are in distress. TLS uses appropriate loss functions to quantify the level company and system's stress.

4 – Data description and main results

4.1. Data description

We work with daily data for 58 US financial institutions on the same four industry groups considered by Girardi and Ergün (2013). Table 1 lists these financial institutions and their type based on two-digit SIC classification code: Depositories Institutions, Securities Dealers and Commodity Brokers, Insurance, and Others. We use the Dow Jones US Financial Index (DJUSFN) as a proxy for the financial system, as Girardi and Ergün (2013). In our case, the full sample goes from January 3rd, 2000 to January 15th, 2018 (4706 observations). A rolling window of 1000 observations is used in the estimation of the model, and 3705 one-day-ahead forecasts are produced since November 3rd, 2003 up to the end of the sample. We compute daily returns as the first difference of log prices, i.e. $[\ln(P_{t+1})-\ln(P_t)]$. We recovered all data from Thomson Reuters Datastream.

4.3. TLS at work

We use TLS to classify the 58 companies used in Girardi and Ergün (2013). Figure 1 shows the companies labelled according TLS. Horizontal axis represents $\left(MD_{t+1}^{i}-\overline{z}_{t+1}^{c}\right)$, i.e. the magnitude of the company's stress state (high stress state when positive, low stress state when negative). Vertical axis represents $(MD_{t+1}^{s|i}-\overline{z}_{t+1}^{s|c})$, i.e. the magnitude of the system's stress state (high stress state when positive, low stress state when negative). The size of the circles is based on the company capitalization in the analyzed period.

5.- Conclusions

The main conclusions are as follows:

1. We provide a new color-based systemic indicator, TLS for systemic risk detection using loss functions frequently used in backtesting to provide a more accurate systemic risk ranking for financial companies.

2.- TLS provides a company ranking system identifying various levels of systemic risk providing a color-based code defined in the line of The Basel Committee's Traffic Light to classify financial institutions. The new TLS is an intuitive, clear and powerful tool for monetary authorities being a complementary to other well-known systemic risk measures.

References

Acharya, V., Pedersen, L.H., Philippon, T., Richardson, M., 2010. Measuring Systemic Risk. Working paper, New York University.

Adrian, T., Brunnermeier, M.K.,

2011. CoVaR. Working paper, Federal Reserve Bank of New York.

2016. CoVaR. American Economic Review 106 (7), 1705-1741

Allen, L., Bali, T.G., Tang, Y., 2010. Does systemic risk in the financial sector predict future economic downturns? Working paper, SSRN.

Barone-Adesi G, F Bourgoin and K Giannopoulos, 1998, "Don't Look Back", Risk, 11, August,100-104.

Barone-Adesi G, K Giannopoulos and L Vosper, 1999, "VaR Without Correlations for Non-Linear Portfolios", Journal of Futures Markets, 19, August, 583-602.

Billio, M., Getmansky, M., Lo, A.W., Pelizzon, L., 2012. Econometric measures of systemic risk in the finance and insurance sectors. Journal of Financial Economics 104, 535–559.

Brownlees, C., Engle, R., 2011. Volatility, Correlation and Tails for Systemic Risk Measurement. Working paper, New York University of CoVaR.

Caporin, M., (2008), Evaluating value-at-risk measures in the presence of long memory conditional volatility, Journal of Risk, 10(3), 79-110

Engle, R.F., 2002. Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroscedasticity models. Journal of Business and Economic Statistics 20, 339–350.

Girardi, G. and Ergün, A. T. (2013), Systemic risk measurement: Multivariate GARCH estimation of CoVaR. Journal of Banking and Finance, 37, 3169-3180.

Glosten, L.R., Jagannathan, R. and Runkle, D.E. (1993), Relationship between the expected value and the volatility of the nominal excess return on stocks, The Journal of Finance, 48(5), 1779-1801.

Huang, X., Zhou, H., Zhu, H., 2009. A framework for assessing the systemic risk of major financial institutions. Journal of Banking and Finance 33, 2036–2049.

Zhou, C., 2010. Are banks too big to fail? Measuring systemic importance of financial institutions. International Journal of Central Banking 6, 205–250.

Figures and Tables

Figure 1. Full period

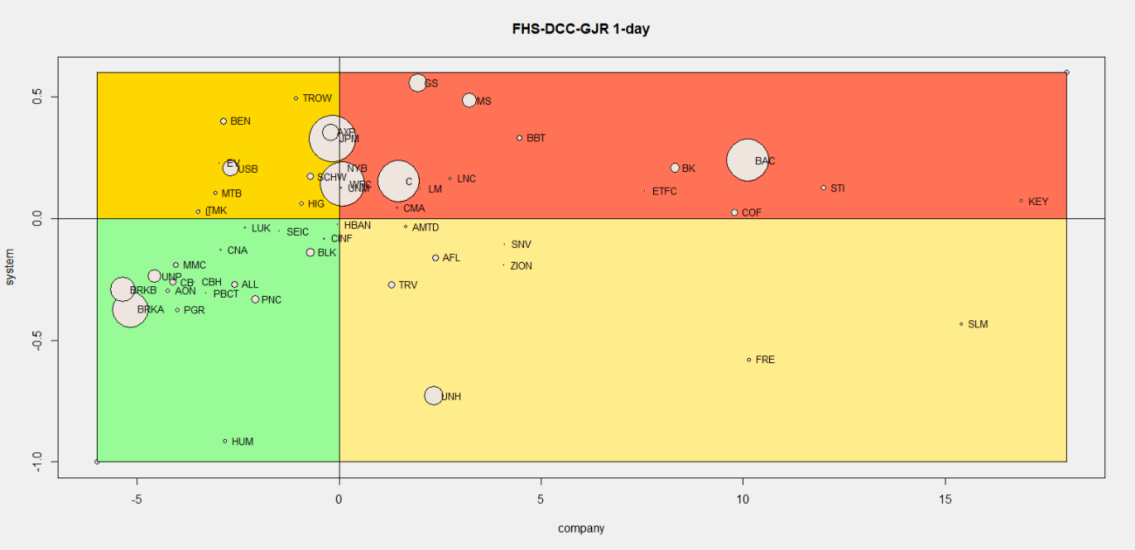


Table 1 .- Names and classifications of 58 US financial institutions

21 (-7)	Depositories	11 (-4)	Others	21(-2)	Insurance	5 (-3)	Broker-dealers
BAC	Bank of America	**AMTD**	Ameritrade Holding	**AFL**	AFLA Inc	**ETFC**	E-Trade Financial
BBT	BB & T	**AXP**	American Express	**AIG**	American International Group	**GS**	Goldman Sachs
BK	Bank of New York Mellon	**BEN**	Franklin Resources Inc	**ALL**	Allstate Corp	**MS**	Morgan Stanley
C	Citigroup	**BLK**	Blackrock Inc	**AON**	AON Corp	**SCHW**	Schwab Charles Corp
CBH	Commerce Bancorp Inc NJ	**COF**	Capital One Financial	**BRKA**	Berkshire Hathaway Inc Del	**TROW**	T Rowe Price
CMA	Comerica Inc	**EV**	Eaton Vance Corp	**BRKB**	Berkshire Hathaway Inc Del		
HBAN	Huntington Bancshares Inc	**LM**	Legg Mason Inc	**CB**	Chubb Corp		
JPM	JP Morgan Chase	**LUK**	Leucadia national	**CINF**	Cincinnati Financial Corp		
KEY	Keycorp New	**SEIC**	Sei Investments Company	**CNA**	Can Financial Corp		
MTB	M & T Bank Corp	**SLM**	S L M Corp	**HIG**	Hartford Financial Svcs Group		
NTRS	Northern Trust Corp	**UNP**	Union Pacific	**HUM**	Humana Inc		
NYB	New York Community Bankcorp			**L**	Loews Corp		
PBCT	People United Financial			**LNC**	Lincoln national Corp		
PNC	PNC Financial Services			**MBI**	MBIA Inc		
RF	Regions Financials			**MBI**	MBIA Inc		
SNV	Synovus Financial Corp			**MMC**	Marsh and Mclennan Cos Inc		
STI	Suntrust Banks Inc			**PGR**	Progressive Corp OH		
STT	State Street Corp			**TMK**	Torchmark Corp		
UB	Unionbancal Corp			**TRV**	Travelers Companies Inc		
USB	US Bancorp Del			**UNH**	United Health Group		
WFC	Wells Fargo			**UNM**	Unum Group		
ZION	Zions Bancorp						

Bayesian Quantile Regression Treed

Mauro Bernardi and Paola Stolfi

Abstract Decision trees and their population counterparts are becoming promising alternatives to classical linear regression techniques because of their superior ability to adapt to situations where the dependence structure between the response and the covariates is highly nonlinear. Despite their popularity, those methods have been developed for classification and regression, while often the conditional mean would not be enough when data strongly deviates from the Gaussian assumption. The approach proposed in this paper instead considers an ensemble of nonparametric regression trees to model the conditional quantile at level $\tau \in (0,1)$ of the response variable. Specifically, a flexible generalised additive model (GAM) is fitted to each partition of the data that corresponds to a given leaf of the tree, allowing an easy interpretation of the model parameters. Indeed, while the trees structure easily adapts to regions of the data having different shapes and variability, the nonlinear part handles parsimoniously the local nonlinear structural relationship of the quantile with the covariates. Unlike the most popular Bayesian approach (BART) that assumes a sum of regression trees, quantile estimates are obtained by averaging the ensemble trees, thereby reducing their variance. We develop a Bayesian procedure for fitting such models that effectively explores the space of B–Spline functions of different orders that features the functional nonlinear relationship with the covariates. The approach is particularly valuable when skewness, fat–tails, outliers, truncated and censored data, and heteroskedasticity, can shadow the nature of the dependence between the variable of interest and the covariates. We apply our model to a sample of US companies belonging to different sectors of the Standard and Poor's Composite Index and we provide an evaluation of the marginal contribution to the overall risk of each individual institution.

Key words: Quantile regression treed, Bayesian inference, Conditional Value–at–Risk.

1 Introduction

In empirical studies, researchers are often interested in analysing the behaviour of a response variable given the information on a set of covariates. The typical answer is to specify a linear regression model where unknown parameters are estimated by OLS, thereby leading to the approximation of the mean function. Although the mean describes the average response path as a function of the covariates, it provides little o no information about the behaviour of the conditional distribution on the tails. As far as the entire distribution is concerned, quantile regression methods adequately characterise the behaviour of the response variable at different confidence levels providing a complete picture of the relationship with the covariates. Moreover, the quantile analysis is particularly suitable when the conditional distribution strongly deviates from

Mauro Bernardi
Department of Statistical Sciences, University of Padova, Via Cesare Battisti, 241, 35121, Padova, e-mail: mauro.bernardi@unipd.it

Paola Stolfi
Institute for applied mathematics "Mauro Picone" (IAC) - CNR, Rome, Italy, e-mail: p.stolfi@iac.cnr.it

the Gaussian assumption because it displays heterogeneity, asymmetry or fat–tails, see, e.g., [9]. Linear quantile regression models have been extensively applied in different areas, such as, finance, engineering, econometrics and environmetrics, as a direct approach to quantify the level of risk of a given event, social sciences and quantitative marketing to find appropriate and effective solutions to specific segments of customers, and many other related fields see, [10]. However, despite their relevance and widespread application in empirical studies, linear quantile regression models provide only a rough "first order" approximation of the relationship between the τ–level quantile of the response variable and the covariates. Indeed, as first recognised by [9], quantiles are linear functions only within a Gaussian world, thereby stimulating many recent attempts to overcome this limitation. [6], for example, consider the copula–based approach to formalise nonlinear and parametric conditional quantile relationships. Although quite flexible in fitting marginal data, the copula approach forgets to consider nonlinear interactions among the covariates. Classification and regression trees (CART, [4]) and their population counterparts ([3]) extensively use recursive partitioning algorithms to perform nonparametric regression and variable selection. The attractive feature of decision trees methods rely in their ability to partition the covariates space into disjoint hyperrectangles, thereby improving the local fit. Therefore, CART adapt to situations where the dependence structure between the response and the covariates is highly nonlinear. Despite their extensive use in a wide variety of fields, those methods have been mainly developed for classification and mean regression. In this paper, we adopt the Bayesian point of view and we extend the Bayesian regression trees approach of [7] to deal with conditional quantiles estimation. Quantile estimation have been previously extended within the related context of random forest by [11]. However, unlike random forests, the Bayesian approach to decision trees learning, being likelihood–based, provides a complete inferential tool for model assessment and selection.

2 Quantile regression treed

The linear quantile regression framework for independent and identically distributed data models the conditional τ–level quantile of the response variable Y, with $\tau \in (0,1)$, as a linear function of the vector of dimension $(q \times 1)$ of exogenous covariates $\mathbf{X}$, i.e., $\mathscr{Q}_\tau(Y \mid \mathbf{X} = \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, thereby avoiding any explicit assumptions about the conditional distribution of $Y \mid \mathbf{X} = \mathbf{x}$. This is equivalent to assume an additive stochastic error term ε for the conditional regression function $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ to be independent and identically distributed with zero τ–th quantile, i.e, $\mathscr{Q}_\tau(\varepsilon \mid \mathbf{x}) = 0$, and constant variance. Following [12] and [2], the previous condition is implicitly satisfied by assuming that the conditional distribution of the response variable Y follows an Asymmetric Laplace (AL) distribution located at the true regression function $\mu(\mathbf{x})$, with constant scale $\sigma > 0$ and shape parameter τ, i.e., $\varepsilon \sim \mathsf{AL}(\tau, \mu(\mathbf{x}), \sigma)$, with probability density function

$$\mathsf{AL}(Y \mid \mathbf{X} = \mathbf{x}, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\frac{1}{\sigma}\rho_\tau(Y - \mu_\tau(\mathbf{x}))\right\} \mathbb{1}_{(-\infty,\infty)}(Y), \tag{1}$$

where $\mu_\tau(\mathbf{x})$ is the quantile regression function and $\rho_\tau(u) = u\left(\tau - \mathbb{1}_{(-\infty,0)}(u)\right)$ denotes the quantile check function at level τ. The quantile regression model postulated in equation (1) assumes the AL distribution as a misspecified working likelihood that correctly identify the conditional quantile function.

Unlike the Bayesian Additive Regression Tree (BART) approach of [8] which considers a sum–of–trees regression where each tree explains only a small portion of the total variance of the dependent variable Y, we model the conditional quantile of the response variable as a function of the covariates as the average of an ensemble of $m \in \mathbb{N}_+$ regression treeds

$$\begin{aligned} \mathscr{Q}(Y \mid \mathbf{X} = \mathbf{x}) &= \mu_\tau(\mathbf{x}) & (2)\\ &\approx \frac{1}{m}\sum_{l=1}^{m} \mathscr{T}_l^{\mathscr{M}_l}(\mathbf{x}), & (3) \end{aligned}$$

where $\mathscr{T}_j^{\mathscr{M}_j}$ denotes the j–th treed of the ensemble, for $j = 1, 2, \dots, m$. In equation (3) each regression trees is composed by a tree structure, denoted by $\mathscr{T}$, and the parameters of the terminal nodes (also called leaves), denoted by $\mathscr{M}$. Therefore, the j–th tree for $j = 1,2,\ldots,m$, denoted by $\mathscr{T}_j^{\mathscr{M}}$, represents a specific combination of tree structure $\mathscr{T}_j$ and tree parameters $\mathscr{M}_j$, i.e., the regression parameters associated to its terminal nodes.

3 Conditional Value–at–Risk estimation

In this section we apply the methodology introduced in the previously section to analyse the tail co–movements between a financial institution j and the whole financial system k. To this aim we consider the Conditional Value–at–Risk (CoVaR) recently introduced by [1], which is defined as the overall VaR of an institution, conditional on another institution being in distress. To be more specific, let (Y_j,Y_k) be the bivariate random variable describing the return of institutions j and k, for $k \neq j$ and assume (Y_j,Y_k) depend on a vector of exogenous covariates $\mathbf{X} = (X_1,X_2,\ldots,X_q)$, then the Conditional Value–at–Risk $\left(\mathrm{CoVaR}_{k|j}^{\mathbf{x},\tau}\right)$ is the Value–at–Risk of institution k conditional on $Y_j = \mathrm{VaR}_j^{\mathbf{x},\tau}$ at the level $\tau \in (0,1)$, i.e., $\mathrm{CoVaR}_{k|j}^{\mathbf{x},\tau}$ satisfies the following equation

$$\mathbb{P}\left(Y_k \leq \mathrm{CoVaR}_{k|j}^{\mathbf{x},\tau} \mid \mathbf{X} = \mathbf{x}, Y_j = \mathrm{VaR}_j^{\mathbf{x},\tau}\right) = \tau, \tag{1}$$

where $\mathrm{VaR}_j^{\mathbf{x},\tau}$ denotes the Value–at–Risk, $\mathrm{VaR}_j^{\mathbf{x},\tau}$ of institution j, i.e., the τ–th level conditional quantile of the random variable $Y_j \mid \mathbf{X} = \mathbf{x}$, defined as

$$\mathbb{P}\left(Y_j \leq \mathrm{VaR}_j^{\mathbf{x},\tau} \mid \mathbf{X} = \mathbf{x}\right) = \tau. \tag{2}$$

Note that both the VaR and the CoVaR corresponds to the τ–th quantiles of the conditional distribution of $Y_j \mid \{\mathbf{X} = \mathbf{x}\}$ and $Y_k \mid \{\mathbf{X} = \mathbf{x}, Y_j = \mathrm{VaR}_j^{\mathbf{x},\tau}\}$, respectively. Therefore, both the VaR and CoVaR equations can be estimated using the Bayesian quantile regression treed models introduced in the previous section.

The financial data we utilise are taken from the Standard and Poor's Composite Index (k) for the U.S market, where different sectors (j) are included. For both the institutions and for the whole system, we consider microeconomics and macroeconomics variables, in order to take into account for individual information and for global economic conditions respectively. Our empirical analysis is based on publicly traded U.S. companies belonging to different sectors of the Standard and Poor's Composite Index (S&P500) listed in Table 1. The sectors considered are: Financials, Consumer Goods, Energy, Industrials, Technologies and Utilities. Financials consists of banks, diversified financial services and consumer financial services. Daily equity price data are converted to weekly log–returns (in percentage points) for the sample period from January 2, 2004 to December 28, 2012, covering the recent global financial crisis. To control for the general economic conditions we use observations of the following macroeconomic regressors as suggested by [1] and [5]: the VIX index (VIX), measuring the model-free implied stock market volatility as evaluated by the Chicago Board Options Exchange (CBOE), a short term liquidity spread (LIQSPR), computed as the difference between the 3-month collateral repo rate and the 3-months Treasury Bill rate, the weekly change in the three-month Treasury Bill rate (3MTB) the change in the slope of the yield curve (TERMSPR), measured by the difference of the 10-years Treasury rate and the 3-month Treasury Bill rate, the change in the credit spread (CREDSPR) between 10-years BAA rated bonds and the 10-years Treasury rate and the weekly return of the Dow Jones US Real Estate Index (DJUSRE). To capture the individual firms' characteristics, we include observations from the following microeconomic regressors: leverage (LEV), calculated as the value of total assets divided by total equity (both measured in book values), the market to book value (MK2BK), defined as the ratio of the market value to the book value of total equity, the size (SIZE), defined by the logarithmic transformation of the market value of total assets, and the maturity mismatch (MM), calculated as short term debt net of cash divided by the total liabilities. To have a complete picture of the contributes from individual and systemic risk we plot the estimated VaR and CoVaR for some of the assets listed in Table 1 in Figure 1. Looking at individual risk assessment, it is clear that the VaR profiles are relatively similar across institutions, displaying strong negative downside effects upon the occurrence of the recent financial crises of 2008 and 2010 and the sovereign debt crisis of 2012. However, the analysis of the time

Name	Ticker Symbol	Sector
CITIGROUP INC.	C	Financial
BANK OF AMERICA CORP.	BAC	Financial
COMERICA INC.	CMA	Financial
JPMORGAN CHASE & CO.	JPM	Financial
KEYCORP	KEY	Financial
GOLDMAN SACHS GROUP INC.	GS	Financial
MORGAN STANLEY	MS	Financial
MOODY'S CORP.	MCO	Financial
AMERICAN EXPRESS CO.	AXP	Financial
MCDONALD'S CORP.	MCD	Consumer
NIKE INC.	NKE	Consumer
CHEVRON CORP.	CVX	Energy
EXXON MOBIL CORP.	XOM	Energy
BOEING CO.	BA	Industrial
GENERAL ELECTRIC CO.	GE	Industrial
INTEL CORP.	INTC	Technology
ORACLE CORP.	ORCL	Technology
AMEREN CORPORATION.	AEE	Utilities
PUBLIC SERVICE ENTERPRISE INC.	PEG	Utilities

TABLE 1: List of companies included in empirical analysis. All listed companies belonged to the Standard and Poor's Composite Index (S&P500) at the start of the trading day of February 15, 2013.

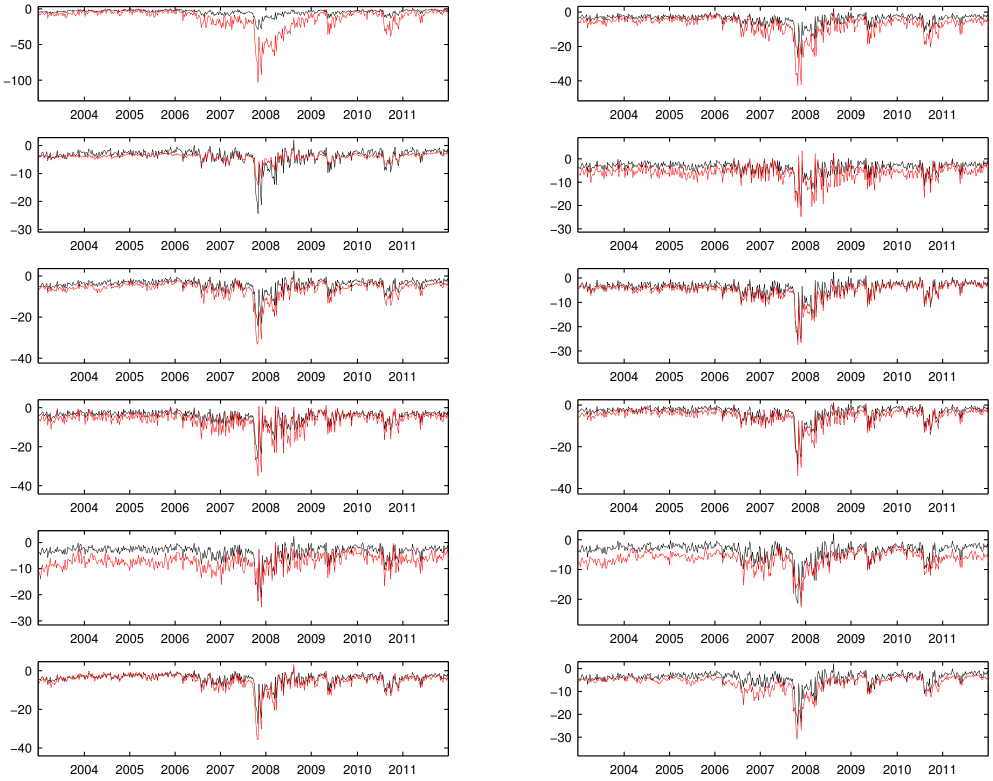


FIG. 1: Time series plot of the $\text{VaR}_j^{\mathbf{x},\tau}$ (red line) and $\text{CoVaR}_{k|j}^{\mathbf{x},\tau}$ (gray line) at the confidence level $\tau = 0.025$, for the following assets: top panel (financial): C (left), GS (right); second panel (consumer): MCD (left) and NKE (right); third panel (energy): CVX (left), XOM (right); fourth panel (industrial): BA (left), GE (right); fifth panel (technology): INTC (left), ORCL (right); bottom panel (utilities): AEE (left), PEG (right).

series evolution of the marginal contribution to the systemic risk, measured by CoVaR, reveals different behaviors for the considered assets. In particular, Citygroup (C), which belongs to the Financials, seems to contribute more to the overall risk than other assets do.

References

1. Adrian, T. and Brunnermeier, M., (2016). CoVaR. *American Economic Review*, 31, 106, 1705-1741.
2. Bernardi, M., Gayraud, G., and Petrella, L. (2015). Bayesian tail risk interdependence using quantile regression. *Bayesian Anal.*, 10(3):553–603.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
4. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.
5. Chao, S.-K., Härdle, W.F. and Chang, W. (2012). Quantile Regression in Risk Calibration. *Handbook for Financial Econometrics and Statistics* in Cheng-Few Lee, ed., Springer Verlag.
6. Chen, X., Koenker, R., and Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *Econometrics Journal*, 12:S50–S67.
7. Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948.
8. Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298.
9. Koenker, P., (2005). *Quantile Regression*. Cambridge University Press, Cambridge.
10. Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017). *Handbook of Quantile Regression*. CRC Press.
11. Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999.
12. Yu, K. and Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54:437–447.

Model Selection in Weighted Stochastic Block models

Selezione del modello per i modelli a blocchi stocastici

Roberto Casarin, Michele Costola, Erdem Yenerdag

Abstract We propose the weighted stochastic block model (WSBM) as generative model for the financial networks and exploit the topological features of its blocks. This model considers both the edge existence and the edge weight of the network and is independent from the methodology implemented on the estimation of the network. In this paper, we discuss three specifications of the model with by analysing the European financial network.

Abstract In questo lavoro, viene proposto il modello a blocchi stocastici pesato (WSBM) come modello generativo per i network finanziari dove vengono esplorate le caratteristiche topologiche dei blocchi. Questo modello considera sia il peso che l'esistenza del legame ed é indipendente dalla metodologia di stima del network. In questo articolo, discutiamo tre specifiche del modello analizzando il network finanziario Europeo.

Key words: financial networks, stochastic block model

1 Introduction

The study of the financial network topologies and connectedness provides useful tools in monitoring financial stability.

Roberto Casarin
Department of Economics, Ca' Foscari University of Venice, Dorsoduro 3246, 30123 Venice (Italy). e-mail: r.casarin@unive.it

Michele Costola
SAFE, House of Finance, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 3, 60323 Frankfurt am Main (Germany). e-mail: costola@safe.uni-frankfurt.de

Erdem Yenerdag
Department of Economics, Washington University in St. Louis, One Brookings Drive Campus Box 1208, Saint Louis, MO 63130 (USA). e-mail: erdemyenerdag@wustl.edu

Nodes exhibit often the propensity to be clustered into groups which are internally densely connected but show fewer connections outside. This feature is known as community structure or network modularity [Newman, 2010]. The aim of this paper is to investigate the network topology with focus on the detection of blocks in a broader sense where nodes represents financial assets. The communities are obtained through the Weighted Stochastic Block Model (WSBM) introduced recently by [Aicher et al., 2014] which allows through the blocks to have a compact characterization of the network's structure. This network generative model considers both the edge existence and the edge weight of the network and represents a generalization of the SBM [Holland et al., 1983] which is independent from the methodology implemented on the estimation of the network. By analysing the European financial institutions, we compare three specifications of the model such as the Stochastic Block Model (SBM), the balanced Weighted Stochastic Block Model (WSBM) and the pure WSBM (pWSBM).

2 Weighted Stochastic Block Models

The likelihood function of the basic Stochastic Block Model (SBM) is:

$$\mathscr{L}(A|z,\theta) = \prod_{ij} \theta_{z_i z_j}^{A_{ij}} (1-\theta_{z_i z_j})^{1-A_{ij}} \tag{1}$$

where A is an adjacency matrix of the network, which contains binary values representing edge existence, $A_{ij} \in \{0,1\}$, z is a vector that contains the group label of each node $z_i \in \{1,\dots,K\}$, where K is the number of latent groups. For example, if A is an $n \times n$ adjacency matrix then z is a $(1 \times n)$ vector such as $z = (z_1, z_2, \dots, z_i, \dots, z_n)$. Therefore, vector z represents the partition of the nodes into K blocks and each pair of groups kk' represents a bundle of edge between the groups. The parameter θ in Equation 1 represents a $(K \times K)$ matrix and its elements represent the edge existence parameters, $\theta_{z_i z_j}$, of each edge bundle. The existence probability of an edge A_{ij} is given by the parameter $\theta_{z_i z_j}$ that depends only the membership of nodes i and j. In Equation 1 A_{ij} is conditionally independent given z and θ. Number of latent groups, K, is a free parameter that must be chosen before the model and it controls model's complexity. The model can be also expressed as an exponential family:

$$\mathscr{L}(A|z,\theta) \propto \exp\left(\sum_{ij} T(A_{ij}) \cdot \eta(\theta_{z_i z_j})\right) \tag{2}$$

where $T(x) = (x,1)$ is the vector-valued function of sufficient statistics of the Bernoulli random variable and $\eta(x) = (\log(x/(1-x)), \log(1-x))$ is the vector-valued function of natural parameters.

This is a basic and classical SBM for unweighted networks since, as they are defined, the functions (T,η) produces binary edge values. With a different and ap-

propriate functions of (T,η), a specific form of Weighted Stochastic Block Model (WSBM) can be established by weights that are drawn from an exponential family distribution over the domain of T. In this case, each $\theta_{z_i z_j}$ denotes the parameters governing the weight distribution of the edge bundle $(z_i z_j)$.
However, the models SBM and WSBM that presented above, produce complete graphs. In order to model sparse networks by SBM and WSBM, [Aicher et al., 2014] assumes $A_{ij} = 0$ as a directed edge from node i to j is existed with zero weight. In this case, to denote absence of an edge from node i to j is $A_{ij} = \text{NaN}$. By this, sparse networks can be modelled with two types of information, edge existence and edge weight values, in together by a simple tuning parameter with the following way:

$$\log \mathscr{L}(A|z,\theta) = \alpha \left(\sum_{ij \in E} T_e(A_{ij}) \cdot \eta_e(\theta^{(e)}_{z_i z_j}) \right) + (1-\alpha) \left(\sum_{ij \in W} T_w(A_{ij}) \cdot \eta_e(\theta^{(w)}_{z_i z_j}) \right) \quad (3)$$

where the pair (T_e, η_e) denotes the family of edge existence distribution and the pair (T_w, η_w) denotes the family of edge-weight distribution, $\alpha \in [0,1]$ is a simple tuning parameter that combines their contributions in the likelihood function. E is the set of observed interactions (including non-edges) and W is the set of weighted edges with $W \subset E$.
It is possible to see in Equation 3 that if $\alpha = 1$ then the model reduces to SBM, to Equation 1, and if $\alpha = 0$ the model ignores edge existence information then we call such models as pure WSBM (pWSBM). When $0 < \alpha < 1$, the likelihood function combines both information set, i.e. it is called balanced WSBM if $\alpha = 0.5$.[1]

3 An application to the European financial market

We provide an application of the model and analyse the European financial institutions using the closing price series from Worldscope lists downloaded in Datastream at a daily frequency from 29th May 1997 to 27th May 1998. The network is estimated through Granger causality tests [Billio et al., 2012]. In SBM ($\alpha = 1$), Figure 1, the most systematically important community is Block 3 where the members have the highest out degree. However, there are not particular edge weight characteristics in SBM ($\alpha = 1$) communities. In balanced WSBM ($\alpha = 0.5$) the most important community is Block 2.

Acknowledgement

The author acknowledges financial support from the Marie Skłodowska-Curie Actions, European Union, Seventh Framework Program HORIZON 2020 under REA

[1] See [Aicher et al., 2014] for more detailed information.

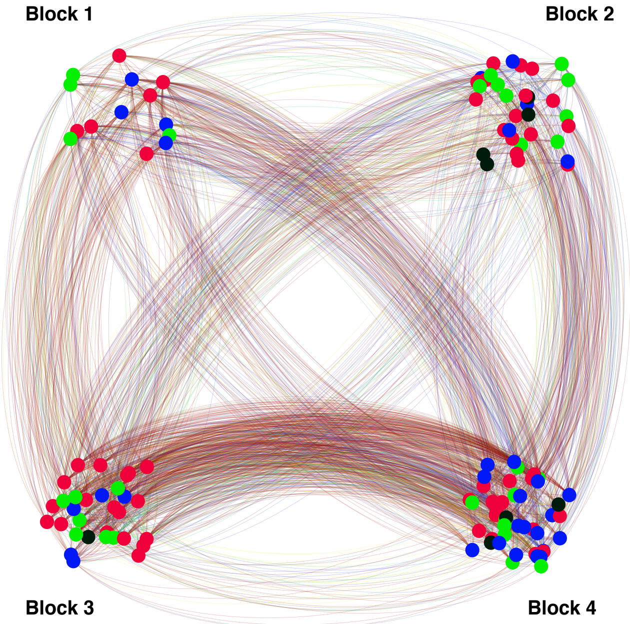


Fig. 1: European Financial Institution Network. Red nodes are banks, blue nodes are insurance companies, green nodes are companies in Financial Service sector and black nodes are Real Estate Companies.

grant agreement n.707070. He also gratefully acknowledges research support from the Research Center SAFE, funded by the State of Hessen initiative for research LOEWE.

References

[Aicher et al., 2014] Aicher, C., Jacobs, A. Z., and Clauset, A. (2014). Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248.

[Billio et al., 2012] Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559.

[Holland et al., 1983] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

[Newman, 2010] Newman, M. (2010). *Networks: an introduction*. Oxford University Press.

Tourism & Cultural Participation

The determinants of tourism destination competitiveness in 2006–2016: a partial least square path modelling approach

I determinanti della competitività di destinazioni turistiche tra 2006 e 2016: un approccio partial least squares a equazioni strutturali

Alessandro Magrini, Laura Grassini

Abstract The present research addresses the analysis of tourism destination competitiveness (TDC) at national level in the period 2006–2016. A partial least square path model (PLS-PM) is developed where TDC and its determinants vary through time according to a second-degree polynomial trend, while their relationships remain constant, thus allowing to draw conclusions on their long-term association. Results show that the most important TDC determinants are cultural heritage, communication technology and tourism infrastructure.
Abstract *Questa ricerca ha l'obiettivo di analizzare la competitività di destinazioni turistiche a livello nazionale nel periodo 2006–2016. È stato sviluppato un modello PLS a equazioni strutturali dove la competitività e i suoi determinanti variano nel tempo secondo un trend quadratico mentre le loro relazioni rimangono costanti, così da poter trarre conclusioni sulla loro associazione a lungo termine. I risultati mostrano che i principali determinanti della competitività sono le risorse culturali, le tecnologie di comunicazione e le infrastrutture turistiche.*

Key words: country-level, formative constructs, PLS, structural equation models, time series.

1 Introduction

As the tourism and travel sector has become an important driver of the contemporary economy, contributing significantly to social, technological and economic develop-

Alessandro Magrini
Dep. Statistics, Computer Science, Applications – University of Florence, Italy
e-mail: alessandro.magrini@unifi.it

Laura Grassini
Dep. Statistics, Computer Science, Applications – University of Florence, Italy
e-mail: laura.grassini@unifi.it

ment [6], the ongoing study of tourism destination competitiveness (TDC) has acquired increasing importance for tourism researchers and policy makers [12]. Over the last decade, several conceptual models for TDC have been proposed [4, 8, 6], and structural equation models (SEMs, see for example [9]) have proved to be a powerful methodology for TDC analysis [11, 10, 2, 7, 1, 16]. The main advantage of SEMs relies in the opportunity to estimate the weights of each indicator and each determinant of competitiveness from data, overcoming the great limitation of constant weights underlying the Tourism and Travel Competitiveness Index [17]. In recent years, particular attention was paid to partial least squares path models (PLS-PMs, see for example [14]), a non parametric formulation of SEMs with weaker sample size requirements, making no assumptions on the distribution of data, and allowing formative constructs. According to several authors, PLS-PMs have introduced a substantial improvement in the methodology for tourism research compared to parametric SEMs, also called covariance-based (CB) SEMs (see the review in [5]). In particular, formative constructs appears more adequate than reflective ones to represent TDC determinants (see the discussion in [11]). However, existing applications focus on one year at a time, failing to capture the substantive (time-invariant) pattern relating TDC and its determinants.

The present research addresses the analysis of TDC at national level in the period 2006–2016. A PLS-PM is developed where TDC and its determinants vary through time according to a second-degree polynomial trend, while their relationships remain constant, thus allowing to draw conclusions on their long-term association.

This paper is structured as follows. In Section 2, a description of the data and the methodology of the research is provided. In Section 3, results are presented. Section 4 includes the discussion of the contribution.

2 Materials and methods

In this research, TDC is understood in the widely accepted definition suggested by [13, page 2]: "*a destination's ability to increase tourism expenditure, to increasingly attract visitors while providing them with satisfying, memorable experiences, and to do so in a profitable way, while enhancing the well-being of destination residents and preserving the natural capital of the destination for future generations*".

The following TDC determinants are considered: core resources and attractiveness (CRA), communication technology (ICT), tourism infrastructure (TOU), demand conditions (DEM).

Data on 20 indicators covering TDC and its determinants (see Table 1 for a description) were gathered from several sources, mostly the World Bank, the World Tourism and Travel Council and the World Economic Forum. The data referred to 264 countries in the period 2006–2016. Destinations were selected in a two-stage procedure. Firstly, the ones with a surface area less than 2000 squared kilometers and less than one million population were merged with a contiguous one, if possible, or excluded. Secondly, the remaining destinations were selected to obtain a dataset with no more than 15% of missing values, as suggested by [15]. The selection procedure led to a total of 130 tourism destinations with a total percentage of missing values of 14%.

A deterministic trend across destinations was taken into account to impute missing values. Due to the limited length of the time series, a total of 15 geographic zones were defined based on physical proximity and economic similarities between the destinations, and each missing datum was replaced with its conditional mean predicted by a linear regression model with destination-specific intercept and geographic zone-specific (instead of destination-specific) second-degree polynomial trend.

The PLS-PM consisted of three parts: a *formative part*, representing the relationships between each TDC determinant and the respective indicators, with the latter determining the former; a *reflective part*, representing the relationships between TDC and the outcomes of tourism activity, with the former determining the latter; and a *structural part* representing the relationship between TDC and its determinants, with the latter determining the former. Destination-specific intercepts and geographic zone-specific second-degree polynomial trends were assumed for each construct.

Let j, k and t indicate the destination, the geographic zone and the year, respectively. The formulation of the PLS-PM was the following:

$$\begin{aligned}
\mathrm{E}[\mathrm{CRA}^{(j,k,t)}] &= \delta^{(j)}_{\mathrm{CRA}} + \gamma^{(k)}_{\mathrm{CRA}} t + \nu^{(k)}_{\mathrm{CRA}} t^2 + \sum_s \lambda_{\mathrm{CRA},s} \cdot X^{(j,k,t)}_{\mathrm{CRA},s} \\
\mathrm{E}[\mathrm{ICT}^{(j,k,t)}] &= \delta^{(j)}_{\mathrm{ICT}} + \gamma^{(k)}_{\mathrm{ICT}} t + \nu^{(k)}_{\mathrm{ICT}} t^2 + \sum_s \lambda_{\mathrm{ICT},s} \cdot X^{(j,k,t)}_{\mathrm{ICT},s} \\
\mathrm{E}[\mathrm{TOU}^{(j,k,t)}] &= \delta^{(j)}_{\mathrm{TOU}} + \gamma^{(k)}_{\mathrm{TOU}} t + \nu^{(k)}_{\mathrm{TOU}} t^2 + \sum_s \lambda_{\mathrm{TOU},s} \cdot X^{(j,k,t)}_{\mathrm{TOU},s} \\
\mathrm{E}[\mathrm{DEM}^{(j,k,t)}] &= \delta^{(j)}_{\mathrm{DEM}} + \gamma^{(k)}_{\mathrm{DEM}} t + \nu^{(k)}_{\mathrm{DEM}} t^2 + \sum_s \lambda_{\mathrm{DEM},s} \cdot X^{(j,k,t)}_{\mathrm{DEM},s} \\
\mathrm{E}[\mathrm{TDC}^{(j,k,t)}] &= \delta^{(j)}_{\mathrm{TDC}} + \gamma^{(k)}_{\mathrm{TDC}} t + \nu^{(k)}_{\mathrm{TDC}} t^2 + \beta_{\mathrm{CRA}} \cdot \mathrm{CRA}^{(j,k,t)} + \\
&\quad + \beta_{\mathrm{ICT}} \cdot \mathrm{ICT}^{(j,k,t)} + \beta_{\mathrm{TOU}} \cdot \mathrm{TOU}^{(j,k,t)} + \beta_{\mathrm{DEM}} \cdot \mathrm{DEM}^{(j,k,t)} \\
\mathrm{E}[X^{(j,k,t)}_{\mathrm{TDC},s}] &= \alpha_{\mathrm{TDC},s} + \lambda_{\mathrm{TDC},s} \cdot \mathrm{TDC}^{(j,k,t)}
\end{aligned} \tag{1}$$

where parameters denoted with letter α, δ, γ, ν, λ and β represent the destination-free intercepts, the destination-specific intercepts, the zone-specific linear trend components, the zone-specific quadratic trend components, the factor loadings and the path coefficients, respectively. The regression with the greatest number of predictors in the PLS-PM has 9 observations per parameter, which allow to detect correlations with absolute value 0.6 with a power of 0.5 [3].

3 Results

Results of PLS-PM estimation are shown in Table 1. The model explains 64% of data variability.

An overall evaluation of the performance of the considered destinations in the period 2006–2016 can be provided by the mean across years of estimated TDC ranks (Table 2). Iceland, with an average rank equal to 1, results the most competitive destination throughout the whole decade. Overall, North and South-West European destinations are the best performing ones (they all appear within the first 25 positions), together with Qatar (17th), Cyprus (19th), United Arab Emirates (23th) and Lebanon (24th).

Table 1 Results of PLS-PM estimation.

Formative part

Indicator	Construct	Std. loading	Variance
Protected areas (% surface area)	CRA	0.1297	1.7%
Number of natural world heritage sites to population	CRA	0.2535	6.4%
Number of cultural world heritage sites to population	CRA	0.8947	80.0%
Number of art museums ($>$ 8000 m^2) to population	CRA	0.5580	31.1%
Number of mobile cellular subscriptions to population	ICT	0.6016	36.2%
Number of individuals using the Internet to population	ICT	0.9837	96.8%
Number of fixed broadband subscriptions to population	ICT	0.9580	91.8%
Number of aircraft departures to population	TOU	0.5089	25.9%
Number of airports to surface area	TOU	0.5814	33.8%
Scheduled available seat kilometers per week	TOU	0.2186	4.8%
Number of hotel rooms to population	TOU	0.2041	4.2%
Number of automated teller machines to adult population	TOU	0.6961	48.5%
Presence of seven major car rental companies	TOU	0.8748	76.5%
Power purchasing parity	DEM	0.5958	35.5%
Consumer price annual inflation	DEM	0.8852	78.4%

Reflective part

Indicator	Construct	Std. loading	Variance
Number of international arrivals to population	TDC	0.8782	77.1%
International tourism receipts to population	TDC	0.9534	90.9%
International tourism expenditure to population	TDC	0.7747	60.0%
Tourism's direct contribution to employment (share)	TDC	0.4716	22.2%
Tourism's direct contribution to GDP (share)	TDC	0.1039	1.1%

Structural part

Path	Std. path coefficient			
	Estimate	Std. error	t-statistic	p-value
CRA ⟶ TDC	0.1287	0.0289	4.4480	0.0000
ICT ⟶ TDC	0.1001	0.0402	2.4876	0.0130
TOU ⟶ TDC	0.1057	0.0214	4.9358	0.0000
DEM ⟶ TDC	−0.0084	0.0096	−0.8703	0.3843

Table 2 Best 25 destinations with respect to the TDC rank averaged in the considered period (2006–2016).

Iceland	1.00	Finland	10.64	Qatar	17.27
Norway	2.64	Greece	11.73	Netherlands	18.27
Denmark	4.18	Latvia	12.18	Cyprus	19.00
Estonia	5.36	Belgium+Luxembourg	13.00	United Kingdom	19.55
Sweden	7.45	Spain	14.73	Germany	20.64
Austria	8.00	Lithuania	15.18	Italy	22.09
Ireland	9.45	France	15.73	United Arab Emirates	22.27
Switzerland	10.45	Portugal	16.73	Lebanon	25.00

4 Discussion

The present research addresses the analysis of tourism destination competitiveness (TDC) at national level in the period 2006–2016. through a partial least square path modelling approach. Differently from existing applications which focus on one year at a time, our contribution is based on time series data and is able to capture the substantive (time-invariant) pattern relating TDC and its determinants.

The main limitation of the present research is represented by the difficulty to find long and almost complete time series. This issue forced us to select 130 on an original number of 260 tourism destinations, and to impute a number of missing values corresponding to almost 15% of total data. The limited length of our time series also precluded a reliable estimation of destination-specific trends, thus we assumed trends to be equal within 15 geographic zones. Being aware that the choice of the geographic zones may significantly affect the results, particular attention was paid to define them so that each included countries with as homogeneous economic characteristics as possible. We hope that future data collection may lead to long enough time series to specify country-specific trends.

The selection of the indicators is a further critical step of our research. In the present contribution, we focused on a limited set of TDC determinants. Future work could consider a broader set of TDC determinants, like public expenditure for the tourism sector, regulation and social aspects.

References

[1] S. Alves and A. R. Nogueira. Towards a Sustainable Tourism Competitiveness Measurement Model for Municipalities: Brazilian Empirical Evidence. *Pasos*, 13(6):1337–1353, 2015.

[2] G. Assaker, R. Hallak, V. Vinzi, and P. O'Connor. An Empirical Operationalization of Countries' Destination Competitiveness Using Partial Least Squares

Modeling. *Journal of Travel Research*, 53(1):26–43, 2014.

[3] W. W. Chin and P. R. Newsted. Structural Equation Modeling Analysis with Small Samples Using Partial Least Square. In R. H. Hoyle, editor, *Statistical Strategies for Small Sample Research*, pages 307–341. Sage Publications, Thousand Oaks, US-CA, 1999.

[4] G. I. Crouch and J. R. B. Ritchie. Tourism, Competitiveness, and Societal Prosperity. *Journal of Business Research*, 44(3):137–152, 1999.

[5] P. O. do Valle and G. Assaker. Using Partial Least Squares Structural Equation Modeling in Tourism Research: A Review of Past Research and Recommendations for Future Applications. *Journal of Travel Research*, 55(6):695–708, 2016.

[6] L. Dwyer and C. Kim. Destination Competitiveness: Determinants and Indicators. *Current Issues in Tourism*, 6:369–414, 2003.

[7] C. Estevao, J. Ferreira, and S. Nunes. Determinants of Tourism Destination Competitiveness: A SEM Approach. *Advances in Culture, Tourism and Hospitality Research*, 10: Marketing Places and Spaces, 2015.

[8] S. S. Hassan. Determinants of Market Competitiveness in an Environmentally Sustainable Tourism Industry. *Journal of Travel Research*, 38(3):239–245, 2000.

[9] R. H. Hoyle. *Handbook of Structural Equation Modeling*. Guilford Press, New York, US-NY, 2012.

[10] J. A. Mazanec and A. Ring. Tourism Destination Competitiveness: Second Thoughts on the World Economic Forum Reports. *Tourism Economics*, 17(4): 725–751, 2011.

[11] J. A. Mazanec, K. Wober, and A. H. Zins. Tourism Destination Competitiveness: From Definition to Explanation? *Journal of Travel Research*, 46:86–95, 2007.

[12] K. Namhyun. Tourism Destination Competitiveness, Globalization, and Strategic Development from a Development Economics Perspective. PhD thesis, 2012.

[13] J. R. B. Ritchie and G. I. Croutch. *The Competitive Destination: A Sustainable Tourism Perspective*. Oxford University Press, Oxford, UK, 2005.

[14] V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang. *Handbook of Partial Least Squares: Concepts, Methods and Applications*. Springer-Verlag, Berlin, DE, 2010.

[15] C. Weismayer. A Replacement Strategy for Missing Values in Destination Competitiveness Data. Diploma Thesis, University of Economics and Business Administration, Vienna, AT, 2006.

[16] H. G. Weldearegay. The Determinants of Tourism Destination Competitiveness: PLS Path Model of Structural Equation Modeling. *Journal of Tourism & Hospitality*, 6(5), 2017.

[17] World Economic Forum. The Travel & Tourism Competitiveness Report. Geneva, CH, 2007.

Participation in tourism of Italian residents in the years of the economic recession

Viaggi e vacanze degli Italiani negli anni della recessione economica

Chiara Bocci, Laura Grassini, Emilia Rocco

Abstract In this study an hurdle model is used to analyze the tourism behavior of Italian residents during the 2004-2013 period. Using the microdata from the quarterly survey on “Trips and holidays in Italy and abroad” carried out by the Italian National Institute of Statistics we investigate the factors that have influenced the tourism participation and the length of stay of residents in Italy in the years of the economic recession. The empirical results show that socio-economic characteristics of the individuals and of their families have an important effect on their tourism participation; that these factors, together with some trips-related characteristics, affect the total number of overnight stays; and that the economic recession impacted negatively on both aspects of tourism behaviour.

Abstract *In questo studio analizziamo il comportamento turistico dei residenti in Italia durante il periodo 2004-2013 mediante un modello hurdle. Utilizzando i microdati dell’indagine trimestrale “Viaggi e vacanze in Italia e all’estero” svolta dall’Istituto Nazionale di Statistica, esaminiamo i fattori che hanno influenzato la partecipazione turistica e la durata delle vacanze degli Italiani negli anni dell’ultima recessione economica. I risultati empirici mostrano che la scelta di fare o non fare un viaggio/vacanza è influenzata da caratteristiche socio-economiche degli individui e delle loro famiglie e che gli stessi fattori, insieme ad altri inerenti i viaggi stessi, influenzano anche il numero totale di giorni di vacanza. I risultati inoltre, mostrano l’impatto negativo della recessione economica su entrambe le fasi decisionali che caratterizzano le scelte turistiche degli italiani.*

Key words: microdata, hurdle models, truncated-at-zero count data models

Chiara Bocci, Laura Grassini and Emilia Rocco
Dipartimento di Statistica, Informatica, Applicazioni “G. Parenti”, Università degli Studi di Firenze, Viale Morgagni, 59 - 50134 Firenze
e-mail: chiara.bocci@unifi.it - laura.grassini@unifi.it - emilia.rocco@unifi.it

1 Introduction

The steady period of economic crisis has seriously affected Italian households that, from 2008 to 2013, have experienced six consecutive years of decrease in purchasing power (available income in real terms), with a -10.4% overall change between 2007 and 2013 [7]. During this recession period, Italian households have shown a reduction in tourism expenditure and a change in travel behavior, as well. In particular, household expenditure surveys show that the expenditure share devoted to hotels and accommodation facilities passed from 2.8% in 2010 to 2.3% in 2013. Household surveys on travel behavior rend us a even worse picture: the annual decrease in the number of trips of residents was nearly -12% in 2010, -19% in 2013. Only in 2015, for the first time after seven years, there has been an increase (+13.5%).

Since tourism is an important driver of economic development, the analysis of the tourism demand of Italian residents is of extreme importance: 1) from an historical perspective, we can be interested in knowing whether, how and why, they have changed their vacation behavior during the years of the economic crisis; 2) for forecasting purposes, the knowledge of the major determinants of household tourism behavior is of extreme usefulness for policy makers. This contribution is concerned with the tourism behavior of Italian residents in the period covering the last economic recession: it investigates whether and how the tourism participation and the total number of overnight stays of Italians have changed in this period. Data on households' and individuals' travel behavior are derived from the survey on *Trips and holidays in Italy and abroad*, currently carried out by the Italian National Institute of Statistics. In the following: Section 2 provides a brief description of the data and the methodology whereas Section 3 discusses the main findings of the analysis.

2 Data and method

From 1997 to 2013 the household survey on *Trips and holidays in Italy and abroad* was carried out quarterly on a national annual sample of about 14,000 households (about 3,500 per quarter for a total annual of about 32,000 individuals). Since 2014 it has become a focus included in the Survey on *Household expenses*, which is carried out monthly on a national theoretical sample of 28,000 families. Given this change, which has been accompanied by several others in the overall survey design, and considering the adoption of the euro currency occurred in 2002, we have limited our analysis to the years from 2004 to 2013. Each year data are observed for the following periods: January-March, April-June, July-September and October-December. In each quarter and for each individual information on vacation and business trips concluded during the quarter and with at least one overnight stay and some socio-demographic characteristic, are recorded. As we are interested in the analysis of the factors that may influence individual tourism choices, we considered only persons at least 15 years old and only the vacation trips. For each trip, the destination, the length of stay, the motivation (leisure and recreation, visiting friends and relatives, and so on), the type of accommodation and the transportation mode are recorded. Available survey micro-data does not include information about the tourism expenditure and the socioeconomic status of the individuals, with the exception of the individual's occupation. Given the data characteristics and the fact that the reduction in the length of stay is one of the main characteristic of current tourism [1], this study examines whether and how the economic recession has affected the total number of overnight stays in a quarter by modelling it through an hurdle model. The hurdle model [5] is a modified count model in which the two processes generating the zeros and the positives are not constrained to be the same. A binomial probability governs the binary outcome of whether a count variate has a zero or a positive realization. If the realization is positive, the hurdle is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero count data model.

The assumptions of the hurdle model are consistent with the phenomenon under study, in which firstly a person decides whether to have a vacation trip and then, conditionally to a positive decision, he decides the number of overnight stays. Therefore the binary process concerning the decision to have at least a vacation in a given quarter has been modelled through a logit regression model in which covariates at both individual level (age, gender, education, occupation, indicator of at least a business trip in the quarter, residential NUTS1 zone) and family level (size, number of children, number of income recipients included retired members) are included. Then the quarterly number of overnight stays, for those who had at least a vacation, has been modelled through a Truncated Negative Binomial regression model which includes, in addition to the variables involved in the first-stage model, trips-related covariates (number of trips for visiting friends and relatives, number of pleasure trips for specific destination: sea, mountain, historical cities, tours and others; number of free accommodation trips; dummy-indicator of at least a trip abroad, total number of vacation trips). Categorical variables for years and quarters are included in both models, as well. Since we consider the number of income recipients as a proxy for household income and we are interest in evaluating its effect on the decision on tourism participation throughout the years, we include a specific interaction term in the first model. Finally, in the second model we allow for different covariates effects on the number of overnight stays for those who only take long vacancies (more than 3 nights at a time) than the others. For this reason, all covariates in the second model are interacted with the dummy indicator *at least one short vacation*. Formally, let y be the number of overnight stays, and $\mathbf{X}$ and $\mathbf{Z}$ the covariates matrices included in the first and the second model respectively, then, the model is:

- I stage: a logit model for the tourism participation

$$P(y_i = 0|\mathbf{X}_i) = exp(\mathbf{X}_i'\beta_1)/(1+exp(\mathbf{X}_i'\beta_1))$$

- II stage: zero-truncated Negative Binomial model for the number of overnight stays given that it is greater than zero

$$P(y_i = j|y_i > 0, \mathbf{Z}_i) = \frac{P(y_i = j \& y_i > 0|\mathbf{Z}_i)}{P(y_i > 0|\mathbf{Z}_i)} = \frac{P(y_i = j|\mathbf{Z}_i)}{[1-P(y_i > 0|\mathbf{Z}_i)]} = \frac{f_{NB}(j)}{[1-f_{NB}(0)]}$$

3 Results and discussion

Looking at previous empirical studies on tourism demand of individuals/households, which mostly investigate tourism expenditures, tourist's choice is conceived as a multi-stage decision process and three categories of travel determinants are generally identified: economic, socio-demographic and trip-related characteristics: they are influential in predicting both the visitors' intention to visit and their willingness to spend money on vacations [8]. According to [1], the same categories of variables affect also the tourist's choice about the vacation length.

Regarding socio-demographic variables our estimates confirm that gender and age may be considered as a proxy of travel preferences and determine travel motivations [8, 3]. Older age groups tend to be less likely to participate in tourism, but show a higher propensity to spent more time in vacation after the decision is make. Moreover, family size is negatively associated with both stages of the tourist's decision process, which can be due to family budget constraints; nonetheless, people with small children (at most 10 years old) seam to be more likely to participate to tourism and to take longer vacations.

Moving on to consider the effect of economic variables, income is one of the most influential and it positively affects all the stages of a tourist's decision process. Tourism individual behavior is also influenced by the business cycle: in an expansion period, people are more inclined to travel on more expensive trips (with greater length of stay) whilst during an economic slowdown more modest domestic trips are preferred [4, 6]. The empirical results, shown in Table 1, agree with these findings. The number of income recipients in a family, which indirectly reflects the household economic condition, have a positive association with the decision to travel and this effect, as shown in Figure 1(a), is more pronounced in the period of economic crisis. Since this variable is also directly related to the occupational status of each family member, it impacts negatively on the quarterly number of overnight stays due to the possible time constraints deriving from the work activity. This agree with the estimated dual effect of the individual occupational status: a stable occupation provide a secure income, increasing the probability of tourism participation, but it reduce the amount of days spent on vacations for those who only take long trips. Moreover, according with [2, 6], unemployed people are less inclined to travel.

Our findings confirm common knowledge that some types of trips strongly determine their length: having more beach holidays or at least one abroad vacation increase the number of total overnight stays for both long and short trips, whereas other motivations have a positive impact only on the total days of short trips. Analogously, most of the overnight stays in long vacations is taken in the Summer trimester, (Figure 1(c)) when the propensity to participate to tourism is higher (Figure 1(b)). This confirms that seasonality is a universal factor in tourism.

Figures 1(b) and 1(c) show that the economic crisis had a negative impact on both tourism participation and on the length of stay, particularly for those tourists that take only long vacations.

The hanging rootogram in Figure 1(d), which compare predicted and observed values, indicates an overall good fitting of the estimated hurdle model.

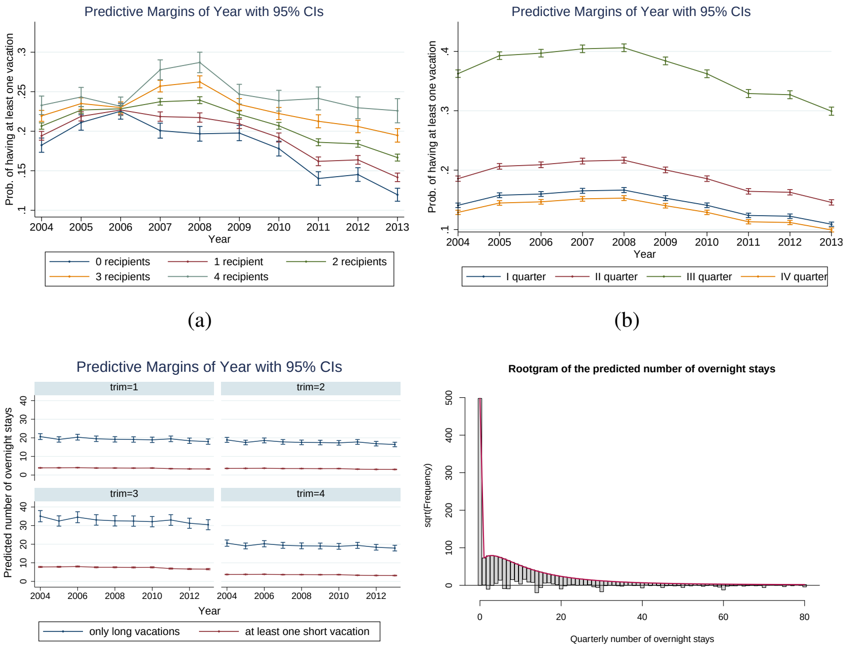
Table 1 Hurdle model estimates

Covariate	Coef.	Covariate	Coef.	Covariate	Coef.
First stage: Tourism participation					
Scaled age	-0.533^{***}	(Scaled age)^2	-0.335^{***}	Female	0.059^{***}
Household size	-0.134^{***}	# of children	0.213^{***}	Univ. degree	0.730^{***}
Business trips	0.424^{***}	OCC:housewife	0.183^{***}	OCC:student	0.912^{***}
OCC:retired	0.442^{***}	OCC:inabile	-0.279^{***}	OCC:managerial staff	0.908^{***}
OCC:office worker	0.612^{***}	OCC:manual worker	-0.063^{*}	OCC:self employed	0.325^{***}
OCC:professional	0.735^{***}	NUTS1:north-east	-0.151^{***}	NUTS1:centre	-0.238^{***}
NUTS1:south	-0.698^{***}	NUTS1:islands	-0.764^{***}	Quarter 2	0.361^{***}
Quarter 3	1.388^{***}	Quarter 4	-0.110^{***}	2005	0.211^{***}
2006	0.310^{***}	2007	0.138^{**}	2008	0.109^{*}
2009	0.116^{*}	2010	-0.034	2011	-0.360^{***}
2012	-0.313^{***}	2013	-0.566^{***}	# income recipients (ir)	0.090^{***}
# ir × 2005	-0.036	# ir × 2006	-0.080^{***}	# ir × 2007	0.036
# ir × 2008	0.057^{*}	# ir × 2009	-0.006	# ir × 2010	0.018
# ir × 2011	0.104^{***}	# ir × 2012	0.073^{**}	# ir × 2013	0.130^{***}
intercept	-1.654***				
Second stage: Quarterly number of overnight stays					
Scaled age	0.160^{***}	(Scaled age)^2	0.067^{***}	Female	0.008
× short vac.	-0.090^{***}	× short vac.	-0.060^{***}	× short vac.	0.020
Household size	-0.017^{***}	# of children	0.064^{***}	Univ. degree	0.057^{***}
× short vac.	-0.021^{**}	× short vac.	0.042^{***}	× short vac.	0.036^{*}
Business trips	-0.069^{***}	OCC:housewife	-0.023	OCC:student	-0.088^{***}
× short vac.	0.080^{**}	× short vac.	0.010	× short vac.	0.218^{***}
OCC:retired	-0.006	OCC:inabile	0.008	OCC:managerial staff	-0.134^{***}
× short vac.	0.057	× short vac.	-0.055	× short vac.	0.177^{***}
OCC:office worker	-0.160^{***}	OCC:manual worker	-0.200^{***}	OCC:self-employed	-0.770^{***}
× short vac.	0.167^{***}	× short vac.	0.054	× short vac.	0.097^{*}
OCC:professional	-0.156^{***}	NUTS1:north-east	-0.080^{***}	NUTS1:centre	-0.062^{***}
× short vac.	0.223^{***}	× short vac.	-0.009	× short vac.	-0.005
NUTS1:south	-0.076^{***}	NUTS1:islands	-0.034^{*}	Quarter 2	-0.036^{**}
× short vac.	-0.040^{*}	× short vac.	-0.120^{***}	× short vac.	0.008
Quarter 3	0.399^{***}	Quarter 4	-0.067^{***}	2005	-0.077^{***}
× short vac.	0.249^{***}	× short vac.	0.024	× short vac.	0.086^{***}
2006	-0.016	2007	-0.059^{***}	2008	-0.074^{***}
× short vac.	0.045°	× short vac.	0.038	× short vac.	0.053^{*}
2009	-0.078^{***}	2010	-0.088^{***}	2011	-0.060^{***}
× short vac.	0.045°	× short vac.	0.063^{*}	× short vac.	-0.060^{*}
2012	-0.115^{***}	2013	-0.140^{***}	# income recipients	-0.029^{***}
× short vac.	-0.040	× short vac.	-0.024	× short vac.	0.011
# family visits	-0.210^{***}	# beach holidays	0.011	# mount. holidays	-0.139^{***}
× short vac.	0.261^{***}	× short vac.	0.091^{***}	× short vac.	0.195^{***}
# visits art towns	-0.324^{***}	# tours	-0.140^{***}	other holidays[a]	-0.104^{***}
× short vac.	0.390^{***}	× short vac.	0.195^{***}	× short vac.	0.155^{***}
# free accom.	0.391^{***}	total trips	0.324^{***}	Abroad	0.120^{***}
× short vac.	-0.441^{***}	× short vac.	0.182^{***}	× short vac.	0.265^{***}
short vacation	-1.781^{***}	intercept	2.052^{***}		

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{\circ}$ $p < 0.1$

Reference levels: Occupation (OCC): unemployed; NUTS1: north-west; Quarter 1; Year 2004

[a] holidays for religious reasons or for health treatments



(a) (b)

(c) (d)

Fig. 1 Results of the hurdle model: (a) Predictive margins of *Year* and of *Number of income recipients* on tourism participation; (b) Predictive margins of *Year* and of *Quarter* on tourism participation; (c) Predictive margins of *Year* and of *Only long vacations* dummy on positive values of the quarterly number of overnight stays; (d) Comparison between predicted and observed values of the quarterly number of overnight stays.

References

1. Alegre, J., Mateo, S., Pou L.: Participation in tourism consumption and the intensity of participation: an analysis of their socio-demographic and economic determinants. Tour. Econ. **15**(3), 531–546 (2009)
2. Alegre, J.,Mateo, S., Pou L.: An analysis of household appraisal of their budget constraints for potential participation in tourism. Tour. Manag., **31**, 45–56 (2010)
3. Bernini, C.,Cracolici, M.F.: Demographic change, tourism expenditure and life cycle behavior. Tourism Manag., **47**, 191–205 (2015)
4. Cafiso, G.,Cellini, R., Cuccia, T.: Do economic crisis lead tourists to closer destinations? Italy at the time of the Great Recession. Papers in Reg. Sci., (2016) doi: 10.111/pirs 12242
5. Cameron, A.C., Trivedi, P.K.: Regression Analysis of Count Data. Cambridge University Press, New York (2013)
6. Eugenio-Martin, J.L., Campos-Soria, J.A.: Economic crisis and tourism expenditure cutback decision. Ann. Tourism Res. **44**, 53–73 (2014)
7. Istat. Rapporto annuale 2014. La situazione del Paese. ISTAT, Rome (2014)
8. Wang, Y., Rompf, P., Severt, D., Peerapatdit, N. Examining and identifying the determinants of travel expenditure patterns. Int. J. Tourism Res., **8**(5), 333–346 (2006)

Cultural Participation in the digital Age in Europe: a multilevel cross-national analysis

La partecipazione culturale nell'era digitale in Europa: un'analisi multilivello transnazionale

Laura Bocci and Isabella Mingo

Abstract Considering a broad spectrum of cultural activities, this study aims to deepening European countries differences in Cultural Participation taking into account micro, meso and macro determinants. For this aim, we used the large dataset provided by the Special Eurobarometer survey n.399 collected in the European countries. The main goals are the following: 1) to outline a multidimensional view of Cultural Participation; 2) to analyse, through a multilevel approach, the determinants of Cultural Participation of European citizens, considering the relationships with both socio-demographic characteristics of people and other relevant contextual features.

Abstract *Considerando un ampio spettro di attività culturali, questo studio mira ad approfondire le differenze della partecipazione culturale nei Paesi europei tenendo conto delle determinanti che possono intervenire, a livello micro, meso e macro. A tal fine, viene utilizzato il dataset dell'Eurobarometro Speciale n.399, rilevato nei paesi europei. Gli obiettivi principali sono i seguenti: 1) delineare una visione multidimensionale della partecipazione culturale nell'era digitale; 2) analizzarne, attraverso un approccio multilivello, le determinanti, considerando sia le relazioni con le caratteristiche socio-demografiche degli individui sia quelle con altre rilevanti caratteristiche di contesto.*

Key words: Cultural Participation, Digital Age, Multilevel analysis.

[1] Laura Bocci, Dipartimento di Comunicazione e Ricerca Sociale, Sapienza Università di Roma; email: laura.bocci@uniroma.it

Isabella Mingo, Dipartimento di Comunicazione e Ricerca Sociale, Sapienza Università di Roma; email: isabella.mingo@uniroma1.it

1 On Cultural Participation

Cultural participation is not only a right enshrined in the United Nations Declaration of Human Rights, but it is also considered a key element for the quality of individual and collective life, as widely recognized by both literature and international Agencies that have contributed to define its concept and setting up of some indicators for its empirical analysis [11,3,1]. At European level, the relevance of this issue is highlighted by the Agenda for Culture adopted in 2007 by the Council of the European Union and the European Council, as well as a number of policy actions set out in the Work Plan for Culture for 2015-2018, adopted by EU Culture Ministers in December 2014.

The concept of Cultural Participation can be made operative on the basis of the main international and European projects: Unesco's Framework for Cultural Statistics in 1986 [10], LEG (Leadership Group on Cultural Statistics) in 2000 [3] and ESSNet-Culture (European Statistical System Network on Culture) in 2011 by Eurostat [1]. These projects, albeit with some differences, agree to adopt a pragmatic definition based on the identification of so-called cultural domains and to include cultural practices that fall into those domains without any distinction in terms of quality and including different types of participation.

Furthermore, these projects emphasize the changes in cultural practices deriving from the rise of ICT (Information and Communication Technology) and especially from the new possibilities offered by the Internet that make cultural participation more complex. Media users have more and more control over the selections of cultural contents offered via different channels, including mobile media. This is the *convergence culture* in which patterns of media use are merging, moving from medium specific content toward content flowing across multiple media channels [6].

In this study, we consider a broad spectrum of cultural activities. The main goal is to deepening European countries differences in Cultural Participation taking into account micro, meso and macro determinants. To this end, the specific objectives are the following:

1) to outline a multidimensional view of cultural participation, considering a broad spectrum of cultural activities;
2) to analyse the determinants of cultural participation of European citizens deepening the direction of the relationships with socio-demographic and other relevant contextual characteristics by using a multilevel approach.

2 Data and Indicators

We analysed the large dataset provided by the Special Eurobarometer survey n.399 collected in the European countries in 2013. This survey detects the attitudes of the European public towards a range of cultural activities, looking at their participation as both consumers and performers of culture. The sample considered in this contribute has a total size of 26,053 individuals aged 15 years and over in 25 european countries[1].

The choice of indicators was driven by the conviction that nowadays Cultural Participation is carried out through traditional activities as well as using the Internet.[2] In the Eurobarometer survey the following 9 variables are dealing with traditional cultural activities: 1- seen a ballet, a dance performance or an opera; 2- been to the cinema; 3- been to the theatre; 4- been to a concert; 5- visited a public library; 6- visited a historical monument or site (palaces, castles, churches, gardens, etc.); 7- visited a museum or gallery; 8- watched or listened to a cultural program on TV or on the radio. Responses modes are: 1- not in the last 12 months; 2- 1-5 times in the last 12 months; 3- more than 5 times in the last 12 months.

Moreover, other variables, which can supplement the previous ones, allow us to take into account some cultural activities performed by the Internet: 1- visiting museum or library websites or other specialized websites ; 2- downloading movies, radio programs (podcasts) or TV programs; 3- watching streamed or on demand movies or TV programs; 4- reading newspaper articles online; 5- downloading music; 6- listening to radio or music; 7- reading or looking at cultural blogs; 8- searching for information on cultural products or events. Responses modes are: 1- not mentioned; 2 - mentioned.

To explain the individual differences of Cultural Participation, other variables were introduced in the analysis. They concern socio-demographic and cultural characteristics of people and some aspects of the community where they live: gender, age, education, occupation, family social class, type of community. Furthermore, since it is reasonable suppose that some features of the country affect significantly Cultural Participation, the following macro level variables were also considered: economic aspects (Gross Disposable Income of Households), political choices (Government Expenditure in Cultural Services), consistency of cultural offer (Employments in Cultural Sector) and level of urbanization (Distribution of Population in the cities)[3].

4 Methods and Results

The previous selected variables were the input for a strategy of analysis that consists in the following two steps: 1) calculate a synthetic index of Cultural Participation at European level; 2) identify the determinants of participation at both individual and country level.

[1] Only countries with a sample size > 1000 were considered. They are: France, Belgium, Netherlands, German, Italy, Denmark, Ireland, Great Britain, Greece, Spain, Portugal, Finland, Sweden, Austria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovakia, Slovenia, Bulgaria, Romania, Croatia.

[2] The set of variables concern only with one form of cultural participation such as attending and receiving. Others two forms (amateur practice and social participation) are not considered in this contribute.

[3] In addition to these variables (source: Eurostat), other ones were also considered in some preliminary analyses, but they resulted not significant: Government expenditure as % GDP in Education, Cultural enterprises % of total of services, Tertiary education (levels 5-8) of the aged population 25-64 (%), Participation rate in education and training (%), Households Broadband coverage (%).

4.1 A Synthetic Index of Cultural Participation at European level

Considering that the variables dealing with cultural activities, including those carried out via the Internet, are all categorical (both nominal and ordinal), Nonlinear Principal Component Analysis (NPCA) was applied to calculate a synthetic index of Cultural Participation.

As it is known, in NPCA, optimal quantification replaces the category labels with category quantifications in such a way that as much as possible of the variance in the quantified variables is accounted for. Specifically, the method maximizes the first p eigenvalues of the correlation matrix of the quantified variables, where p indicates the number of components that are chosen in the analysis. The aim of optimal quantification is to maximize the Variance Accounted For (VAF) in the quantified variables [4,8].

The first dimension (alpha=0.87, VAF=32.12%) has positive correlations (between 0.5 and 0.7) with all variables. The choice of ordinal analysis levels for the variables has been evaluated by examining their transformation plots: for ordinal variables, they indicate that the categories are in the right order and the difference between categories 1 and 2 is slightly larger than that between 2 and 3. For nominal variables, straight lines indicate that they are linearly related to the other variables. The stability of results have been confirmed according 95% bootstrap confidence regions for eigenvalues, component loadings, person scores and category quantifications [7] estimated using balanced bootstrap with 1000 bootstrap samples.

The first component can be interpreted as a Cultural Participation Synthetic index (CPS). The index (ranging from -1.2 to 3.6; mean=0; median=-0.19; standard deviation=1; asymmetry=0.72; Kurtosis=-0.23) assumes different mean values by gender, age, education and country. It is highest among women, decreases as the age increases, it steps up with higher levels of education and is quite different among European countries: its minimum pertains to Portugal (-0.59), the maximum to Sweden (0.9) with different distributions inside each country.

What are the determinants of those different levels of Cultural Participation in the European countries?

4.2 The determinants of Cultural Participation in the European countries: a multilevel approach

The structure of the data to be analysed is clearly hierarchical, since individuals are nested within countries. Given the intrinsically hierarchical nature of the data set and hypothesizing that the variability of the CPS index can depend on both the people characteristics and the different contexts in which they live, a multilevel approach was used [9].

In our case, data consist of the values of CPS index (dependent variable) and several explanatory variables, both social–demographic individual features and countries variables, referred to i-th respondent in j-th country ($i = 1, \dots, N_j$, $j = 1, \dots ,25$ and $\sum_j N_j = 26{,}053$). Therefore, there are two levels of analysis: level two, the highest, is that of countries, and level one, the lowest and nested within the higher level, is that of the individuals.

Since it is reasonable assume that countries can have a systematic effect on the Cultural Participation of individuals, CPS index values within the same countries are dependent or correlated. In this context a multilevel analysis with no explanatory variables at all, the so called *intercept-only model*, was firstly applied

$$Y_{ij} = \square_{00} + u_{0j} + e_{ij}, \qquad i = 1, \dots, N_j \text{ and } j = 1, \dots, 25 \qquad (1)$$

where e_{ij} represents some individual-dependent residual while u_{0j} is a random country-dependent deviation. Model (1) provides a partitioning of the variance between the first (e_{ij}) and second (u_{0j}) level (residual effect and random effect covariance, respectively) and allows us to evaluate the Intraclass Correlation Coefficient (ICC) for the country effect. ICC can be considered both a measure of the between countries variability and the degree of the non-independence of individuals nested into countries. In our case intercepts vary significantly across countries (Wald Z=3.44, p-value=0.001) and ICC=0.136 (Table 1) shows that about 13.6% of the CPS index variability is due to the variability between countries.

Since these previous results justify a multilevel approach [10,5], we performed an analysis in two steps: first, q level-one (individual) X_q explanatory variables were introduced in the multilevel model and then p level-two (country) Z_p explanatory variables were also put in defining the following final model (3)

$$Y_{ij} = \square_{00} + \gamma_{q0}X_{qij} + u_{0j} + e_{ij}, \qquad i = 1, \dots, N_j \text{ and } j = 1, \dots, 25 \qquad (2)$$

$$Y_{ij} = \square_{00} + \gamma_{q0}X_{qij} + \gamma_{0p}Z_{pj} + u_{0j} + e_{ij}, \qquad i = 1, \dots, N_j \text{ and } j = 1, \dots, 25 \qquad (3)$$

In model (2), taking into account only the individual covariates, ICC= 0.113 was still high, even if less than one obtained from model (1), so there still was a fair amount of variation (11.3%) across countries that can be explained by level-two covariates. Actually, in model (3) ICC drastically decreases to 0.031, a value lower than a cut-off (0.05) fixed by most of the researchers [10,5] (Table 1).

The results from model (3), which is the best of the three models according to AICC and BIC (Table 2), show that all predictors are significant. Compared to the reference value, CPS index increases in youngest aged 15-20 people, who are still studying, is self-employed, is living in cities and belongs to high social class (Table 3). Nonetheless, the results show that differences at individual level are not sufficient to explain all the determinants of Cultural Participation. It is necessary to take into account some characteristics of the countries: all the predictors - economic, cultural, and political - considered in model (3) have a positive influence on CPS index, especially the Employment in cultural sector (that is a proxy of cultural offer) and Gross disposable Income of household (Table 3).

Table 1: Residual and Random effects

	Model 1		Model 2		Model 3	
	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.
Residual Effect	0.872	0.000	0.618	0.000	0.618	0.000
Random effect Covariance	0.137	0.001	0.079	0.001	0.020	0.001

ICC	0.136	-	0.113	-	0.031	-

Table 2: Models fit statistics

	Model 1	Model 2	Model 3
Akaike Corrected (AICC)	70507,095	59427,695	59410,96
Bayesian (BIC)	70523,431	59443,957	59427,22

Table 3: Model 3 Fixed effects

Model Term		Coeff.	Std. Error	t	Sig.
Intercept		-1.074	.0706	-15.219	.000
Gender (reference: Female)	Male	-.114	.0190	-5.999	.000
Age (reference: Over 65)	15-29	.297	.0422	7.049	.000
	30-49	.215	.0288	7.480	.000
	50-65	.080	.0206	3.877	.000
Education (reference: No full education)	Still studying	1.299	.0786	16.529	.000
	20 years and older	.974	.0720	13.531	.000
	16-19 years	.422	.0684	6.168	.000
	Up to 15 years	.119	.0627	1.904	.057
Occupation (reference: Not working)	Self-employed	.307	.0326	9.404	.000
	Employed	.218	.0197	11.065	.000
Community (reference: Rural area)	Large town	.261	.0255	10.252	.000
	Small/middle town	.101	.0187	5.382	.000
Family Social Class (reference: Low level)	High level	.331	.0291	11.385	.000
	Middle level	.160	.0142	11.252	.000
Gross Disposable Income		.108	.0218	4.959	.000
Government Expenditure in Cultural Service (%GDP)		.090	.0197	4.566	.000
Employment in cultural sector as % of total employment		.134	.0367	3.650	.000
Distribution of population in the cities (%)		.048	.0222	2.178	.029

References

1. ESSnet Project: *ESSnet Culture Final Report*. Luxembourg (2011).
2. European Commission: Special Eurobarometer 399, Cultural access and participation Report, (2013)
3. Eurostat: *Cultural Statistics in the EU, «Final report of the LeG»* – Population and social conditions. Working papers 3/2000/E/n.1., European Commission, Luxemburg (2000).
4. Gifi, A.: Nonlinear multivariate analysis. John Wiley & Sons, New York (1999).
5. Hox, J.: *Multilevel analysis: Techniques and applications*. Routledge (2010).
6. Jenkins, H.: Convergence Culture: Where Old and New Media Collide. New York University Press, New York, London (2006).
7. Linting, M., & Van der Kooij, A. J. (2012). Nonlinear principal components analysis with CATPCA: A tutorial.Journal of Personality Assessment, 94(1), 12–25.
8. Meulman, J.J., Van Der Kooij, A.J., Heiser, W.J.: Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. The Sage Handbook of Quantitative Methodology for the Social Sciences. Elsevier Science B.V., Amsterdam (2003).
9. Raudenbush, S.W., Bryk, A.S.: Hierarchical linear models: Applications and data analysis methods (Vol. 1). Sage Publications, Thousand Oaks, London, UK (2002).

10. Snijder, T., Bosker, R.: *Multilevel analysis*. *An introduction to basic and advanced modelling*. SAGE Publications (1999).
11. Unesco: *Framework for Culture Statistics*. Paris (1986).

Tourist flows and museum admissions in Italy: an integrated analysis

Flussi turistici e visitatori dei musei in Italia: un analisi integrata

Lorenzo Cavallo, Francesca Petrei, Maria Teresa Santoro

Abstract The main aims of this paper are to analyse cultural tourism in Italy and its relationship with the territory and to provide a tool for policy makers to identify the local dynamics and choose the best development policies. The cultural tourism is studied trough the relation between tourism and cultural statistics.
The analysis uses a spatial approach to identify well-defined geographical areas having significant different - touristic and cultural - characteristics. In order to identify the areas with the greatest cultural-tourism vocation, it is proposed a classification of the municipalities, focusing on some case studies.
Abstract *Il principale obiettivo di questo lavoro è analizzare il turismo culturale e la sua relazione con il territorio, in modo da fornire uno strumento ai policy makers per conoscere le dinamiche del proprio territorio e di intervenire con le adeguate politiche di sviluppo. Attraverso l'analisi spaziale sono individuate le aree geografiche con differenti caratteristiche dal punto di vista turistico e culturale ed è proposta una categorizzazione del territorio italiano in base a una più o meno accentuata vocazione turistica o culturale.*

Key words: cultural tourism, heritage tourism, tourism flows, museums, visitors

1 Introduction

The most visible contribution of culture to local development lies in its ability to attract tourists and the consequent positive effects on spending, incomes and employment (Streeten P., 2006. Leslie D., Sigala M., 2005).

This viewpoint has inspired many studies since the early 1980s that have attempted to demonstrate the impact of culture on local development (OECD, 2005).

[1] Lorenzo Cavallo, Istat; email: cavallo@istat.it; Francesca; Petrei, Istat; email: petrei@istat.it; Maria Teresa Santoro, Istat; masantor@istat.it.

Cultural tourism is one of the largest and fastest growing global tourism markets (OECD, 2009), but what exactly is cultural tourism and how can we measure it? One important obstacle to supplying answers to these questions is both the lack of a standard definition and of data on cultural tourism (Noonan D.S., Rizzo I., 2017). The identification of cultural tourism as a growth market has based more on assertion than hard information and more on isolated observations than systematic analysis. It has therefore been difficult to demonstrate just what cultural tourism is, how important is, how fast it is growing, or to identify the reasons why it has grown (Patuelli R., Mussoni M., Candela G., 2013) (Borowiecki K.J, Castiglione C., 2012) (Cuccia T., Cellini R., 2007 - 2013).

Keeping all that in mind, the main aims of this paper are to analyse cultural tourism in Italy and its relationship with the territory and to provide a tool to policy makers to know the territorial dynamics and identify the right development policies.

2 Sources and methodology

Giving a quantitative measurement of cultural tourism is even more difficult and, considering the difficulty in defining cultural tourism in a one simple way that covers the whole concept, the most practical approach, therefore, seems to be to choose the definition that is most suited to the task at hand (ATLAS, 2005).

In this paper we decided to start with two assumptions: to use only *official statistics* as source, in order to consider data with high and uniform standards of quality; to have the territory as fundamental element for the analysis, because culture, tourism and local development are strongly interconnected with the territory.

2.1 Data

Regarding culture, data on museums and their admissions can be considered, in Italy, as a proxy of the cultural heritage of the territory and are available at a very high territorial detail (at municipality level)[2]. Regarding tourism, we chose official statistics on the capacity of accommodation establishments[3] and the occupancy of accommodation establishments[4], which are available at municipal level.

2.2 Spatial models and zoning operation

[2] http://www.istat.it/it/archivio/6656

[3] http://www.istat.it/it/archivio/14517

[4] http://www.istat.it/it/archivio/15073

Since the early 1950s tourism has been considered as a purely geographical phenomenon and it is proposed an interpretive scheme of the phenomenon, elaborating a spatial model configuration of tourism (Toschi, 1948). Since the late 1960s, the models of interpretation of the spatial structure of tourism have begun to appear in the geographic field and various models are studied (Pearce, 1989), (Zabbini, 2007).

For this analysis, we used a method that well suits the applications supported by GIS systems, that is to divide the territory into homogeneous areas according to some criteria: the zoning operation.

In our analysis, we proceeded in zoning to identify homogenous areas for cultural and/or touristic characteristics trying to give an immediate and clear picture of the Italian territory. Therefore, through a geographical approach, mapping the variables on tourism (arrivals and bed-places) and culture (museums/cultural exhibits and admissions) at Italian municipality level, the analysis allowed us to identify well-defined geographical areas, having significant different tourism and cultural characteristics.

3 The integrated data analysis: classification of the Italian municipalities

For our integrated analysis, we used the final data of the year 2015 on number of establishments, number of arrivals at tourist accommodation establishments (A.E.), number of museums and similar institutions and number of admissions at these institutions. According to the combination of the chosen variables, we classified each municipality in the following categories (table 1):

1. Municipality without A.E. and without museums;
2. Municipality with no. of museums > 0 but without A.E.;
3. Municipality with no. of A.E. > 0 but without museums;
4. Municipality with no. of A.E. and no. of museums > 0.

We attributed the municipalities of this last category to two sub-categories:

4.a Municipality with balanced presence of admissions and arrivals;

4.b Municipality with unbalanced presence of admissions and arrivals.

For this sub-classification, we created an indicator for Italian municipalities[5] based on two variables: arrivals in accommodation establishments and admissions in museums and cultural exhibits.

In summary, for balanced municipalities the two considered flows - admissions and arrivals - are proportionate, highlighting a strong connection between tourism

[5] The first step was to consider the arrivals and admissions for 1,000 inhabitants and standardize the variables. Then we generated a synthetic indicator, ranging from 0 to 1, that, when it presented values around 0.5 (between 0.47 and 0.53, with a range of 5%) the municipality is assigned to the category "4.a", else to the "4.b." (denoting a municipality with an "imbalance" towards one of the two variables).

and culture[6]; for unbalanced municipalities there is a more or less relevant imbalance towards one type of flow.

Table 1: Distribution of the Italian municipalities by the five categories

Category	*N. of Municipalities*	*% on Total*
1	1,005	12.5
2	94	1.2
3	4,608	57.2
4	2,339	29.1
4.a	*652*	*8.1*
4.b	*1,687*	*21.0*
Total	**8,046**	**100**

The municipalities in category 1 are mainly located in the industrial area of the northwest and in the Apennine area of the south; the municipalities in category 2 in the internal areas of the south (including Sardegna) and in the centre-north (see Figure 1). In Toscana, Umbria and Valle d'Aosta, these two categories are completely absent. Municipalities in category 3 are mainly located in the regions of Northern and Southern Italy and, despite being the larger category, there are very few in central Italy.

The balanced municipalities in category 4.a are mainly concentrated in Toscana and Umbria (see Figure 1); then they are along the Adriatic coast of Marche and Emilia-Romagna and in the internal part of Pianura Padana between Bologna and Piacenza. In addition, there are some balanced municipalities in the two provinces of Bolzano and Trento and in southern Italy, along the coast of Puglia and the south coast of Sicilia. Among the balanced municipalities, there are many "provincial capitals", including Milan, which is one of the municipalities with the largest number of arrivals in Italy. In addition, there are some municipalities usually known for non-cultural tourism, such as Rimini (beach tourism) or Merano (winter and thermal tourism), that seem to have clearly realized good local policies to profitably valorize their cultural resources.

The unbalanced municipalities in category 4.b are uniformly spread all over the national territory, with concentrations in the Alpine area and in the internal area of Marche, Lazio, Puglia, Sicilia and Sardegna.

This analysis shows that in Italy, accommodation establishments are widespread on the territory and it is clear that when a municipality has at least one museum, in most cases, there is at least one accommodation establishment. Therefore, the presence of a cultural resource, such as a museum is, really attracts tourists and services for them, encouraging the development of economic activities on the territory.

Moreover, the territorial differences are quite clear: 1) the regional administrative boundaries (Regions have legislative power over tourism matters); 2) the historical disparities among North, Centre and South of Italy and among non-urban areas, metropolitan areas and coastal areas.

[6] The museum offer in these municipalities is an important and "functional" resource for the territory because it actually attracts tourists: these latter not only visit museums but they also spend - nights there. Probably the main motivation of tourists going to these destinations is the culture.

Among the unbalanced municipalities, we define as "municipalities with cultural prevalence (CP)" those having a greater imbalance towards admissions (n. 1,004 municipalities) and as "municipalities with touristic prevalence (TP)" those having imbalance to arrivals (n. 673 municipalities). The unbalanced municipalities are uniformly spread over the territory and more detailed analysis, realized thanks to the knowledge of the territory, reveals important signals (see Figure 2).

Figure 1: Categories of municipalities - Year 2015

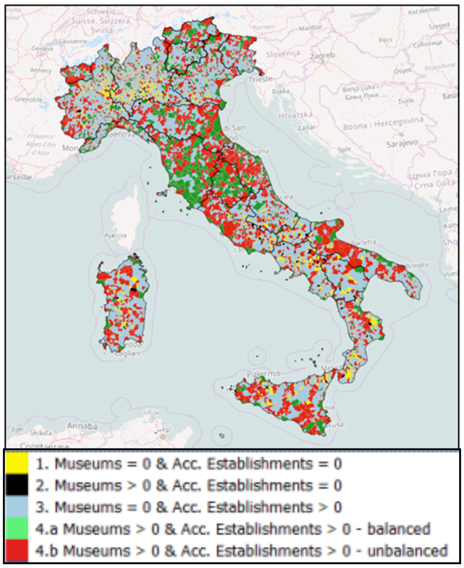
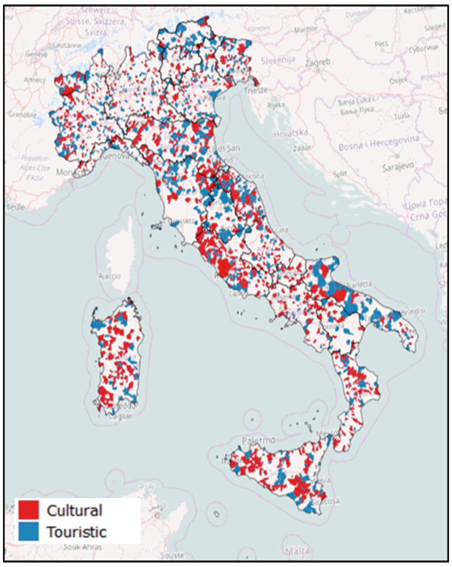

🟨 1. Museums = 0 & Acc. Establishments = 0
⬛ 2. Museums > 0 & Acc. Establishments = 0
🟦 3. Museums = 0 & Acc. Establishments > 0
🟩 4.a Museums > 0 & Acc. Establishments > 0 - balanced
🟥 4.b Museums > 0 & Acc. Establishments > 0 - unbalanced

Figure 2: Unbalanced municipality: classification by prevalence (cultural or touristic) - Year 2015



Among the CP municipalities, there are municipalities well known for their specific and attractive cultural resources: Pompei (with its archaeological park), Taormina (with the Greek theater) or Ferrara and Siena, "art cities" with many museums. In these destinations, the phenomenon of "cultural same-day visits" and quick trips are evident. Furthermore, there are also many important provincial capitals, well known as tourist destinations: Roma, Napoli, Pisa, Firenze and Agrigento. These destinations record large numbers of arrivals (they are among the most relevant municipalities from a tourist point of view) but also even a greater presence of same-day visitors for cultural purposes which do not spend nights in accommodation establishments.

Among the TP municipalities, there are all those destinations that have other resources attracting tourists, as well as cultural ones, like as the main destinations for winter activities, thermal, lakeside or beach tourism. In these municipalities there are cultural resources but they are not the main reason for choosing these destinations. In this category, there are also many important municipalities for number of arrivals, as Venice, a destination mainly visited for cultural reasons.

Table 2: Summary of key themes

	Summary of key themes
1	museums and accommodation establishments are widely spread over the Italian territory, although they reveal considerable density in some specific areas
2	arrivals and admissions are highly concentrated in few regions that collect more than half of both flows
3	part of the structural resources of the country, both touristic and cultural, are underused
4	the balanced municipalities are a minority, but they could be an example to follow
5	the unbalanced CP municipalities use well their cultural resources but they have to change to a more structured tourism that leads to greater revenue and local development
6	the unbalanced TP municipalities have big results in tourism sector but could increase more by the development of their cultural resources

4 Final remarks and future outlooks

The main aim of this analysis is to create a tool for policy makers and territorial stakeholders in order to provide them quality statistical information and objective suggestions to make decisions. This is at the initial stage with the intention to develop it using other data to better qualify all the aspects of the territory (e.g. urbanization, coastal/non-coastal, inner areas, parks, etc.). In our opinion, it is essential to start from the territory, in order to define the right local development policies; moreover, a coordinated and efficient governance system is essential in order to plan the actions and follow their developments.

References

1. ATLAS: Cultural tourism in Europe. Wallingford, (2005).
2. Borowiecki, K. J., Castiglione, C.: Cultural participation and tourism flows: an empirical investigation of Italian provinces. In: Tourism Economics, pp. 241-262, 20(2), (2012).
3. Cellini, R., Cuccia, T.: Is cultural Heritage really important for tourists: a contingent rating study. In: Applied Economics, pp. 261-271, n. 39 (2), (2007).
4. Cellini, R., Cuccia, T.: Museum and monument attendance and tourism flow: a time series analysis approach. In: Applied Economics, pp. 3473-3482, n. 45, (2013).
5. Leslie, D., Sigala, M.: International Cultural Tourism: management, implications and cases. Elsevier Ltd., (2005).
6. Noonan, D.S., Rizzo, I.: Economics of cultural tourism: issues and perspectives. In: Journal of Cultural Economics, Issue 2, pp 95–107, Vol. 41, (2017).
7. OECD: Culture and Local Development. Paris, (2005).
8. OECD: The Impact of Culture on Tourism. Paris, (2009).
9. Patuelli, R., Mussoni, M., Candela, G.: The effects of world heritage sites on domestic tourism: A spatial interaction model for Italy. In: Journal of geographical system, pp. 369-402, n. 15, (2013).
10. Pearce, D.: Turismo oggi. Ulisse Edizioni, Torino, (1989).
11. Streeten, P.: Culture and economic development. In: Ginsburgh, V., Throsby, D. (eds.): Handbook of economics of art and culture, pp. 399- 412, Elsevier, Amsterdam, (2006).
12. Toschi, U.: Corso di Geografia economica generale. Macrì, Firenze-Bari, (1948).
13. Zabbini, E.: Modelli spaziali dell'evoluzione dei territori turistici. Quaderni - Working Papers DSE 585, marzo, (2007).

Posterior Predictive Assessment for Item Response Theory Models: A Proposal Based on the Hellinger Distance

Valutazione Predittiva A Posteriori per i Modelli di Item Response Theory: Una Proposta Basata sulla Distanza di Hellinger

Mariagiulia Matteucci and Stefania Mignani

Abstract Bayesian posterior predictive assessment has received considerable attention for investigating specific aspects of fit of item response theory models. In fact, this approach is easy to apply within Markov chain Monte Carlo estimation, it is flexible and free from distributional assumptions. In its classical implementation, the method is based on graphical analysis and the estimation of posterior predictive *p*-values to investigate the degree to which observed data are expected under the model, given a discrepancy measure. In this work, we propose to quantify the distance between the realized and the predictive distributions of the discrepancy measure based on the Hellinger distance. The results show that this measure is able to provide clear recommendations about the investigated aspects of model fit.

Abstract *Lo studio di aspetti specifici dell'adattamento dei modelli di item response theory è stato affrontato di recente con successo in ambito bayesiano usando strumenti della valutazione predittiva a posteriori. Questo approccio infatti è di facile applicazione quando si utilizza il metodo Markov chain Monte Carlo, è flessibile e non dipende da assunzioni distributive. Nella sua implementazione classica, il metodo si basa sull'analisi grafica e sulla stima dei p-value predittivi a posteriori basati su una particolare misura di discrepanza. In questo lavoro, si propone di quantificare la distanza tra la distribuzione realizzata e quella predittiva della misura di discrepanza utilizzando la distanza di Hellinger. I risultati mostrano che questa misura di distanza è in grado di fornire indicazioni chiare circa i particolari aspetti dell'adattamento considerati.*

[1] Mariagiulia Matteucci, University of Bologna; email: m.matteucci@unibo.it
Stefania Mignani, University of Bologna; email: stefania.mignani@unibo.it

Key words: posterior predictive model checks, item response theory models, Hellinger distance.

1 Introduction

In educational and psychological measurement, item response theory (IRT) models (see, e.g., van der Linden and Hambleton, 1997) are commonly used to estimate the characteristics of both the categorical items and the test takers. Several IRT unidimensional and multidimensional models have been proposed to account for different data structures. While unidimensional models assume the presence of a single latent variable underlying the response process, the multidimensional ones allow for multiple abilities. In this setting, the issue of model goodness-of-fit is crucial to investigate both absolute and relative fit.

Due to the increasing model complexity, a considerable amount of literature has been recently focused on Bayesian estimation of IRT models via Markov chain Monte Carlo (MCMC) methods due to its flexibility. Starting from a MCMC output, one possibility for examining model fit is using Bayesian posterior predictive model checks (PPMC; Rubin, 1984). Considerable advantages of the method are that it does not rely on distributional assumptions, and it is relatively easy to implement, given that the entire posterior distribution of all parameters of interest is obtained through MCMC algorithms.

The first proposals on the use of PPMC for IRT models deal with differential item functioning, person fit, fit of unidimensional models and item fit (see, e.g., Sinharay, 2006). Later, there was an increasing interest in checking specifically for the behavior of unidimensional models fitted to potential multidimensional data (see, among others, Sinharay, Johnson, and Stern, 2006; Levy, Mislevy, and Sinharay, 2009; Levy and Svetina, 2011). In these studies, PPMC has been implemented with graphical analyses and the estimation of the posterior predictive *p*-values (PPP-values) to investigate the degree to which observed data are expected under the model, given a discrepancy measure. Moreover, Wu, Yuen, and Leung (2014) proposed the use of relative entropy (RE) within PPMC to quantify the information the realized distribution loses when it is approximated by the predictive distribution.

The aim of this study is to propose the Hellinger distance, based on the Hellinger integral (Hellinger, 1909), to measure the distance between the realized and the predictive distributions. Unlike the relative entropy, the Hellinger distance is symmetric, it does obey the triangle inequality and it goes from zero to one. The use of the Hellinger distance is investigated for detecting the misfit of an IRT unidimensional model when response data are multidimensional with both simulated and real data.

2 Posterior Predictive Assessment of IRT Models

PPMC techniques are based on the comparison of observed data with replicated data generated or predicted by the model by using a number of diagnostic measures that are sensitive to model misfit (Sinharay, Johnson, and Stern, 2006). Substantial differences between the posterior distribution based on observed data and the posterior predictive distribution indicate poor model fit.

Given the data **y**, let p(**y**|**ω**) and p(**ω**) be the likelihood for a model depending on the set of parameters **ω** and the prior distribution for the parameters, respectively. In the IRT context, **ω** consists of the item parameters, person parameters, and trait correlations. To examine the differences between the observed and the replicated data, the latter are drawn from the posterior predictive distribution (PPD) of replicated data $\mathbf{y}^{rep}$

$$p(\boldsymbol{y}^{rep}|\boldsymbol{y}) = \int_{\boldsymbol{\omega}} p(\boldsymbol{y}^{rep}|\boldsymbol{\omega})p(\boldsymbol{\omega}|y)\partial\boldsymbol{\omega}. \qquad (1)$$

From a practical point of view, one should define a suitable discrepancy measure D(·) and compare the posterior distribution of D(**y**,**ω**), based on observed data, to the posterior predictive distribution of D($\mathbf{y}^{rep}$,**ω**). Discrepancy measures should be chosen to capture relevant features of the data and differences among data and the model. As a first step in PPMC, a graphical analysis is conducted to investigate the differences among realized and replicated discrepancy measures. Then, the PPP-value is defined as

$$\text{PPP-value} = p\big(D(\boldsymbol{y}^{rep}, \boldsymbol{\omega}) \geq D(\boldsymbol{y}, \boldsymbol{\omega}|\boldsymbol{y})\big). \qquad (2)$$

The PPP-value is estimated by computing the proportion of MCMC replications which satisfy Equation (2). The PPP-values provide a measure of the degree to which observed data would be expected under the model: values close to 0 or 1 mean that the realized values fall far in the tails of the distribution of the discrepancy measure based on PPD, indicating misfit; conversely, values of approximately 0.5 mean that the realized values fall in the middle of the distribution, indicating good fit. As underlined by Levy, Mislevy, and Sinharay (2009), PPMC has several advantages over traditional techniques. The method is easy to apply and flexible because the reference distribution is built empirically and it does not require regularity conditions or asymptotic results. Moreover, PPMC relies on Bayesian estimation, which is based on the full posterior distribution: compared with maximum likelihood techniques, which are based on a point estimate, the method is able to directly incorporate uncertainty into the estimation. However, using PPMC is not equivalent to conducting a classical hypothesis test, and the method should be treated as a diagnostic tool (Gelman, Meng, and Stern, 1996; Sinharay, Johnson, and Stern, 2006).

The choice of a suitable discrepancy measure is crucial in PPMC. Effective diagnostic measures in checking for unidimensionality or multidimensionality are based on the association or on covariance/correlation among item pairs. In the first group, the Mantel-Haenszel (MH) statistic is based on the odds ratio conditionally to the rest score *s*, i.e., the raw test score obtained by excluding the two items. For each couple of items *j* and *j*’, with *j*, *j*’=1,…,*k*, the MH statistic is defined as

$$\mathrm{MH}_{jj'} = \frac{\sum_s n_{11s} n_{00s}/n_s}{\sum_s n_{10s} n_{01s}/n_s}, \tag{3}$$

where $n_{tt's}$ is the number of subjects with rest score *s* who score *t* on item *j* and *t*′ on item *j*′, with *t*, *t*′=0,1, and n_s is the number of subjects with rest score *s*. In the second group, the model-based covariance (MBC) is defined as follows

$$\mathrm{MBC}_{jj'} = \frac{\sum_{i=1}^{n}(Y_{ij}-E(Y_{ij}))(Y_{ij'}-E(Y_{ij'}))}{n}, \tag{4}$$

where Y_{ij} is the response variable for individual *i* to item *j*, with *i*=1,…,*n* and *j*=1,…,*k*, and E(Y_{ij}) is its expected value depending on the specific IRT model and the estimated parameters.

2.1 The Hellinger Distance for PPMC

While the PPP-value counts the number of replications for which the predictive discrepancy exceeds the realized one, the researcher may be interested in measuring the size of the difference itself. For this reason, Wu, Yuen, and Leung (2014) proposed the use of the relative entropy (RE), also known as Kullback-Leibler divergence or information, to evaluate the magnitude of the differences between the realized and the predictive measures with limited information statistics based on low-order margins. However, the RE is asymmetric and it is not upper bounded so it is difficult to establish proper threshold levels for assessing absolute model fit or making comparisons.

To overcome these limitations, we propose the use of the Hellinger distance which is symmetric, it does obey the triangle inequality and its range is 0-1. Since the Hellinger distance is used to quantify the distance between two probability measures, it can be used to measure the distance between the realized and the predictive distribution within PPMC as follows

$$\mathrm{H}(P,Q) = \sqrt{1 - \int \sqrt{p(D(\boldsymbol{y},\boldsymbol{\omega}))p(D(\boldsymbol{y}^{rep},\boldsymbol{\omega}))}d\boldsymbol{y}d\boldsymbol{\omega}}. \tag{5}$$

The direct calculation of (5) is computationally demanding and it is usually done via MCMC. Specifically, it is calculated by using the normal kernel density estimates to represent the probability density functions of the realized and the predictive discrepancy measures, given the MCMC replications. In order to check for model unidimensionality, we propose the use of the Hellinger distance with the MBC discrepancy measure, which is based on both data and model parameters, to take into account a fit measure for each item pair. A MATLAB code was written by the Authors to implement the proposal.

3 Main Results

A simulation study is conducted to investigate the performance of the PPP-values and the Hellinger distance at detecting the misfit of a unidimensional IRT model when the data structure is multidimensional. Two different multidimensional IRT models are considered, namely the multi-unidimensional and additive models (see Sheng and Wikle, 2009). Within a confirmatory approach, the multi-unidimensional model relates each item response to a single latent variable, by allowing for trait correlations. In the additive model, a further overall latent trait is assumed underlying all item responses. All traits may be correlated as well. The corresponding IRT unidimensional model is built under the assumption of unidimensionality.

In the study, response data for tests with 10 items and 1,000 respondents are simulated. The trait correlations are manipulated. A number of 5,000 MCMC iterations are conducted, where 1,000 are used for PPMC. Finally, 100 replications are done for each simulation condition. The MH statistic and the MBC are used as discrepancy measures, with 45 item pairs to be considered. Given the mean of the PPP-values for each item pair over the replications, the proportion of extreme PPP-values (below 0.05 or above 0.95) is estimated pooling the results on all item pairs. Given the mean of the Hellinger distance for the MBC (MBC-H) for each item pair, some summary descriptive measures are computed by pooling the results on all item pairs.

Data are generated from the multidimensional models with a bidimensional structure and then analysed with the unidimensional approach. Due to space limit, the results are not reported in the paper but only briefly discussed in the following. The results on the PPP-values show that, for both discrepancy measures, the proportion of extreme values is above 0.75 for most conditions suggesting bad fit. In particular, the MH statistic outperforms the MBC. In the cases of strong and very strong trait correlations, data are conceived as unidimensional and, consequently, the proportion of extreme PPP-values decreases. The average MBC-H is estimated above 0.8 for most cases showing bad fit. The results are coherent and easily interpretable, in fact, the more the MBC-H is close to one, the bad the fit is. For strongly correlated traits, the MBC-H is estimated on average around 0.65. This means that the data are conceived as unidimensional but the distance measure is still able to catch the discrepancy between the generating model and the one used to analyse response data.

Data coming from a survey conducted by the University of Bologna (Bernini, Matteucci, and Mignani, 2015) to investigate the residents' perceptions toward tourism in terms of perceived benefits and costs are used. A total of 5 items on benefits and 5 items on costs are administered, suggesting a bidimensional latent structure. The unidimensional, multi-unidimensional and additive models are fitted. The results on the PPP-values show that the unidimensional approach is associated to a proportion of about 80% of extreme values. On the contrary, about 30% and 16% of extreme

PPP-values are reported for the multi-unidimensional and the additive model, respectively. Clearly, the additive model shows the best fit. These results are confirmed by the analysis with the MBC-H. On average, the estimated distances are about 0.8, 0.5, and 0.4 for the unidimensional, multi-unidimensional and additive model, respectively.

The approach based on the Hellinger distance seems to be promising to evaluate model fit within posterior predictive assessment. In particular, all measures could be used to investigate misfit due to specific items. A more comprehensive simulation study is needed to check the performance of the method for different simulation conditions.

References

1. Bernini, C., Matteucci, M., Mignani, S.: Investigating heterogeneity in residents' attitudes toward tourism with an IRT multidimensional approach. Qual. Quant. **49**, 805-826 (2015).
2. Gelman, A., Meng, X.L., Stern, H.S.: Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. **6**, 733-807 (1996).
3. Hellinger, E.: Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik (in German) **136**, 210–271 (1909).
4. Levy, R., Svetina, D.: A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. Br. J. Math. Stat. Psychol. **64**, 208-232 (2011).
5. Levy, R., Mislevy, R.J., Sinharay, S.: Posterior predictive model checking for multidimensionality in item response theory. Appl. Psychol. Meas. **33**, 519-537 (2009).
6. Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applies statistician. Ann. Stat. **12**, 1151-1172 (1984).
7. Sheng. Y., Wikle. C.: Bayesian IRT models incorporating general and specific abilities. Behaviormetrika **36**, 27-48 (2009).
8. Sinharay, S.: Bayesian item fit analysis for unidimensional item response theory models. Posterior predictive assessment of item response theory models. Br. J. Math. Stat. Psychol. **59**, 429-449 (2006).
9. Sinharay, S., Johnson, M.S., Stern, H.S.: Posterior predictive assessment of item response theory models. Appl. Psychol. Meas. **30**, 298-321 (2006).
10. van der Linden, W. J., Hambleton, R.K. Handbook of Modern Item Response Theory. Springer-Verlag, New York (1997).
11. Wu, H., Yuen, K.V., Leung, S.O.: A novel relative entropy-posterior predictive model checking approach with limited information statistics for latent trait models in sparse 2^k contingency tables. Comput. Stat. Data Anal. **79**, 261-276 (2014).

Well-being & Quality of Life

Is Structural Equation Modelling Able to Predict Well-being?

È possibile stimare il livello di benessere per mezzo di modelli ad equazioni strutturali?

Daniele Toninelli and Michela Cameletti[1]

Abstract The well-being (WB) measurement is an important and challenging task. Quality of life is a multifaceted topic, thus its measure cannot rely anymore on one or few indicators only. Data from large-scale survey projects, such as the European Social Survey (ESS), are a solid basis for testing new methods aimed at measuring the phenomenon. We apply Structural Equation Modelling (SEM) to ESS wave 8 data. Our research aims at evaluating if SEM is a reliable method for estimating the WB and the relative importance of its dimensions in some European countries.

Abstract *Misurare il benessere è una sfida metodologica di rilievo. La qualità della vita è un concetto multidimensionale, la cui misura non può basarsi solo su uno o pochi indicatori. La disponibilità di dati da progetti di indagine su larga scala (come l'European Social Survey - ESS), è una solida base per sperimentare nuovi metodi di stima. In questo lavoro si applica la metodologia SEM (Structural Equation Modelling) per stimare il livello di benessere soggettivo rilevato nei Paesi Europei coinvolti nella wave 8 della ESS e l'importanza delle sue componenti.*

Key words: European Social Survey, quality of life, structural equation models, subjective well-being.

1 Introduction: Well-being & Structural Equation Modelling

Measuring the well-being (WB) became one of the key priorities, for national statistical institutes; nevertheless, this task is challenging, due to both the multi-

[1] Daniele Toninelli, University of Bergamo; email: daniele.toninelli@unibg.it
Michela Cameletti, University of Bergamo; email: michela.cameletti@unibg.it

dimensional and the latent nature of the phenomenon. GDP or other macro-economic indicators are not enough anymore (Stiglitz, Sen & Fitoussi, 2010): they are not able to detect all factors that actually affect the citizens' quality of life. Nevertheless, obtaining a reliable measure of the WB is extremely important, mostly in the framework of official statistics. On the one hand, a low level of WB can highlight critical areas politicians should turn investments towards and, on the other hand, observed changes in the WB could be used to evaluate the impact of implemented policies. This is why several national statistical institutes started developing projects aimed at measuring the WB[1].

In this work, we study the subjective WB (see Diener, 2013) with the main objective of testing the capability of SEM in measuring such a concept in different European countries. The estimates rely on several indicators (obtained from groups of items of the ESS questionnaire) that are supposed to cover the main WB dimensions. SEM allows us to evaluate how each dimension is affecting the latent concept of subjective WB. Moreover, we evaluate if SEM is, generally, able to estimate the WB level of the studied countries, making a direct comparison with measures of WB given by two specific items of the ESS questionnaire (our benchmark). Examples of previous SEM applications to WB include Oliver et al. (2009), Lin & Yeh (2014), Turashvili & Turashvili (2015) and Warner & Rasco (2016). These papers are focused on very specific WB aspects and are based on relatively small samples (mostly made by students) and/or on cross-sectional studies. Our approach, instead, is broader and based on a steady large-scale survey that allows us to develop a cross-country study at the European level. Moreover, we focus on a wide set of items that potentially covers all the main WB dimensions.

2 Data & Method

Our work starts from data collected in 2016 among 18 European countries[2] within the framework of the ESS wave 8. The dataset includes 34,836 records, but we estimate the final model on 26,455 units (8,381 units were excluded due to the listwise deletion method used by the Lavaan R package; missing values are mainly observed for sensitive items such as the ones asking about the household income). In particular, we focus on the measurement of subjective WB, because "measures of subjective well-being provide key information about people's quality of life" (Stiglitz, Sen & Fitoussi, 2010, p. 58). We do not focus, as suggested by Diener (1984) on the three main components of subjective WB (life satisfaction, positive experiences and negative experiences). We rather analyse, independently from their positive/negative nature, a set of indicators that could measure and cover as much as possible all WB dimensions. Table 1 shows how we grouped the items of the wave 8

[1] Such as: the "Better live initiative" (OECD: http://www.oecd.org/statistics/measuring-well-being-and-progress.htm), the BES project (Istat: http://www.misuredelbenessere.it/index.php?id=51), or other studies (UK Office for Nat. Stat.: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing).

[2] Austria, Belgium, Switzerland, Czech Rep., Germany, Estonia, Finland, France, United Kingdom, Ireland, Iceland, Israel, Netherlands, Norway, Poland, Russia, Sweden, Slovenia.

ESS questionnaire in order to take into account these dimensions (listed in the first column; the full list of indicators is shown within Figure 1). The second column of Table 1 shows the correspondence to the WB dimensions suggested by Stiglitz, Sen & Fitoussi (2010), referenced as "SSF" (the new dimension we introduced are in bold). The third column shows the number of items taken into account for each dimension. Some couples of dimensions (e.g. *Country attachment and trust* and *Public involvement*) were merged into a common latent dimension (e.g. *atchtrst*) of the SEM model. Religion was excluded because the available variables, even if combined together to obtain two separate items, were too correlated to be included in the SEM model as measurements of a latent religion dimension (and this caused estimation problems). We also avoided to specify a SEM model with WB being directly measure by one religion variable (computed as an average of all the religion variables) without specifying a latent religion construct.

Table 1: *WB dimensions covered by the selected ESS questionnaire items (batt.=battery of items)*

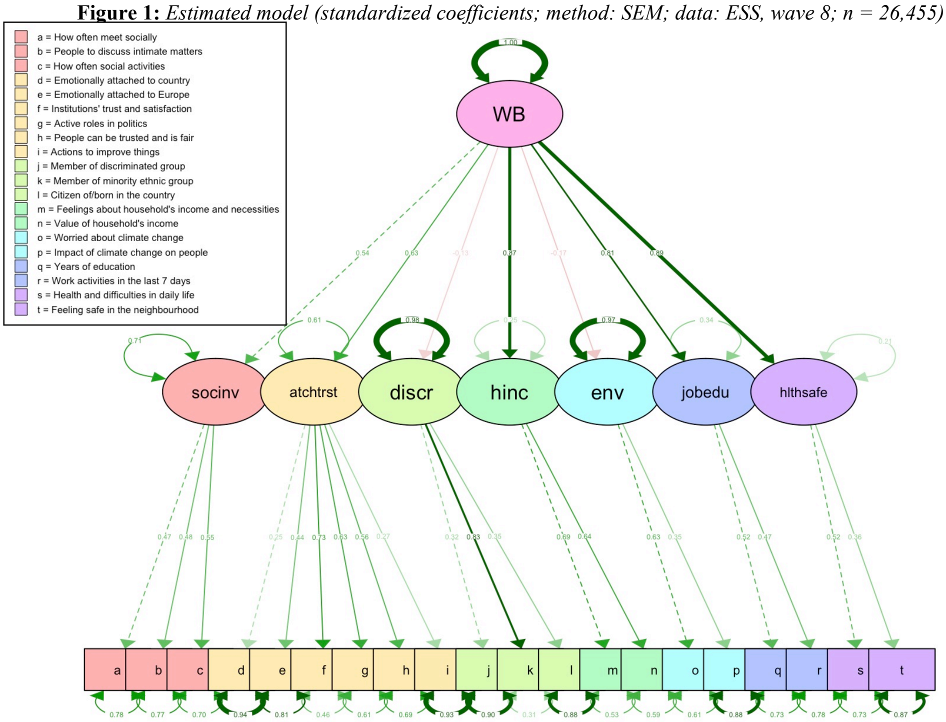
WB dimension *(model name)*	***SSF***	***No. items***
Social involvement (*socinv*)	Social connection	3
Feeling safe (*hlthsafe*)	Insecurity	2
Health conditions (*hlthsafe*)	Health	2
Country attachment & trust (*atchtrst*)	---	4+2 batt.
Public involvement (*atchtrst*)	Political voice	2+1 batt.
Religion	---	4
Discrimination/citizenship (*discr*)	---	4
Environment (*env*)	Environment	2
Income perception (*hinc*)	Material living standards	3
Work status (*jobedu*)	Personal activities / work	1+1 batt.
Education (*jobedu*)	Education	1

We aim at estimating a model based on SEM (see Hox & Bechger, 1999, and Kline, 2011), able to predict the WB level for the studied European countries. For this study, we used the `R lavaan` package (Rosseel, 2012; http://lavaan.ugent.be/). Our objective is twofold. First, we want to understand if and how the WB dimensions differently affect the phenomenon in European countries, studying the standardized coefficients. Second, we evaluate if SEM is a reliable method for estimating the country subjective WB, comparing observed data and SEM estimates. For this second objective, we study three different benchmark, considered by Dolan & Metcalfe (2012) as part of "experience measures" and of "evaluation measures", respectively. These items are: 1) "**Happy**" ("*Taking all things together, how happy would you say you are?*"; 0 to 10 answering scale, with extremes corresponding to "extremely unhappy" and "extremely happy"); 2) "**Satisf**" (*"All things considered, how satisfied are you with your life as a whole nowadays?"*; same scale, with extremes labelled as "extremely dissatisfied" and "extremely satisfied"); 3) "**H-SAvg**": average of the two previous items. For this last variable a high variability was observed, ranging from 5.95 of Russia to 8.16 of Switzerland, with a global (not weighted) average equal to 7.52 (st.dev.=0.565).

3 Results

Applying the SEM on the whole group of countries, we obtained the model shown in Figure 1, where the WB is defined by 7 latent dimensions each composed by a set of at least two indicators or batteries of variables.

Figure 1: *Estimated model (standardized coefficients; method: SEM; data: ESS, wave 8; n = 26,455)*


a = How often meet socially
b = People to discuss intimate matters
c = How often social activities
d = Emotionally attached to country
e = Emotionally attached to Europe
f = Institutions' trust and satisfaction
g = Active roles in politics
h = People can be trusted and is fair
i = Actions to improve things
j = Member of discriminated group
k = Member of minority ethnic group
l = Citizen of/born in the country
m = Feelings about household's income and necessities
n = Value of household's income
o = Worried about climate change
p = Impact of climate change on people
q = Years of education
r = Work activities in the last 7 days
s = Health and difficulties in daily life
t = Feeling safe in the neighbourhood
1.00
WB
0.54
0.63
-0.13
0.87
-0.17
0.81
0.89
0.71
0.61
0.98
0.25
0.97
0.34
0.21
socinv
atchtrst
discr
hinc
env
jobedu
hlthsafe
0.47
0.48
0.55
0.25
0.44
0.73
0.63
0.56
0.27
0.32
0.83
0.35
0.69
0.64
0.63
0.35
0.52
0.47
0.52
0.36
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
0.78
0.77
0.70
0.94
0.81
0.46
0.61
0.69
0.93
0.90
0.31
0.88
0.53
0.59
0.61
0.88
0.73
0.78
0.73
0.87

The model converges after 139 iterations. It does not generally provide an adequate fit ($\chi^2(163) = 16{,}429$, $p<.001$); nevertheless, this is likely to happen with a very large sample (Kline, 2011, p. 209). The model does not perform well also according to the Comparative Fit Index (CFI=.764), indicating a not satisfying improvement in fit over a baseline independence model (the usual threshold is >0.90). RMSEA (=.061) shows a good fit (it should be smaller than .10), instead.

Despite the difficulties of the model in reproducing the empirical covariance matrix, all path coefficients are statistically significant ($p<.001$). According to the completely standardized solution and considering *socinv* as the reference dimension, the dimensions that mostly affect the subjective WB are (standardized loadings within parentheses): *discr* (-0.130), *env* (-0.170); *atchtrst* (0.626); *jobedu* (0.813); *hinc* (0.869); *hlthsafe* (0.888). Thus, both being worried about the climate change and being discriminated negatively affect the perceived WB, whereas an active work status, the household income, the current health conditions and the perceived safety positively affect the WB. To a lower extent, we also note a positive linkage with how much the respondent has an active role in politics or trusts in/is satisfied about public institutions (such as government, police).

Note that the model presented in this paper is not the optimal one: for example, excluding the latent dimensions *env*, *jobedu* and *hlthsafe*, both the CFI (=.830) and the AIC index (=1,582,693 vs 2,158,450 of the full model) show a better fit than using the full model. Nevertheless, we decided to keep the full version, in order to be able to explore the complete list of generally recognized WB dimensions.

In order to check if the SEM works in detecting the subjective WB level, starting from its dimensions, we compare the perceived WB level (measured through the *H-SAvg* variable, third column of Table 2) and the estimates obtained using the SEM (fourth column of Table 2) for all European countries.

Table 2: *Countries rankings: by observed WB (H-SAvg) and by model estimates (WB-SEM)*

Rank H-SAvg	***Country***	***WB H-SAvg***	***Rank Happy***	***Rank Satisf***	***WB SEM***	***Rank WB-SEM***
1	Russian F.	5,949	1	1	-0,194	**1**
2	Czech Rep.	6,803	2	3	-0,100	**2**
3	France	6,887	3	2	-0,094	**4**
4	Estonia	7,172	4	4	-0,080	**5**
5	Slovenia	7,299	5	5	-0,071	**6**
6	Poland	7,401	7	6	-0,065	**7**
7	Ireland	7,404	6	7	-0,049	**8**
8	UK	7,555	8	8	-0,010	**10**
9	Belgium	7,614	10	9	-0,005	**11**
10	Germany	7,644	11	10	0,035	**13**
11	Austria	7,743	9	12	-0,018	**9**
12	Iceland	7,827	13	11	-0,096	**3**
13	Netherlands	7,876	14	13	0,048	**14**
14	Sweden	7,902	12	14	0,110	**17**
15	Norway	8,043	15	15	0,115	**18**
16	Israel	8,107	16	17	0,107	**16**
17	Finland	8,128	18	16	0,034	**12**
18	Switzerland	8,165	17	18	0,061	**15**

The Spearman’s rank correlation coefficient shows that all associations between our model estimates and the WB variables are significant (α=0.01) and high: ρ=.804 for *Happy*; ρ=0.851 for *Satisf*; ρ=0.835 for *H-SAvg*. The Kendall’s tau coefficient confirms these results (with associations equal or higher than 0.686 and significant, for α=0.01). The Pearson correlation coefficients between the WB model estimates by country and the three observed variables are all also quite high and significant (p<0.001): 0.841 for *Happy*; 0.859 for both *Satisf* and *H-SAvg*.

4 Conclusions

SEM is able, taking into account several latent dimensions, to estimate the relative WB level of different European countries and to help in detecting, at an aggregate level, the dimensions actually more important. Nevertheless, the estimated model shows fitting issues. Since we estimate a general model for all the European countries, this could be caused by the diversity of studied contexts: determinants of the subjective WB can be different across countries. However, estimating country-specific models should probably fix this issue. This strategy should also depict more closely the relative importance of different dimensions in affecting subjective WB. Moreover, we did not use any weights in estimating our model, thus estimates could be mainly driven by the most populated countries. This preliminary work was developed in order to understand if SEM is able to estimate WB, on the basis of items measuring different WB dimensions. This is just a first step. For further research, we suggest to use data coming from social networks and/or obtained using web-scraping in order to create indexes useful to update the information made available by large-scale survey projects, such as the ESS.

References

1. Diener, E.: Subjective well-being. Psychological Bulletin *95*, 542--575 (1984)
2. Diener, E.: The remarkable changes in the science of subjective well-being. Perspectives on Psychological Science *8*, 663--666 (2013)
3. Dolan, P., Metcalfe, R.: Measuring subjective wellbeing: recommendations on measures for use by national governments. Journal of social policy *41(2)*, 409--427 (2012)
4. Hox, J.J., Bechger, T.M.: An introduction to structural equation modeling. Family Science Review, *11*, 354—373 (1999)
5. Kline, R.B.: Principles and Practice of Structural Equation Modeling. The Guildford Press, New York (2011)
6. Lin, C.C., Yeh, Y.C.: How gratitude influences well-being: a structural equation modeling approach. Soc Indic Res *118(1)*, 205—217 (2014)
7. Oliver, A., Navarro, E., Meléndez, J.C., Molina, C, Tomás, J.M.: Structural equation model for predicting well-being and functional dependency of the elderly in the Dominican Republic. Rev Panam Salud Publica *26(3)*, 189--196 (2009)
8. Rosseel, Y.: `lavaan`: An `R` Package for Structural Equation Modeling. Journal of Statistical Software *48(2)*, 1--36 (2012)
9. Stiglitz, J.E., Sen, A., Fitoussi, J.-P.: Report by the Commission on the Measurement of Economic Performance and Social Progress. (2010). http://hdl.voced.edu.au/10707/47743. Cited 21 Feb 2018
10. Turashvili, T., Turashvili, M.: Structural equation model of psychological well-being, a Georgian Exploration. Procedia - Social and Behavioral Sciences *190*, 4--9 (2015)
11. Warner, R.M., Rasco, D.: Structural equation models for prediction of subjective well-being: Modeling negative affect as a separate outcome. The Journal of Happiness & Well-Being 2(1), 161-176 (2014)

The well-being in the Italian urban areas: a local geographic variation analysis

Il benessere nelle aree urbane italiane: un'analisi della variabilità locale

Eugenia Nissi and Annalina Sarra

Abstract Following the place-based well-being literature, this paper is aimed at assessing inequality between Italian province capital cities in terms of their performance in promoting human and ecosystem well-being. The case study rely on the theoretical framework adopted by ISTAT within the Ur-Bes project. The available indicators are used to derive a multidimensional urban well-being index. To this end we adopt a two-steps procedure. Firstly, by using the geographically weighted PCA we assess the spatial variability for each Ur-Bes pillar data and obtain for each dimension a composite index. In the second stage, the ranking of the Italian province capital cities according to their efficiency in promoting equitable and sustainable well-being is facilitated by DEA.

Abstract *In questo lavoro ci si propone di valutare il benessere equo-sostenibile delle province italiane. Il caso studio fa riferimento al modello concettuale di benessere adottato dall'ISTAT nell'ambito del progetto Ur-Bes. Gli indicatori disponibili sono utilizzati per ricavare un indice multidimensionale del benessere urbano. A tal fine, viene impiegata una procedura a due fasi. Inizialmente, attraverso la ACP ponderata geograficamente, si valuta la variabilità spaziale per ciascun dominio dell' Ur-Bes e si ottiene per ogni dimensione un indice composito. Nella seconda fase, l'impiego della tecnica DEA consente di ottenere un indice globale di misura del benessere equo-sostenibile e di confrontare l'efficienza delle province italiane nel promuoverlo.*

Key words: well-being, unitary input DEA, GW PCA, efficiency ranking

Eugenia Nissi
University "G.d'Annunzio" of Chieti-Pescara, Department of Economics, e-mail: nissi@unich.it

Annalina Sarra
University "G.d'Annunzio" of Chieti-Pescara, Department of Legal and Social Sciences e-mail: asarra@unich.it

1 Introduction

The last two decades have witnessed a growing interest on the measurement of well-being and quality of life, as documented in many theoretical and empirical studies. Some of these researches focus on the well-being assessment at local level (see, among others, Bai et al., 2012). Basically, the root of understanding local well-being (in regions and cities) lies in the intersection of well-being and public policy. In this respect, more fine-grained measures of well-being will help policy-makers to enhance the design and targeting of policies and improve their capacity to respond to the paramount and varied needs of residents. Well-being is a multidimensional phenomenon, whose definition and theorization requires the specification of a conceptual framework for its assessment at national as well as at local level. The conceptual framework providing grounds for the discussion in this paper has been that adopted by ISTAT within the “Equitable and Sustainable Well-Being" project, whose Italian acronym, used hereafter, is BES (see ISTAT- Cnel, 2012), which, in turn, is based on the conceptual model published by OECD (Hall et al. 2010). This theoretical framework reflects the conceptual complexity of well-being and highlights its dependency upon attributes specific to each person and on attributes shared with other people or revealing the relations between them or how a society is peaceful, resilient, cohesive.

Our case study considers the Italian Province capital cities as units of analysis and employs the urban Bes (Ur-BES) report data, which refers to 64 particular indicators, belonging to 11 dimensions, identified within the equitable and sustainable well-being initiative (BES). The paper sheds light on the construction of a multidimensional urban well-being index for the Italian Province capital cities and, following the place-based well-being literature, on assessing inequality between Italian province capital cities in terms of their performance in promoting human and ecosystem well-being. In our analysis, a special focus is placed on the importance of surveying the spatial dimension of the local well-being indicators and their related variables. Most of the existing literature on the construction of composite indicators neglects to consider the spatial heterogeneity of the units in the computation of their relative composite indicators scores. As matter of fact, it may happen that the value of a composite indicator may be more dependent on a certain sub-indicator in a given location, and another sub-indicator in different location. To ascertain this kind of spatial dependence can reveal useful for policy decision makers in tackling problems in an efficient way, and distinguishing their causes at local level. With regards to this research issue, we propose a two-step approach. Firstly, for each of the well-being dimensions we employ the Geographically Weighted (GW) Principal Component Analysis (PCA). The GW PCA, introduced by Harris et al. 2011, can be deemed a local version of the traditional PCA in that it takes spatial variations across a study region into account and produce maps of spatial variations of each local principal component and local variance at each place. This variant of global PCA is chosen due its merits in assessing the spatial variability of each Ur-BES pillar data dimensionality and checking how the elementary indicators influence the corresponding spatially-varying component. In the second stage of our empirical procedure, the synthesis of Ur-Bes elementary indicators obtained through the GW PCA, is included in a unitary Data Envelopment Analysis (DEA) model to derive a spatial composite index. The employment of DEA facilitates the ranking of the Italian province capital cities according to their efficiency in promoting equitable and sustainable well-being. The rest of paper proceeds as follows. In Section 2 we give some details of the theoretical background of the GW PCA technique, then we present the basic of DEA model, as well as the specific model selected for our case study. The results are discussed in Section 3.

2 Methodological approach

In our two step-procedure we start by reducing the dimensionality of Ur-Bes elementary indicators by using GW PCA. Next, the reduced set of variables is employed in a unitary input DEA model to assess the relative efficiency of the Italian Province capital cities in producing equitable and sustainable wellbeing. The following sub-sections describe both GW PCA and DEA techniques.

2.1 Geographically Weighted PCA

GW PCA is a local spatial form of the PCA able to provide locally derived sets of principal components for each location (Harris et al. 2011). GW PCA adapt PCA for spatial effects with respect to spatial heterogeneity. We assume that the vector of observed well-being variables at location i have a multivariate normal distribution, with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ ($x_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). The mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ are now function of location i, with coordinates (u,v). This implies that each element of the mean vector and the variance matrix is, in turn, function of position and are expressed as $\boldsymbol{\mu}(u,v)$ and $\boldsymbol{\Sigma}(u,v)$, respectively. The geographically weighted principal components are obtained through the decomposition of the geographically weighted variance-covariance matrix:

$$\boldsymbol{\Sigma}(u,v) = \mathbf{X}^T \mathbf{W}(u,v) \tag{1}$$

where $\mathbf{W}(u,v)$ is a diagonal matrix of weights. As for any geographically weighted methods, diverse kernel functions (gaussian, exponential, bi-square) can be employed to generate the diagonal matrix of weights, under the control of a parameter known as bandwidth. Geographically weighted principal components are obtained using the decomposition of the variance-covariance matrix. More specifically, the local principal component at location (u_i, v_i) can be written as:

$$\mathbf{L}(u_i, v_i)\mathbf{V}(u_i, v_i)\mathbf{L}(u_i, v_i)^T = \mathbf{\Sigma}(u_i, v_i) \tag{2}$$

where $\mathbf{L}(u_i, v_i)$ is the matrix of the geographically weighted eigenvectors and $\mathbf{V}(u_i, v_i)$ is the diagonal matrix of the geographically weighted eigenvalues. For p variables, the GW PCA provides p components, p eigenvalues, p set of component loadings and p set of component scores for each data location in the study area. Full technical details underlying GW PCA are described in Harris et al (2011).

2.2 Unitary input DEA model

Data Envelopment Analysis (DEA) is a widely used non-parametric method of measuring the efficiency of organisational units, termed Decision Making Units (DMUs), within production contexts, characterised by multiple outputs and inputs (Charnes et al., 1978). Over the past three decades, the scope of DEA has broadened considerably. In particular, DEA has been employed as a valid instrument to construct social and economic well-being indicators (see, among others, Despotis 2005). The adaption of DEA to the measuring of environmental and social aspects has required the changing of the objective function in the standard model in order to recognize the change in focus. To deal with the case of the production of human well-being and ecosystem, it is possible conceptualise a production process where each city is a “firm" which uses government resources to produce well-being outputs, such as better education, improvement of health status, greater access to labour markets, reduction of environmental pollution and so on. Accordingly, each city is assumed to have one “government" and hence one unit of input, and it produces the aforementioned outputs. Because we do not have the classic production context, but we can only rely on secondary variables, obtained as rates or combinations of primary variables, a DEA model with a single constant input can be suitably adopted. For the purpose of our work, we make use of the approach proposed by Lovell and Pastor (1999), in which the CCR and BCC models are equivalent. By adopting the output orientation, the linearized unitary input DEA-model is expressed by the following linear programming:

$$\max \quad h_0 \tag{3}$$

$$s.t \quad \sum_{k=1}^{n} \lambda_k y_{jk} \geq h_0 \, y_{j0} \quad \forall j \tag{4}$$

$$\sum_{k=1}^{n} \lambda_k \leq 1 \tag{5}$$

$$\lambda_k \geq 0 \quad \forall k \tag{6}$$

In equations (3-6) h_0 denotes the inverse of efficiency of the DMU under analysis (DMU_0), y_{jk} is the j_{th} output ($j = 1 \dots s$) of the DMU_k ($k = 1 \dots n$) and λ_k is the individual contribution of each DMU in the formation of DMU_0's target. Nissi and Sarra (2018) propose an integrated DEA-entropy approach to strength the discrimination power of that model.

3 Results and Conclusions

Our analysis is restricted to 103 province capital cities and takes into account eight out of eleven domains of the original Ur-Bes dataset: "*Health*", "*Education and Training*", "*Work and Life Balance*", "*Economic well-being*", "*Social Relationships*", "*Security*", "*Landscape and Cultural Heritage*", "*Environment*". Following the methodology described in the previous section, we first compute a spatial composite index for each pillar of Ur-Bes through the GW PCA. The GW PCA analyses have been carried out in R using the GWmodel package (Gollini et al.2013). We used a bi-square kernel function with adaptive bandwidths, whose sizes are selected automatically and objectively via cross-validation and not based on a priori decision. The output of the GW PCA allows to highlight the local change in the structure of multivariate data and how the original well-being indicators influence the local principal components retrieved for each of the Ur-Bes pillars. The GW PCA makes possible to display the localized proportions of the total variance (PTVs) and ascertain if the spatial patterns in the PTVs vary significantly across the study region. In general, for the majority of the urban well-being domains, the maps of PTVs, not displayed here, reveal that the highest PTVs are often located in the province capital cities of the South of Italy. Some exceptions are recorded for the "Security" pillar, for which the PTVs data are lower in the province capital cities of Central Italy. In the areas where highest PTVs are detected the local correlation (or local collinearity) among well-being data is assumed high, suggesting that not all Ur-Bes indicators need to be considered. By retaining only the first component, that accounts for a substantial proportion of the variability in the original data, it is possible, for a given local well-being domain, to extricate how each of the elementary indicators influences the selected pillar. For instance, in the domain "Health", the local principal components reveal multifaceted geographical variations in the variables with the largest loadings. From GW PCA results, we see that first component is mainly represented, in the province capital cities of Piemonte and Trentino

Alto Adige, by the age-standardised cancer mortality rate (19-64 years old); by life expectancy at birth (male) in most cities of central Italian regions and in Sardinia, and by life expectancy at birth (female) and mortality rate for road accidents (15-34 years old) for the Southern province capital cities. The spatial variation of the first local component of the “Education and Training" domain, reveals that it mainly considers the participation to primary school and this elementary indicator dominates in the most urban areas, with some exceptions for a number of province capital cities of Calabria and Sicily, where the leading variable is represented by the early leavers from education and training. This analysis is replicated for each dimension of Ur-Bes. Once weights are obtained for each variable in each location, the spatial composite indicator for the urban well-being dimensions, has been computed as the weighted sum (linear combination) of the variables, location by location. For the arising composite indices a data transformation has been undertaken to respect their positive o negative linkages with equitable and sustainable well-being and assure strictly positive data. In the second stage of the analysis, the overall well-being index is obtained via a unitary input DEA model, with entities defined only by outputs. We found that efficiencies are between 78.8% and 100%, while the mean efficiency is 97.2%. A large number of efficient cities are located in the North and Central part of Italy. Thirty-four cities achieve a well-being efficiency score which is below the average. In the last positions of the ranking we find Napoli, Salerno, Foggia, Bari, Benevento, Isernia, Caserta, Ascoli Piceno, Chieti, Macerata and Agrigento.

References

1. Bai, X., Nath, I., Capon, A., Hasan, N., Jaron, D.: Health and wellbeing in the changing urban environment: complex challenges, scientific responses, and the way forward. Curr. Opin. Environ. Sustain. **4**, 1–8 (2012)
2. Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. Eur. J. Oper. Res. **2**, 429–444 (1978)
3. Despotis, D.K.: A reassessment of the Human Development Index via Data Envelopment Analysis. J Oper Res Soc. **56**, 969–980 (2005)
4. Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P.: GWmodel: an R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models (2013)
5. Hall, J., Giovannini, E., Morrone A., Ranuzzi, G.: GA framework to measure the progress of societies. OECD statistics working papers, 2010/5, OECD Publishing (2010)
6. Harris, P., Brunsdon, C, Charlton, M.: Geographically weighted principal component analysis. Int. J. Geogr. Inf. Sci. **25**, 1717–1739 (2011)
7. Istat-Cnel: Comitato sulla misura del progresso in Italia. La misurazione del Benessere Equo e Sostenibile, CNEL, Roma (draft), www.misuredelbenessere.it (2012)
8. Nissi, E., Sarra, A. : A Measure of Well-Being Across the Italian Urban Areas: An Integrated DEA-Entropy Approach. Soc. Indic. Res. **136**, 1183–1209 (2018)
9. Lovell, C. A.K., Pastor, J.T.: Radial DEA models without inputs or without outputs. Eur. J. Oper. Res. **118**, 46–51 (1999)

Comparing Composite Indicators to measure Quality of Life: the Italian “Sole 24 Ore” case

Un confronto tra indicatori compositi per misurare la qualità della vita: il caso italiano de “Il Sole 24 Ore”

Gianna Agrò, Mariantonietta Ruggieri and Erasmo Vassallo

Abstract The measure of Quality of Life is still a topic widely discussed in the literature. In Italy, the newspaper "Il Sole 24 Ore" publishes a famous ranking that highlights a strong disparity between Northern and Southern Italy areas. In this paper, some methods are compared in order to show how different types of normalization and aggregation can influence the results.

Riassunto *La misura della qualità della vita continua ad essere un argomento ampiamente discusso in letteratura. In Italia, il quotidiano "Il Sole 24 Ore" pubblica una nota graduatoria che evidenzia una forte disparità tra le aree del Nord e del Sud Italia. In questo lavoro, si confrontano alcune procedure per mostrare come diversi tipi di normalizzazione e di aggregazione possano influenzano i risultati.*

Key words: Quality of Life, Composite Indicators, BoD approach, standardization

1 Introduction

The newspaper "Il Sole 24 Ore" has been publishing for 25 years a ranking of Quality of Life (QoL) for Italian provinces according to the NUTS3 classification. This survey produces every year a great media resonance, recalling a long-debated issue about the so-called "Questione Meridionale"; economic and social territorial disparities are highlighted by a strong polarization among the provinces of North Italy, with higher levels of QoL than the provinces of the South (Felice, 2013;

[1] Gianna Agrò, DSEAS, University of Palermo; email: gianna.agro@unipa.it

Mariantonietta Ruggieri, DSEAS, University of Palermo; email: mariantonietta.ruggieri@unipa.it

Erasmo Vassallo, DSEAS, University of Palermo; email: erasmo.vassallo@unipa.it

IlSole24Ore, 2017). Actually, this result is strongly influenced by the choice of the basic dimensions, the kind of data and indicators, the standardization and aggregation techniques to obtain a Composite Indicator (CI). There is no doubt that the economic and the social context presents a greater level of discomfort in the South of Italy (ISTAT, 2017a), but this can not immediately be related to lower levels of QoL. It is very difficult to provide a univocal definition of QoL, as this concept invests personal and/or community aspects, depending on subjective and/or objective well-being and happiness (Cummins, 1998; Nussbaum and Sen 2003). Certainly, the concept of QoL is much more than "standard of living" one, that is basically connected to income levels (ISTAT, 2017b). The absence of a clear and univocal definition makes the concept more difficult to translate into data, indicators and measures that are free of criticism. Nevertheless, the aim of this paper is not to discuss about the different definitions of QoL used in the theoretical and applied literature, but to focus on the statistical techniques used to standardize and synthesize data, influencing the final ranking of the Italian provinces. Although we highlight many perplexities on the redundancy of some data and their ability to correctly represent some dimensions, here we accept the choice of "Il Sole 24 Ore" about pillars and indicators used in its QoL dossier, well-known by journalists and politicians. In particular, we compare four procedures, described in the Section 2, including the BoD-DEA method, which provides endogenous weights emphasizing the strength of each province (Giambona and Vassallo, 2014). The results and comments are reported in Section 3, while in Section 4 we draw some conclusions.

2 "Il Sole 24 Ore" quality of life and BOD approach

The 2017 "Il Sole 24 Ore" dossier uses 42 indicators divided into 6 dimensions (income and consumption, labor and innovation, environment and services, demography and society, justice and safety, culture and leisure), each consisting in 7 indicators to measure QoL among the 110 Italian provinces, according to the NUTS3 classification of the territory. In this paper, we do not discuss the choice of indicators and dimensions of the analysis that, moreover, is not constant over time, but the procedure used to aggregate the data. In particular, the dossier proposes to standardize each of the 42 indicators for the *i*-th province with $std.value_i=(indicator\ value_i\ /\ max(indicator\ values))\cdot 1000$ if the indicator polarity is positive (higher values correspond to a better condition) and $std.value_i=(min(indicator\ values)/indicator\ value_i\ )\cdot 1000$ if the polarity is negative (lower values correspond to a better condition). In this way, the best province (benchmark) always assumes a maximum score of 1000 and the other provinces assume decreasing values up to a minimum; then a simple arithmetic mean among indicators/dimensions provides an overall score for each province. This method has the advantage of being very simple and intuitive, but it ignores some statistical features of data, for example the different variability exhibited by the different indicators. Besides, the simple arithmetic mean is a fully compensatory aggregation procedure; this implies that the indicators and dimensions have the same importance and they can also be considered perfectly complementary. A better way to standardize the indicators is to use the min-max procedure, i.e. *std.value*$_i$ *= (indicator value*$_i$ *- min(indicator values))/(max(indicator values)-min(indicator values))* if the polarity is positive and *std.value*$_i$ *= (max(indicator values) - indicator value*$_i$*)/(max(indicator values)-min(indicator values))* if the indicator polarity is negative; in this way, the standardized values always range between 0 and 1 for all indicators (OECD, 2008). Obviously, the compensatory nature of the mean is not resolved.

Among other proposals in literature (Munda and Nardo, 2009), the AMPI index is widely used by some ISTAT reports; it presents a simple solution to the full compensation of the arithmetic mean through the use of a penalty based on a function of variability (Mazziotta and Pareto, 2017). So, the indicators are standardized via min-max rescaled between 70 and 130 and the corresponding mean is reduced if the province has different values (higher variability) among the dimensions, i.e. $AMPI_i = M_i - S_i \cdot cv_i$, where M_i, S_i and cv_i are the mean, the standard deviation and the coefficient of variation, respectively, of the indicators for the i-th province.

Recent developments aim at removing any kind of subjectivity in the choice of right weights for aggregating indicators and/or dimensions. In this framework, the Benefit of Doubt (BoD) based methods have received many consents. These methods exploit the Data Envelopment Analysis (DEA), a frontier technique that has been usually used for measuring the efficiency in production. Several variants of BoD models have been proposed in literature (Rogge and Van Nijverseel, 2018). In this work, we use an appropriate geometric mean of the indicators, in which the weights are endogenously defined, depending on the characteristics of the data according to the principle of BoD. A Composite Indicator $CI_i = \prod_{j=1}^{n} I_{ij}^{w_j}$ is defined, where I_{ij} is the *j*-th indicator of QoL ($j = 1,...,n$) for the *i*-th province ($i = 1,...,m$) with weight w_j determined endogenously by an algorithm based on a multiplicative optimization model, that solves the following problem: $SI_i = \max_{w} \prod_{j=1}^{n} I_{ij}^{w_j}$ with constraints $\prod_{j=1}^{n} I_{ij}^{w_j} \le e$ and $w_j \ge 0$, where e is the Napier's constant (Zhou et al., 2007; Giambona and Vassallo, 2016). In this way, the CI is obtained by multiplying the basic indicators of QoL with weights calculated in the best possible conditions, i.e. increasing as much as possible the composite score for a given province. In short, a low value of the CI for the *i*-th province is due to low values of the indicators that compose it and not to specific weights, calculated to obtain the best, i.e. the maximum possible, result for the *i*-th unit compared to the benchmark province. At the end, we obtain scores between 1 and 2.71 (the Napier's constant), attributable to the most intuitive interval between 0 and 1 by applying the antilogarithm. However, the optimization problem could determine zero weight to some indicators and attribute too much weight to other indicators, and this is not desirable if all the dimensions are relevant. So, we add specific constraints on the weights; in particular, we add proportion constraints to the model: $\left(\prod_{j=1}^{n} I_{ij}^{\,w_j}\right)^{L} \leq I_{ij}^{\,w_j} \leq \left(\prod_{j=1}^{n} I_{ij}^{\,w_j}\right)^{U}$, where *U* and *L* ranging between 0 and 1 to represent the lower and the upper bound (in percentage terms) for the contribution of the *j*-th indicator. In fact, without constraints on the weights, the model could ignore the contribution of the under-performing indicators or dimensions to maximize the best solution, and this is not admissible. In our case *L*=10% (*U* is defined accordingly); we note that bounds slightly lower or slightly higher lead to similar results, so an intermediate value has been chosen among the possible alternatives, taking into account that *L* (to guarantee sufficient flexibility to the method) must be relatively low.

3 Results and comments

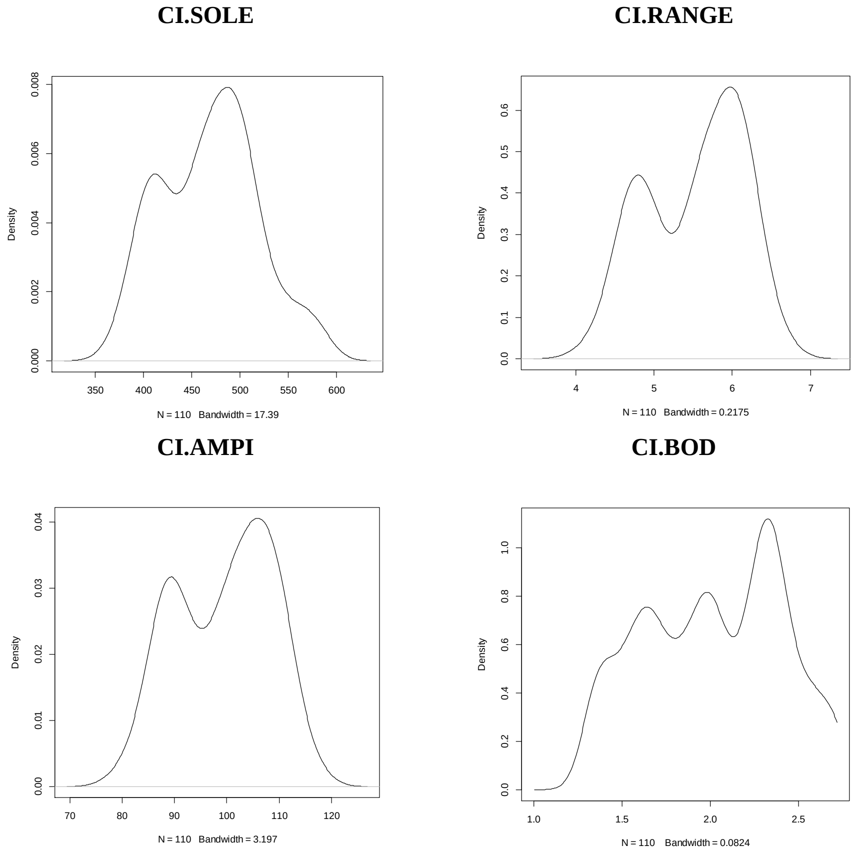
This Section presents the results obtained by means of the methods described in Section 3, that is: 1) the original "IlSole24Ore" with min or max standardization on 1000 and use of arithmetic mean (CI.SOLE); 2) the "range" with min-max standardization rescaled between 1 and 10 and arithmetic average (CI.RANGE); 3) the "ampi" with min-max standardization between 70 and 130 and use of a penalized arithmetic mean (CI.AMPI); 4) the "BoD" with min-max standardization, rescaled for calculation purposes between 2-10, and geometric mean with weights endogenously defined (CI.BOD).
In Table 1 the correlation matrix among the considered CIs is reported, showing a very high correlation between the series.

Table 1: Correlation matrix among CIs

	CI.SOLE	CI.RANGE	CI.AMPI	CI.BOD
CI.SOLE	1.00	0.94	0.93	0.82
CI.RANGE	0.94	1.00	0.99	0.93
CI.AMPI	0.93	0.99	1.00	0.94
CI.BOD	0.82	0.93	0.94	1.00

Figure 1 reports the kernel density estimates of the four CIs, where the BoD kernel density estimates uses the Silverman reflection method to avoid bias and inconsistency near the boundaries (Silverman, 1986; Scott, 1992). We note some “twin peaks” distributions, substantially in correspondence with the provinces of the southern and northern Italy. A similar polarization can be seen in BoD, although this distribution suggests a more detailed interpretative framework. In fact, the scores of the different methods are consistent each other, but the rankings are sometimes quite different. In particular, the BoD ranking shows a less intense polarization and the distance North-South appears less severe. In this regard, it is interesting to note the strong change of position of some provinces compared to the original classification of the "IlSole24Ore"; for example, some northern provinces lose many positions:

Monza e Brianza loses 55 positions (from 29 to 84), Verbano-Cusio-Ossola loses 49 positions (from 7 to 56), Sondrio loses 41 positions (from 3 to 44), and so on. On the contrary, other provinces acquire positions, such as Brescia (from 46 to 14), Padova (from 42 to 12), Venezia (from 43 to 15), etc.



CI.SOLE

CI.RANGE

CI.AMPI

CI.BOD

Figure 1: Kernel density estimates of the considered CIs

4 Conclusions

In this paper four different procedures, able to supply a unique composite indicator measuring the QoL, are applied to data reported on the well-known "IlSole24Ore" QoL dossier. We do not discuss the choice of indicators and dimensions, even if there are many critical aspects, but we focus on the possible consequences in applying different aggregative procedures. The original “IlSole24Ore” procedure uses a simple standardization and a fully compensatory arithmetic mean, similarly the "range" procedure that, instead, uses a better min-max standardization. The "ampi" technique uses again an arithmetic mean, but penalizes the provinces with greater variability among indicators/dimensions. Finally, the BoD method applies different weights to dimensions and provinces according to the specific characteristics of the data; besides, it uses a geometric mean to avoid the problem of a full compensation among indicators or dimensions. The BoD method is more specific and advanced from a technical point of view and addresses some critical aspects compared to simpler methods; nevertheless, it is not possible to assert that an endogenous choice of the weights is better compared to an exogenous one or to a choice of identical weights for units and dimensions. Certainly, the techniques of standardization and aggregation strongly influence the rankings, and this generates many doubts about the utility of these classifications. Finally, this work highlights that all the rankings maintain a strong North-South distinction, but this depends on strongly heterogeneous indicators between northern and southern provinces: with less heterogeneous variables, the ranking differences would be even more evident.

References

1. Cummins R.A. (1998). *Quality of Life Definition and Terminology: A Discussion*. Document from the International Society for Quality-of-Life Studies. International Society for Quality-of-Life Studies. *(mimeo).*
2. Felice E. (2013). *Perché il Sud è rimasto indietro*, Il Mulino, Bologna.
3. Giambona F. and Vassallo E. (2014). Composite Indicator of Social Inclusion for European countries. *Social Indicators Research*, 116, 1, 269-293.
4. Giambona F. and Vassallo E. (2016). *Composite Indicator of Social Inclusion for the EU Countries*. In: Alleva G. and Giommi A. (eds.) "Topics in Theoretical and Applied Statistics", Springer, Berlin.
5. IlSole24Ore (2017). *Dossier Qualità di vita 2017*, Il Sole 24 Ore, Milano.
6. ISTAT (2017a). *Rapporto Annuale*, Istat, Roma.
7. ISTAT (2017b). *Rapporto BES*, Istat, Roma.
8. Mazziotta M. and Pareto A. (2017). Measuring Well-Being Over Time: The Adjusted Mazziotta–Pareto Index Versus Other Non-compensatory Indices. *Social Indicators Research*, first online, doi: 10.1007/s11205-017-1577-5.
9. Munda G. and Nardo M. (2009). Noncompensatory/nonlinear composite indicators for ranking countries: a defensible setting. *Applied Economics*, 41, 12, 1513-1523.
10. Nussbaum M. and Sen A. (2003). *The Quality of Life*, Oxford Press, Oxford.
11. OECD (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Oecd, Paris.
12. Rogge N. and Van Nijverseel I. (2018). Quality of Life in the European Union: A Multidimensional Analysis. *Social Indicators Research*, first online, doi: 10.1007/s11205-018-1854-y.
13. Scott D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.
14. Silverman B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman $ Hall, London.
15. Zhou P., Ang B.W. and Poh K.L. (2007). A mathematical programming approach to constructing composite indicators. *Ecological Economics*, 62, 291–297.

Quality of working life in Italy: findings from Inapp survey

La qualità del lavoro in Italia: i risultati di un'indagine Inapp

Paolo Emilio Cardone[1]

Abstract Job quality is important to all citizens and is also central to policy concerns as Europe tries to boost economic growth and address the demographic challenge and the threats to the welfare systems. Making work sustainable and keeping people in work for longer are two key aspirations of many European countries. Maintaining and developing job quality is crucial for attaining these goals.

The paper aims at analyzing the behavior of the workers as regard the job quality in Italy in terms of working conditions, working time, work-life balance, health and well-being at work, pay and income.

More in detail, the paper provides a statistical picture of job quality in Italy, investigating its variability and relevant inequalities. The analysis is carried out using the quantitative research "*Indagine campionaria sulla Qualità del Lavoro*" (QdL), provided by Inapp ("*Istituto Nazionale per l'Analisi delle Politiche Pubbliche*" – Public Policy Innovation). It provides a rich portrait of workplace trends in Italy over the last fifteen years. Using logistic regression model, it is possible to estimate the job satisfaction level among workers more accurately. Overall the data confirm the existence of strong inequalities among workers and the findings also underline the importance of company and workplace practices in safeguarding health and safety (including against psychosocial risks), improving work–life balance, supporting career development, promoting skills use and development, managing workload and designing meaningful jobs. This requires policy attention, to increase the focus on job quality as part of active labour market policies.

Abstract *Il presente contributo fornisce un quadro statistico della qualità del lavoro in Italia, indagando la sua variabilità e le disuguaglianze rilevanti, analizzando il comportamento dei lavoratori in termini di condizioni di lavoro, orario di lavoro, equilibrio tra vita lavorativa, salute e benessere al lavoro, retribuzione e reddito. L'analisi è stata effettuata utilizzando "l'indagine*

[1] Paolo Emilio Cardone, INAPP-Statistical Office; email: p.cardone@inapp.org

campionaria sulla qualità del lavoro" (QdL), una ricerca quantitativa condotta dall'Inapp (Istituto nazionale per l'Analisi delle politiche pubbliche). Utilizzando un modello di regressione logistica, è stato possibile stimare più accuratamente il livello di soddisfazione del lavoro tra gli occupati. Nel complesso i dati confermano l'esistenza di forti disuguaglianze tra i lavoratori e sottolineano anche l'importanza delle prassi aziendali.

Key words: Ageing workforce; Job quality; Pay and income; Work-life balance; Working conditions; Logistic regression model.

1 Introduction

Job satisfaction for all is a goal for European education, training and employment policies. In particular, the Treaty on the Functioning of the European Union (TFEU) underlines as significant objectives the "promotion of employment, improved living and working conditions…proper social protection, dialogue between management and labour, the development of human resources with a view to lasting employment and the combating of exclusion" (Article 151 TFEU).

In the main slogan of the Europe 2020 strategy – 'smart, sustainable and inclusive growth' – the ideas of working conditions and job quality are implicit rather than explicit; however, 'improving the quality of work and working conditions' is a pivotal objective of the European Commission's 2010 'Agenda for new skills and jobs' initiative[2].

Improving working conditions and job quality continues to be a significant goal in European policies, underpinning Europe's capacity to compete. It is a cross-cutting issue that both influences and is influenced by many other European policies. For example, the Commission's recent industrial renaissance and enterprise policies have implications for working conditions and job quality (European Commission 2017). And conversely, the improvement of working conditions is important for the implementation of other European policies (i.e. innovation and gender equality).

Current European policy concerns include work–life balance (in particular for working parents), fighting undeclared and fraudulent work, extending working life, addressing the challenge of segmented labour markets and ensuring a proper balance between flexibility and security, investing in human capital, preparing individuals for potential risks over their life course, and addressing the significant inequalities that people face in the labour market (Judge and Watanabe 1993).

[2] "*High quality of work goes hand in hand with high employment participation. This is because the working environment plays a crucial role in enhancing the potential of the workforce and is a leading competitiveness factor. In order to innovate and to deliver promptly and efficiently, EU companies depend for their survival and expansion on a committed workforce, thriving in a high-quality working environment, with safe and healthy working conditions*".

The current policy debates on new forms of employment and undeclared and fraudulent work highlight the importance of monitoring working conditions and of providing data and analysis that can both enhance understanding of the common challenges faced by Europe and the Member States and support policymaking in these areas.

Numerous changes (demographic, structural and technological) are affecting the composition of the workforce, employment levels, job content and how workers experience their working lives. These developments challenge the role that work plays in our societies, with working life actors responding in different ways, and have a knock-on effect on working conditions and job quality (Eurofound 2017).

Furthermore, demographic ageing is an irreversible process. The direct effect of population ageing is the increasing share of elderly people, who are in retirement age, compared to the decreasing share of young people (Eurofound 2015).

The ageing of the European working population calls for policy attention to two issues: ensuring that demanding working conditions can be undertaken by an older workforce and ensuring that working conditions are sustainable over the life course to allow people to remain in work longer.

2 Data and methods

In order to achieve this goal, the analysis is carried out using microdata from the quantitative research "*Indagine campionaria sulla Qualità del Lavoro*" (QdL)[3], provided by Inapp[4].

The survey aims to measure working conditions in Italy, analyse the differences among workers, identify groups at risk, highlight issues of concern and areas of progress and, ultimately, contribute to developing policy aimed at improving job quality. In 2015, QdL interviewed over 15.000 workers, both employees and self-employed people, in Italy (with appropriate weights provided by Istat they are 22 million, exactly the workers' amount in Italy). Workers were asked a range of questions concerning employment status, work organisation, learning and training, working time duration and organisation, physical and psychosocial risk factors, health and safety, work–life balance, worker participation, earnings and financial security, as well as work and health (Clark 1998).

Going beyond the objective measures of job quality, the report also looks at workers' own assessment of their working lives. It finds associations between the different dimensions of job quality and factors such as engagement, financial

[3] For more details: http://inapp.org/it/dati/qualitadellavoro

[4] National Institute for Public Policy Analysis, former ISFOL - National Research Institute for Vocational Education and Training Employment, that changed its company name in INAPP (*Istituto Nazionale per l'Analisi delle Politiche Pubbliche – Public Policy Innovation*) on December 1st 2016 (www.inapp.org)

security, the development of skills and competences, health and well-being, the reconciliation of work and private life, and the sustainability of work.

The first part describes the main characteristics of the workforce in Italy. Apart from traditional aspects, such as employment levels broken down by occupation, sector or employment status, it also looks at indicators such as sex, age, educational attainment, country of origin, seniority, health status and household circumstances.

Subsequently it focuses on developments in job quality in Italy. The current analysis is based on the following seven indices: physical environment; work intensity; working time quality; social environment; skills and discretion; prospects; and earnings. The indices cover extrinsic and intrinsic job features captured from an objective perspective. They are based on positive and negative self-reported features of the job, which measure the concrete experiences of work and have been proven to have a causal effect (either positive or negative) on the health and well-being of workers. The analysis of each index's components is supplemented by other features of the job or the working environment, such as dealing with customers or place of work. Other organisational resources provided through employee representation at the workplace are also considered.

Special attention is paid to the perspective of the individual job-holder: how their skills match their job, what their level of engagement is with their job, whether it provides them with financial security, what their work–life balance and time preferences are like, and how they juggle their different roles as worker, family member and citizen. Finally, issues around health and well-being, as well as workers' views on the sustainability of work, are explored.

QdL paints a wide-ranging picture of Italy at work, across occupations, sectors and age groups. Its findings highlight actions for policy actors to help them address the challenges facing Italy today and provide detailed information on a broad range of issues, including exposure to physical and psychosocial risks, work organisation, work–life balance, and health and well-being.

The topics covered include employment status, working time duration and organisation, work organisation, learning and training, physical and psychosocial risk factors, health and safety, work-life balance, worker participation, earnings and financial security, as well as work and health, with a special attention to gender mainstreaming.

Since its launch in 2002 the "*Indagine campionaria sulla Qualità del Lavoro*" has provided an overview of working conditions in Italy in order to:

- assess and quantify working conditions of both employees and the self employed;
- analyse relationships between different aspects of working conditions;
- identify groups at risk and issues of concern as well as of progress;
- monitor trends by providing homogeneous indicators on these issues;
- contribute to national policy development in particular on quality of work and employment issues.

Using multivariate analysis (logistic regression models with Stata software) it is possible to estimate the job satisfaction level more accurately (Liu 2016). The model has been developed for employed adults only and includes, first of all, adults' socio-demographic characteristics (age, gender and citizenship), secondly, job and size enterprise.

In order to achieve this goal, we have used "Satisfaction" as the dependent variable (weighted model). Satisfaction = 1 if the worker is satisfied.

Concretely, in our study the following variables are considered (tab. 1):

- ***Gender***. Categorical. Dummy variable: Female, Male (reference cat.).
- ***Age group***. Categorical. Six intervals. From 18 to 24 (reference cat.); 25 to 34; 35 to 44; 45 to 54; 55 to 64; 65+.
- ***JobISCO***. Categorical. Nine levels. Elementary occupations (reference cat.); Managers; Professionals; Technicians and associate professionals; Clerical support workers; Services and sales workers; Craft and related trades workers; Plant and machine operators and assemblers; Military forces.
- ***Education level***. Categorical. Three levels. Secondary school (reference cat.); High school; University.
- ***Training***. Four levels: No training (reference cat.), Yes, paid by me; Yes, paid by other; Yes, paid by me and others.

Table 1: *Logistic regression model*

Variables		Beta	ODDS	Sign.
• Gender				
Male (ref.)	Female	0,28	1,33	0,003
• Age group				
18-24 (ref.)	25 - 34	-0,45	0,63	0,142
	35 - 44	-0,66	0,51	0,031
	45 - 54	-0,43	0,65	0,155
	55 - 64	-0,69	0,50	0,025
	65 - W	-0,17	0,84	0,666
• Education level				
Secondary school (ref.)	High school	-0,31	0,73	0,009
	University	-0,41	0,66	0,007
• Job ISCO				
Elementary occupations (ref.)	Managers	1,19	3,27	0,000
	Professionals	0,83	2,29	0,000
	Technicians and associate professionals	0,77	2,16	0,000
	Clerical support workers	0,71	2,02	0,000
	Services and sales workers	0,34	1,41	0,057
	Craft and related trades workers	0,68	1,96	0,000
	Plant and machine operators and assemblers	0,31	1,37	0,121
	Military forces	1,72	5,59	0,007
• Training				
No (ref.)	Yes, paid by me	0,34	1,40	0,021
	Yes, paid by other	0,62	1,86	0,000
	Yes, paid by me and others	0,65	1,92	0,005
	Intercept	2,02	7,56	0,000

Number of obs = 15,059
Wald chi2(19) = 94.85
Prob > chi2 = 0.0000
Log pseudolikelihood = -7266817.1
Pseudo R2 = 0.0283

Source: own elaboration on QdL data (Inapp survey)

3 Conclusions

Overall, the survey finds, structural inequalities and differences in terms of gender, employment status and occupation are still significant. In the last 10 years, there has been limited progress in some aspect of job quality.

Going beyond the objective measures of job quality, the report also looks at workers' own assessment of their working lives. It finds associations between the different dimensions of job quality and factors such as engagement, financial security, the development of skills and competences, health and well-being, the reconciliation of work and private life, and the sustainability of work.

Looking at the findings through the lens of the job quality profiles, jobs in the 'poor quality' profile would benefit most from actions to support the various dimensions of job quality and labour market policies focused on moving workers into better-quality roles (Cedefop 2015).

More generally, job quality can be supported by a wide-ranging set of policies and actions aimed at addressing the issues raised in the survey's analysis of job quality indices and profiles and that support workers throughout their working lives. In addition to policy initiatives at EU level, by national authorities and social partners, progress can also be achieved through workplace practices and policies at company level.

The improvement of working conditions takes place in a context of subsidiarity. Governments and social partners, companies and workers all have a role to play. Yet experience has shown that the EU is also a key player and has contributed to improving working conditions through its various measures with regard to the improvement of health and safety at work and gender equality, and its wider coordination of employment policies

Finally, vocational training plays an important role because the nature of jobs is changing, necessitating changes in the skills that are required of workers and adapting lifelong learning systems to the needs of an ageing workforce. The recent crisis has also highlighted the importance of education and training at all stages of life, in particular for older adults to avoid unemployment, vindicating the messages that "it is never too late to learn" and learning must be for all. This requires older people to maintain and update the skills they have, particularly in relation to new technologies. Continuous learning and development of an ageing workforce are important for employers' survival in competitive markets, as well as for maintaining older people's employability.

References

1. Cedefop (2015), *Skills, qualifications and jobs in the EU: The making of a perfect match? Evidence from Cedefop's European skills and jobs survey*, Publications Office of the European Union, Luxembourg.

2. Clark, A. (1998), “*Measures of Job Satisfaction: What Makes a Good Job? Evidence from OECD Countries*”, OECD Labour Market and Social Policy Occasional Papers, No. 34, OECD Publishing, Paris.
3. Eurofound (2015), *Sustainable work over the life course*: Concept paper, Publications Office of the European Union, Luxembourg.
4. Eurofound (2017), *Sixth European Working Conditions Survey – Overview report (2017 update)*, Publications Office of the European Union, Luxembourg.
5. European Commission (2017), *Employment and social developments in Europe* 2017, Publications Office of the European Union, Luxembourg.
6. Judge T., Watanabe S. (1993), *Another look at the job satisfaction-life satisfaction relationship*, “Journal of Applied Psychology”, n. 78.
7. Liu, X. (2016). *Applied Ordinal Logistic Regression using Stata*. Sage Publications. http://www.stata.com/bookstore/applied-ordinal-logistic-regression-using-stata

Well-being indices: what about Italian scenario?

Indici di benessere: lo scenario italiano

Silvia Facchinetti and Elena Siletti

Abstract This paper proposes a review of the Italian well-being scenario, introducing its peculiarities and defects. The aim is to provide a detailed image, also to policy-makers, describing the accuracy, timeliness and territorial coverage. All these characteristics are a challenge for a good planning, for a correct international comparison or to integrate well-being indices with new data, represented, as example, by big data.

Abstract *Questo lavoro propone una rassegna delle misure del benessere in Italia, presentandone peculiarità e difetti. L'obiettivo è fornire un'immagine dettagliata, anche ai policy-makers, descrivendone l'accuratezza, la tempestività e la copertura territoriale. Tutte queste caratteristiche rappresentano una sfida per una buona pianificazione, per un corretto confronto internazionale o per l'integrazione con nuove fonti di dati, rappresentate, ad esempio, dai big data.*

Key words: well-being indices, social indicators, quality of life, official statistics

1 Background

The importance of well-being has been widely acknowledged and its measurement is a matter that scholars have been tackling for a long time. Traditionally the most commonly employed indicator was the Gross Domestic Product (GDP) thanks to its ability to connect goods and services with different nature, to its linear methodology, to its objectivity and clearness, and the usefulness in international comparisons.

Silvia Facchinetti
Department of Statistical science, Università Cattolica del Sacro Cuore, largo Gemelli 1 Milano, e-mail: silvia.facchinetti@unicatt.it

Elena Siletti
Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, via Conservatorio 7 Milano, e-mail: elena.siletti@unimi.it

Many issues have been raised about the adequacy of GDP as an indicator of well-being, concluding that the use of GDP as a proxy of well-being can lead to misleading conclusions [3]. On the one hand negative events such as natural disasters, earthquake or foods, or big car accidents, reduce wealth of society but can increase GDP; on the other hand it doesn't take into account distribution of income, so that a great disparity and poverty of capabilities wouldn't be noticed. To overcome the GDP issues, alternative approaches have been developed across the years and they are still nowadays a core topic. A recent theoretical framework about well-being is the capability approach proposed by [7]. Following this approach, in 2007, the European Commission, European Parliament, Club of Rome, Organization for Economic Co-operation and Development (OECD) and WWF, hosted the conference titled "Beyond GDP". In August 2009, the European Commission released its road map, the Communication "GDP and beyond: Measuring progress in a changing world", when the so-called Stiglitz Commission [8] suggested to build a complementary statistical system, focused on social well-being and suitable for measuring sustainability. Shortly, going beyond GDP implies to develop indicators that are as clear and appealing as GDP, but more inclusive of environmental and social aspects of progress [6]. The proposed measure accounts a wide set of indicators representing both objective and subjective assessment, and including also people's perception of quality of life. The commission suggests to take into consideration the following aggregated dimensions: Material living standards, Health, Education, Personal activities including work, Political voice and governance, Social connections, Environment and insecurity. Following this advise, have been developed a huge number of indicators [3], with different structures, considering a great variety of dimensions and for many purposes. International examples of those indices are: the Human Development Index, the Better Life Index, the Happy Planet Index, the Canadian Index of Well-being and the Gross National Happiness Index.

In the field of subjective well-being, new measure have been recently proposed using big data: [9] developed a system to predict the life satisfaction of Facebook users based on lexical and topical features, many studies used Twitter data for the same purpose ([4]). In Italy, [2] proposed the iHappy index, an indicator of happiness that, using information from million of tweets, measures the average level of idiosyncratic happiness in the 110 Italian provinces. [5] developed the Social Well-being Index (SWBI), a multidimensional well-being composite indicator relying always on Twitter data and derived from a new human supervised technique of sentiment analysis (see [1] for details). All these proposals try to overcome the lack in official statistics, they use this new kind of information represented by big data whether by integrating them with statistics or by themselves, in any case for this kind of proposals the methodological challenges are still many and open.

In this paper we review the traditional indices used to measure well-being in Italy, their purposes, limitations, and their progressive evolution. This simple description is the starting point to understand well-being measurement in our Country, how it is addressed, and what is the quality of this kind of data. In future steps of our research we will propose new approaches to maximize advantages and to cut down limitations of the treated indices.

2 The Italian case

The interest in Italy towards well-being is constantly increasing, also due to the fact that there are still a number of open questions related to its measurement. First of all, we need accurate indices to face the European and international comparison in the best possible way. But above all it is increasingly necessary to inform policy-makers about the perceived effect of local policies. Then availability of sub-national data would be extremely useful, and due to the active well-being process that rapidly changes over time, high frequencies information would be even more welcome. For all these reasons the knowledge, and the characteristics of information sources about well-being in Italy is a straightforward and actual interest.

In the following subsections we describe: the Better Life Index (BLI), used for comparison in international context, the Fair Sustainable Well-Being Index (BES), elaborated by the Italian Institute of Statistics (ISTAT) as the Italian answer to the Stiglitz Commission proposal, and the “Il sole 24 ore” Quality of Life (QoL) index, that has a long history in the Italian society.

2.1 The Better Life Index - BLI

The BLI initiative, launched in 2011, is built around eleven topics that reflect what the OECD has identified as essential to well-being in terms of material living conditions (Housing, Income and wealth, Jobs and earnings) and quality of life (Social connections/community, Education and skills, Environmental quality, Civic engagement and governance, Health status, Life satisfaction, Personal security or safety, Work and life). Each dimension is built on one to four specific indicators that are averaged with equal weights, and have been chosen on the basis of a number of statistical criteria such as relevance and data quality and in consultation with OECD member countries. As the BLI gathers many indicators, expressed on different units, in order to compare and aggregate the values a normalisation, according to a standard formula, is performed. Moreover, thanks to its web platform, to visualise and compare the performances of the members, it is also possible to mix the set of dimensions, giving different weights, in order to elaborate an index coherently with one’s preferences (see [6]).

Data mostly come from official sources such as the OECD or National Accounts, United Nations Statistics, National Statistics Offices. Instead, some indicators are based on data from the Gallup World Poll: a division of the Gallup Organization that regularly conducts public opinion polls in more than 140 countries around the world. Unfortunately, about time frequency, almost always information is delayed. With reference to the Italian 2017 index only the 21% refers to 2016, sometimes the measure is the 3-year average 2014-2016, and for all the other indicators there is a greater lag, that, for example, for ”time devoted to leisure and personal care” arrives to 18 years. To aggravate the situation of this index, essentially born for comparisons between countries, the fact that these delays change with the country considered.

At the moment, most of the indicators in the BLI are not available at a more disaggregated level for a great number of countries, but there is a regional edition (https://www.oecdregionalwellbeing.org). However, in this case, the situation worsens even more, it should be noted that in the June 2016 release, currently on line as the latest available, the most updated data for Italy dates back to 2014. Perhaps because of these source disparities between countries, in all the OECD publications, they talk about "first year available" and "last year available", without specifying, if not in the most specific material, the true reference for each country.

2.2 The Fair Sustainable Well-Being Index - BES

The BES (Benessere Equo Sostenibile) is setting up from a dashboard of twelve dimensions: Economic well-being; Education and training; Environment; Health; Landscape and culture heritage; Politics and institutions; Research and innovation; Security; Service quality; Social relationship; Subjective well-being; Work and life balance. Although it is clear conceptually and statistically similar to the Better Life Index, in the first two years the BES does differ it avoided any form of aggregation: in the periodical report, for each dimension, the entire set of proxies, with 130 indicators, is presented and discussed. For this reason, from 2013 to 2014, the only available aggregation form was provided by IRES Piemonte (the Regional Institute for Economic and Social Research of Piemonte). Since 2015, ISTAT provided by its self a general and domain specific composite indicators for Italian regions, nevertheless, for the composite indicator of the single dimensions often not all the indicators are accounted. Concerning time frequency, also in this case quite always the information is delayed. Only for instance, for Health in 2015 no one indicators refers to 2015, the 36% refers to 2014, and the 64% is more delayed. In addition to the regional edition, there are, for some years and some Italian provinces or chief towns, some pilot versions of the BES in which the indices considered are not the same and the subjective component is not provided for.

2.3 The "Il Sole 24 Ore" Quality of Life Index - QoL

Since 1990, the Italian business newspaper "Il Sole 24 Ore" publishes a well-being index of the Quality of Life (QoL) for all the Italian provinces.This composite indicator is strictly objective and it is defined along six categories, namely: Income, savings and consumption; Environment, services and welfare; Business, work and innovation; Justice, security and crime; Demographics, family and integration; Culture, leisure and participation. Since 2016, the considered indicators for every topic pass from six to seven, achieving a total number of 42 indicators. Note that, as the QoL index cover only material quality of life, it becomes for media a benchmark indicator for Italian well-being. Despite efforts to improve the quality, the index, in addition to having a low frequency with only an annual data, often shows delayed information. This is a serious flaw when decision makers want to base their choices on such information. Moreover, the indicators associated to the different dimensions change over the time, and in some cases the same indicator correspond to different dimension year by year. For example, the ”broadband coverage”, traditionally in Environment, services and welfare topic, from 2013 to 2014 pass to Culture, leisure and participation. Furthermore, also when the indicator is the same, sometimes is detected with different measures. For instance, the information about ”shows” in Culture, leisure and participation some years has been detected as number of shows, number of cinemas, or number of cinema seats per inhabitants. For this reasons with time series we are obliged to use the score and not the real value of the variables.

3 Negative and positive aspects of Italian well-being indices

This review shows that well-being information actually provided for Italy has been improved according to the suggestions of scientific international community. In Table 1 a synthesis of their principal traits. Especially, BES takes into account both subjective and objective dimensions, and with its great number of indicators, it performs a good multifaced source of information. Nevertheless, despite all efforts made to improve the measures of well-being, all the indices still presents several gaps. First of all, concerning time frequency, in addition to be designed as low frequency data the info is often delayed. Moreover, some indicators associated to the different dimensions change over time or the same indices correspond to different dimension year by year, or they are detected with different measures. Finally many indicators are not available at a more disaggregated level: they do not allow comparisons of disparities within a country or between social groups.

4 Conclusion

All the indices flaws, summarised above, are serious issues when decision makers want to base their choices on them, or when you need to make national or international ranking. To overcome this matter and to improve the quality of well-being data, the efforts could be address to a better data planning or to the integration of official statistics with more frequent and light surveys with new kind of information like social networks or big data. Before doing this, we can’t forget to solve the linked methodological issues, and official statistics must however be more accurate, more frequent and considered local level too, without forgetting the subjective dimension.

Table 1 Synthesis of the principal traits for the three considered indices

Index	*Starting year*	*Frequency*	*Dimensions*	*Indicators' number*		*Subjective*	*Level*
BLI	2013	Annual	Housing	3	24	Subjective dimension "life satisfaction" with only 1 indicator	National Regional level with delays and disparities
			Income and wealth	2			
			Jobs and earnings	4			
			Social connections/community	1			
			Education and skills	3			
			Environmental quality	2			
			Civic engagement and governance	2			
			Health status	2			
			Life satisfaction	1			
			Personal security/safety	2			
			Work and life	2			
BES	2013	Annual The 19% of the indicators are surveyed with more delay	Work and life balance	14	130	The subjective dimension, with its 4 indicators is available at regional and not for provincial level	Regional There is a local pilot experience only for few self-selected provinces from 2013 to 2015
			Economic well-being	10			
			Education and training	11			
			Environment	15			
			Health	14			
			Landscape and cultural heritage	12			
			Politics and institutions	12			
			Service quality	11			
			Research and innovation	7			
			Security	11			
			Social relationships	9			
			Subjective well-being	4			
QoL	1990	Annual	Income, savings and consumption	7	42	Unconsidered	For the 110 Italian provinces
			Environment, services and welfare	7			
			Business, work and innovation	7			
			Justice, security and crime	7			
			Demographics, family and integration	7			
			Culture, leisure and participation	7			

References

1. Ceron, A., Curini, L., Iacus, S.M.: iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. Information Sciences, **367-368**, 105–124 (2016)
2. Curini, L., Iacus, S.M., Canova, L.: Measuring Idiosyncratic Happiness Through the Analysis of Twitter: An Application to the Italian Case. Social Indicators Research, **2**, 525–542 (2015)
3. Fleurbaey, M.: Beyond GDP: The Quest for a Measure of Social Welfare. Journal of Economic Literature **47**, 1029–75 (2009)
4. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M.: Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. PLoS ONE, **6**, 1–26 (2011)
5. Iacus, S.M., Porro, G., Salini, S., Siletti, E.: Social networks data and subjective well-being: an innovative measurement for Italian provinces. Italian Journal of Regional Science, in press.
6. OECD: How's Life? 2013. Measuring Well-being, OECD Publishing (2013)
7. Sen, A.: Development as Capability Expansion. Readings in Human Development, Oxford University Press (2003)
8. Stiglitz, J., Sen, A., Fitoussi, J.P.: Report by the Commission on the Measurement of Economic Performance and Social Progress, INSEE (2009)
9. Schwartz, H.A., Sap, M., Kern, M.L., Eichstaedt, J.C., Kapelner, A., Agrawal, M.,. . . Ungar, L.H.: Predicting Individual Well-Being Through the Language of Social Media. Pacific Symposium on Biocomputing, **21**, 516-527 (2016).

How can we compare rankings that are expected to be similar? An example based on composite well being indicators.

Confronti tra graduatorie simili. Un esempio basato su indicatori compositi di benessere.

Silvia Terzi e Luca Moroni[1]

Abstract We compare 5 different well-being rankings derived from the following composite indicators: Human Development Index, Inequality Adjusted Human Development Index, Legatum Prosperity Index, Good Nation Index, Development Sustainable Goals Index.

Abstract *Confrontiamo tra loro 5 diverse graduatorie di benessere basate sui seguenti indicatori composti: Human Development Index, Inequality Adjusted Human Development Index, Legatum Prosperity Index, Good Nation Index, Development Sustainable Goals Index.*

Key words: well being, concordance, local concordance

1 Well being indicators and derived rankings

The most widely-used and well-known composite indicator of well being is the Human Development Index (HDI). It is a summary measure of average achievement in key dimensions of human development: *a long and healthy life,*

[1] Silvia Terzi Department of Economics RomaTre University; silvia.terzi@uniroma3.it
Luca Moroni, Department of Economics RomaTre University; luca.moroni@uniroma3.it

being knowledgeable and have a decent standard of living [3]. The HDI is the geometric mean of normalized indices (life expectancy, education, Gross National Income) for each of the three dimensions. However this index does not account for disparities (inequalities) within each dimension across the population; thus United Nations Development Programme also computes an Inequality adjusted HDI (IHDI). The IHDI combines a country's average achievements in health, education and income with how those achievements are distributed among country's population by "discounting" each dimension's average value according to its level of inequality.[2]

Other well being indicators available for a large number of countries worldwide are: the Legatum Prosperity Index, the Good Country Index, the Sustainable Development Goals Indicator. The Legatum Prosperity Index (LPI) [8], based on 104 indicators , is an aggregation of nine sub-indexes: (1) economic quality, (2) business environment, (3) governance, (4) education, (5) health, (6) safety & security, (7) personal freedom, (8) social capital, (9) natural environment.

The Good Country Index (GCI) [1] is based on 35 indicators related to the following 7 dimensions: (i) Science, Technology & Knowledge; (ii) Culture; (iii) International Peace and Security; (iv) World Order; (v) Planet and Climate; (vi) Prosperity and Equality; and (vii) Health and Wellbeing.

The Sustainable Development Index (SDGI) is based on several normalized indicators for each of the 17 Sustainable Development Goals (see [6] for the list of indicators and goals).

A careful insight into these different indicators goes beyond the standards allowed for our communication. Otoiou et al [5] explore whether the variables used in computing three of the most widely known indicators of well-being and social progress, the HDI, LPI, and Happy Planet Index, can be used to develop a relevant cluster structure, which can then be used to assess the validity and reliability of the country rankings obtained by these indicators. Among other comments and conclusions, they argue that the optimal cluster structure is very close to HDI country rankings. Making moves from the assumption that all the quoted indicators are well being indicators, and that what receives attention from media and policy makers are the country rankings of these different composite indicators, our interest focuses on the ranking comparison. How can we measure the similarities or the distances or the agreement between the different rankings? Is one of these rankings to be preferred with respect to the others? Are our comparisons in line with Otoiou et al.?

[2] A different kind of adjustment has been suggested by Terzi [7]: to correct HDI by means of a Multidimensional Poverty Indicator. This gives rise to the *corrected HDI*. However, MPI is available only for a subset of the countries for which HDI is computed so we decided not to consider MPI; not as a well-being indicator nor as a correction factor.

2 Ranking comparisons

First of all let us examine the correlations among the rankings of the 5 chosen indexes (Good Country [1], Human Development [3], Inequality Adjusted HDI [3], Legatum Prosperity [8], Sustainable Development Goals [6]) for n = 120 countries, for the most recent available years (2017 for GCI and SDGI; 2016 for the others).

	GCI	HDI	IHDI	LPI	SDGI
GCI	1.000	0.843	0.852	0.851	0.836
HDI	0.843	1.000	0.983	0.896	0.939
IHDI	0.852	0.983	1.000	0.888	0.958
LPI	0.851	0.896	0.888	1.000	0.872
SDGI	0.836	0.939	0.958	0.872	1.000

Table 1: correlations among rankings

Also, we have the concordance:

Kendall's $W = 0.913$.

From the correlations and from Kendall's coefficient, the 5 rankings appear to be very close (in particular HDI, IHDI, SDGI).

However what could interest mostly the policy makers are the differences in rankings of the best and of the worst performing countries. For this reason we resort to Top-Down concordance coefficient [4].[3] The Top-Down concordance coefficient is derived by computing Kendall's W not on ranks R$_i$ but instead on Savage scores

$$S_{R_i} = \sum_{j=R_i}^{n} \frac{1}{j}$$

In particular we have a top-down *low* concordance coefficient based on W (W_{TDL}) by substituting each rank R_i with the respective Savage score S_{R_i} when our interest is for the concordance among the lowest ranks, i.e. the top of the distribution; and a top-down *high* concordance coefficient when our interest is for the highest ranks, i.e. to bottom of the distribution. In this case we substitute each rank R_i with the Savage score:

$$S_{n-R_i+1} = \sum_{j=n-R_i+1}^{n} \frac{1}{j}$$

thus obtaining the Top-Down (high) concordance coefficient W_{TDH}. By the way, this is the same as calculating W_{TDL} on the descendent rankings of the objects.

For our rankings we have:

[3] If interested in the agreement among two rankings this could be achieved by means of weighted rank correlation, derived from Spearman's ρ (see for example Dancelli et al [2]).

Top-Down Kendall high $W_{TDH} = 0.892$,
Top-Down Kendall low $W_{TDL} = 0.839$.

Thus the 5 rankings are closer in the tail of the distributions than for the best performing countries.

To have an even closer insight of the concordance in the top or bottom or in the central rankings, we consider a local headcount function. We partition the set of the first $n = 120$ natural numbers (the ranks of the 120 observed countries) in contiguous subsets of fixed size s (for example 10% of the observations; 12 in our case) of consecutive naturals, and count for each subset how many units (countries), for at least one of the $d = 5$ rankings, rank in that subset. A local headcount of $s = 12$, the minimum of its range, means that in the interval spanned by the subset there are exactly s units, whose ranks are thus very close on all indicators. Conversely, a local headcount of $min(n, s.d) = 60$, the maximum of its range, means that in the span of the subset no unit ranks more than once.

The interpretation we suggest for the local headcounts is that they are inverse indicators of local concordance, because the headcounts are smaller when the ranks are closer and bigger when the ranks are spaced; so we derive a local concordance function as the maximum, $min(n, s.d) = 60$, minus the value of the local headcount. Moreover, we calculate the local headcount and the local concordance for each subset of s consecutive naturals in $(1, \ldots, n)$, and associate the value to the central rank of the subset, so that we can plot a smooth curve (Figure 1); the local concordance of the partition subsets are enhanced in red.

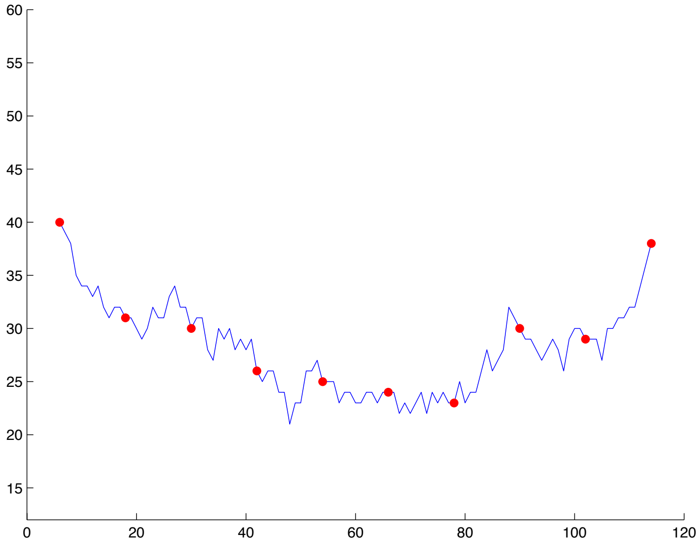


Figure 1: local concordance function

Alternatively we could be interested in finding out which well being indicator best represents all the others.

Let us define a centrality measure $C_{1(J)}$ for each ranking J that counts to how many units it assigns the median rank among the 5. Let us also define a de-centrality measure $C_{2(J)}$ as the sum of the differences between the unit's rank (in the J-th ranking) and the median rank. These values are summarized in Table 2.

Rankings	C_1	C_2
GCI	13	1365
HDI	36	449
IHDI	40	411
LPI	24	1079
SDGI	29	721

Table 2: centrality and de-centrality

Thus the ranking that best represents all the five well-being rankings is IHDI.

The same conclusion appears from Figure 2, in which on the *x*-axis we represent the countries ordered by means of their median rank and on the *y*-axis the ranks and their median.

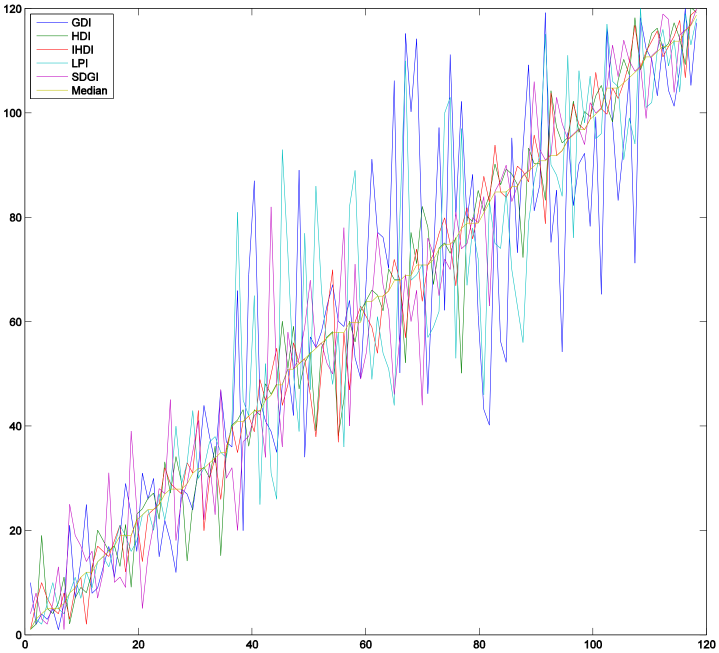


Figure 2: rankings and median by country vs ranking of medians

3 Concluding remarks

Although for the chosen well-being indicators Kendall's coefficient is high (W = 0,913), there are differences among the five rankings. In particular, there are differences among the units ranked in the center of their ranking (as can be seen by means of our local concordance function) and asymmetries in concordance among highest and lowest ranks (as can be seen by computing top-down concordance coefficients). From Figure 2 we can draw similar considerations: the greatest variability among ranks is in the centre of the plot; moreover, it can be seen that the farthest-away-from-the-others ranking is the GCI based ranking, whereas the most intermediate is based on IHDI, and this conclusion is reinforced by our centrality and de-centrality measures in Table 2. Overall, our findings seem consistent with Otoiu's findings [5].

References

1. Anholt, S., Govers, R (2014) The good country index. Tech. rep., The Good Country Party, http://www.goodcountry.org
2. Dancelli L., Manisera M., Vezzoli M. (2013) On Pinto da Costa & Soares' and other Weighted Rank Correlation Measures deriving from the Spearman's rho. Advances and Applications in Statistics, 36, 2, 83-104.
3. Human Development Report (2016) - Human Development for Everyone – UNDP.
4. Iman R.L., Conover W. J. (1987) A Measure of Top – Down Correlation. Technometrics, 29, 3, 351-357.
5. Otoiu A., E. Titan, R. Dumitrescu (2014) Are the variables used in building composite indicators of well-being relevant? Validating composite indexes of well-being. Ecological Indicators, 46, 575-585.
6. SDG Index & Dashboards Report (2017) Global responsibilities. Bertelsmann Stiftung and Sustainable Development Solutions Network.
7. Terzi, S. (2013) How to Integrate Macro and Micro Perspectives: An Example on Human Development and Multidimensional Poverty. Social Indicators Research, 114, 3, 935-945.
8. The Legatum Prosperity Index (2017) Legatum Institute, http://www.prosperity.com.

6. Poster Sessions

A distribution curves comparison approach to analyze the university moving students performance

L'analisi delle preformance degli studenti universitari in mobilità, secondo un approccio basato sul confronto tra curve di distribuzione

Giovanni Boscaino, Giada Adelfio, Gianluca Sottile

Abstract Nowadays in Italy we observe a one-directional migration flow of university students, typically from the South to the North. It represents the new millennium migration flow: people migrate looking for better job opportunities already during their educational path, believing that northern universities may provide more opportunities for being more successful. This paper aims to study the performance of those Sicilian students that move to the northern universities, to take the second level degree, in comparison with those remain in Sicily. We want to test the empirical evidence that shows a similar performance between the two groups of students. We use different measures of performance and follow a new methodology based on the comparison among the distribution curves. Results seems to confirm our idea, highlighting some difference.

Abstract *Da qualche anno anche l'Italia sta assistendo a un flusso migratorio di studenti, che da un povero Sud si dirige verso un più ricco Nord. Se un tempo la migrazione avveniva nel momento di cercare lavoro, adesso questa è anticipata da studenti che ritengono di avere meggiore successo se conseguono il titolo al Nord. L'obiettivo di questo studio è verificare l'ipotesi secondo la quale gli studenti che restano per iscriversi alla Laurea Magistrale hanno un percorso simile rispetto a quelli che si iscrivono a un percorso magistrale del Nord. Considerate diverse misure di performance, e svolto i confronti attraverso una nuova procedura di raffronto tra curve di distribuzione, i risultati non mostrano sostanziali differenze.*

Key words: performance, student migration, distribution comparison

Giovanni Boscaino, Giada Adelfio, Gianluca Sottile
Dipartimento di Scienze Economiche, Aziendali e Statistiche, viale delle Scienze, Edificio 13, Palermo (Italy) e-mail: giovanni.boscaino; giada.adelfio; gianluca.sottile @unipa.it

1 Introduction

The intra-European students' mobility is one of the targets of the Bologna Process to promote EU integration and foster a higher quality educational system. As known, the individual student mobility is a great opportunity to gain experience, to increase the knowledge of other cultures, to enhance human capital formation, networks, and relational abilities. Literature offers a lot of papers on the international student mobility [7] [4] [11] and its determinants, and they underline the influence of the socio-economic and cultural conditions of both the areas of origin and the destination. Actually, other factors could influence the wish of moving abroad, such as the quality of the receiving universities and the migration costs.

Studies on internal student mobility are rather scant on UK [8] and on Netherlands [13], where domestic student mobility flows are almost entirely unidirectional – from South to North – mirroring the internal economic migration. This is the case of Italy too: in fact, the migration flows from the South to the North of Italy, historically due to the different job opportunities, has changed in students migration flows. Probably, this is due to the expectation of students who think they will be more successful if they graduate in a more economically developed area [3].

Literature offers several studies about student performance, mainly devoted to find its determinants, and often based on different measures [16] [1] [5]. On the other hand, it is not so common to find papers about the performance of the university moving students [12] [3] [6]. Our research focuses on this new point of interest, and in particular on students that enroll in a university second level degree at institutions outside the students region of origin.

Comparison is made using a new method for curve clustering [15], since we compare the distribution curves (marginal and conditional to some given socio-demographic characteristics) of students, both in terms of performance and time to the degree.

2 The new method for clustering of curves

The clustering of distribution curves here performed is based on a new method to find similarities of curves in a quantile regression coefficient modeling framework, also multivariate, in which the effect of covariates on a response variable is represented by curves in the space of percentiles [9] [10]. In particular, let y be the response of a dependence model problem, in [15] the authors first estimate the regression coefficient functions $\beta_1(p \mid \boldsymbol{\theta}),\ldots,\beta_q(p \mid \boldsymbol{\theta})$, namely effects curves, and then assess if these q curves, that describe the effects of each covariate on the response, can be clustered based on similarities of effects, as a variable selection procedure.

The proposed clustering approach is based on a new measure of dissimilarity that uses both the shape of a curve and its distance with respect to other curves:

- the *shape* of a curve is evaluated using its second derivative. Moreover, two different curves are similar in shape if, at any given point, the signs of the second derivatives are concordant;
- the *distance* between two curves is evaluated as their with respect to other curves. Two curves are said close if their distance at any given point is lower than a fixed value.

Let i and i' be two different curves, $\boldsymbol{p} \in (0,1)$ the vector of percentiles. Then

$$d^{ii'}_{\text{shape}}(\boldsymbol{p}) = I(\text{sign}(\beta''_i(\boldsymbol{p} \mid \boldsymbol{\theta})) \times \text{sign}(\beta''_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})) = \mathbf{1})$$
$$d^{ii'}_{\text{distance}}(\boldsymbol{p}) = I(|\beta_i(\boldsymbol{p} \mid \boldsymbol{\theta}) - \beta_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})| \leq f(\alpha, \text{dist}(\boldsymbol{p}))),$$

where $f(\cdot,\cdot)$ is a cut-off function, that depends on a probability value α, and on dist($\boldsymbol{p}$), the vector of the distances between two curves across all percentiles. Finally, the proposed dissimilarity measure between two curves is defined as

$$d(i,i') = 1 - \int_0^1 \left[d^{ii'}_{\text{shape}}(p) \cdot d^{ii'}_{\text{distance}}(p)\right] \mathrm{d}p. \tag{1}$$

In [15], the new measure is used, to account for their concordance at each point, and any hierarchical clustering method can be applied. The method has been developed in the R package `clustEff` [14]. The proposed approach is very flexible and can be generalized to different contexts.

3 Data, Methodology and results

Thanks to a partnership with the Italian Ministry of Education and Research (MIUR) and four Italian Universities (University of Cagliari, University of Palermo, University of Siena, University of Torino), we refer to the whole dataset of the students enrolled in 2008 at any Degree Course offered by any Italian University, and followed for 8 years. In such a way, it is possible to follow the whole career of the students and their eventual mobility along the Italian area.

The dataset is very large, more than 100,000 records, where the record is the student. Many variables are available: gender, age, nationality, residence, diploma mark, high school diploma type, university and degree course attended at each year of observation, residence, average mark and total credits gained at each year.

In particular, the statistical unit is the student that graduates at the first level degree and that enrol in a second level degree course. The comparison regards three groups of students: i) the group of the Southern graduates that enrol in one of the universities in the South of Italy; ii) the group of the Southern graduates that enrol in the Northern university; iii) the group of resident in the North of Italy and still studying in the North.

Our specific question is “is there any difference in performance between moving and unmoving students?”. We refer to the Southern students as a first step analysis of the consistent migration flow from the South to the North of Italy (fig. 2). The chords represent the migration flows, linking the leaving and destination regions. As thicker is the chord, as more significant is the migration flow. The Fig. 1 highlights that main attractive universities are in the Centre-North of Italy, in particular in Lombardia, Piemonte, Emilia Romagna and Lazio.

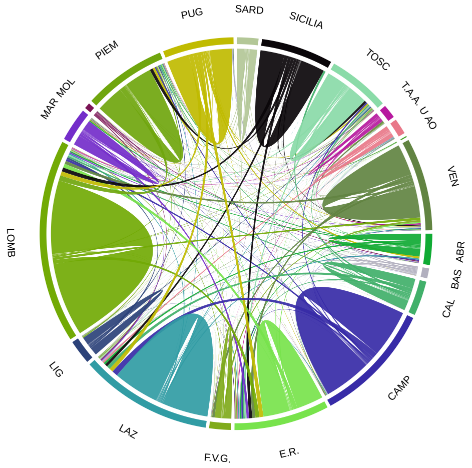


Fig. 1 Chord Diagram: migration student flows within Italy

The performance, in this paper, is measured accounting for two different perspectives: i) marks, ii) gained educational credits. In order to account both for the information of marks and credits, we use a new indicator that combines marks and credits in a unique measure, keeping unchanged the mark-scale [2]. In such a way, there is no need to analyze the two distributions of marks and credits because the new measure weights each mark by the correspondent exam credit.

We report the analysis of the distribution of the transformed mark conditioned to the High School Diploma Type (binary variable “Lyceum”, “No-Lyceum”), and our procedure identifies four clusters, together with interesting results, suggesting the necessity of further analysis (fig. 2). Indeed, Northern No-Lyceum students still studying in the North perform worse

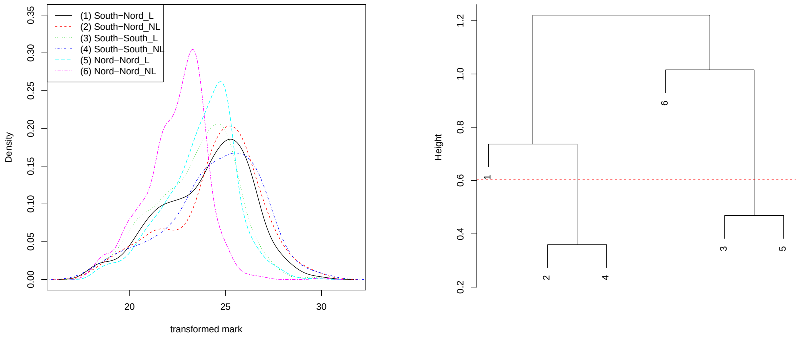


Fig. 2 Output of the proposed algorithm: the distribution curves (on the left) and the corresponding dendrogram (on the right)

than others and stay separate from the other students. Southern and Northern Lyceum student performance is similar (density curves are in the same cluster) when they do not migrate, and stay studying in their macro-area of residence (i.e. South or North, respectively). The third cluster is the one of the Lyceum Southern students migrating in the North, that perform on average a slightly better than the students of the two previous clusters. Finally, No-Lyceum Southern student performance does not change both if they migrate to the North or they continue their studies in the Southern Universities, and they seem to perform better, on average, than the students of the three previous cluster.

We perform the analysis of the distribution of the transformed mark conditioned to the gender too but, although the clustering approach identifies three clusters, results do not show difference in performance.

These results, just provisional and partial, suggest the use of the proposed approach such as an exploratory method for studying this kind of dependence, based on the comparison of distribution functions. Indeed, further analysis (not here reported for the sake of brevity) considering a quantile regression model approach for the student performance, confirms our exploratory analysis results and highlights some differences with respect to other variables, like topic area of the degree course and going into detail of the geographical areas for students of the three groups.

Acknowledgements

This paper has been supported by the national grant of the Italian Ministry of Education University and Research (MIUR) for the PRIN-2015 program “Prot. 20157PRZC4. PI: G. Adelfio".

References

1. Adelfio, G., G. Boscaino, and V. Capursi. A new indicator for higher education student performance. Higher Education 68 (5), 653-668 (2014).
2. Adelfio, G.; Boscaino, G., and Capursi, V. Further considerations on a new indicator for higher education student performance. Proceedings XLVIII Scientific Meeting of the Italian Statistical Society (2016)
3. Attanasio, M.; Enea, M. La mobilitÃă studentesca In Italia: un'analisi dei flussi dal Sud d'Italia verso il Centro-Nord. Proceedings of Popdays 2017. (2017)
4. Beine, M.; Bertoli, S; and Fernandez-Huertas Moraga, J. A practitionersâĂŹ guide to gravity models of migration, The World Economy, vol 6, n- 4, 496-512. Wiley Blackwell. (2016)
5. Birch, E. R. and P. W. Miller. Student outcomes at university in Australia a quantile regression approach. Australian Economic Press 45 (1), 1-17.(2006)
6. Boscaino, G.; Vassallo, P.. La migrazione studentesca dalla laurea triennale alla laurea magistrale. Proceedings of Popdays 2017. (2017)
7. Caruso, R.; de Wit, H.. Determinants of Mobility of Students in Europe. Empirical Evidence for the period 1998-2009. Journal of Studies in International Education, vol 19, n. 3, 265-282. Sage Publishing. (2015)
8. Faggian, A.; McCann, P.; and Sheppard, S.. Human Capital, Higher Education and Graduate Migration: An Analysis of Scottish and Welsh Students. Urban Studies. Volume: 44 issue: 13, 2511-2528 (2007)
9. Frumento, P., Bottai, M., Parametric modeling of quantile regression coefficient functions, Biometrics, 72, 74-84 (2015)
10. Frumento, P. (2017). qrcm: Quantile Regression Coefficients Modeling. R package version 2.0, https://CRAN.R-project.org/package=qrcm.
11. Kahanec, M.; and Kralikova R.. Higher Education Policy and Migration: The Role of International Student Mobility. CESifo DICE Report 9(4):20-27 (2011)
12. Ordine, P.; Rose, G.. StudentsâĂŹ mobility and regional disparities in quality and returns to education in Italy. Giornale degli Economisti e Annali di Economia, vol. 66, n. 2, 149-176 (2007)
13. SÃă, C., Florax, R.J.G.M.; Rietveld, P.. Determinants of the regional demand for higher education in the Netherlands: A gravity model approach. Regional Studies, 38, 375-392 (2004)
14. Sottile, G., Adelfio, G. (2017). clustEff: Clusters of Effects Curves in Quantile Regression Models. R package version 0.1.2 GPL (General Public Licence): https://CRAN.R-project.org/package=clustEff
15. Sottile, G., Adelfio, G. Clusters of effects curves in quantile regression models. Submitted (2018)
16. Van Bragt, C. A. C.; Bakx, A. W. E. A.; Bergen, T. C. M.; and Croon, M. A.. Looking for students personal characteristics predicting study outcome. Higher Education 61, 59-75. (2011)

A Partial Ordering Application in Aggregating Dimensions of Subjective Well-being

Una applicazione dell'ordinamento parziale nella sintesi di dimensioni del benessere soggettivo

Paola Conigliaro

Abstract Starting from the interest for the relationships between labour status and subjective well-being, we focused analysis upon European Union Statistics on Income and Living Conditions 2013 ad-hoc module on subjective well-being. Following the OECD Guidelines on Measuring Subjective Well-being (2013), we assumed the multidimensionality of well-being as a premise. In order to preserve the multidimensionality of the concept of SWB, we applied an aggregating method based on partial ordering. This note compare results of Partial Ordering methodology with results of other aggregating methods.

Abstract *Muovendo dall'interesse per il rapporto tra stato lavorativo e benessere soggettivo, questo studio si concentra sul modulo ad-hoc sul benessere soggettivo dell'Indagine europea sul reddito e sulle condizioni di vita (Eu-SILC) 2013. In accordo con le Linee guida OCSE sulla misurazione del benessere soggettivo (2013), assumiamo in premessa la multidimensionalità del benessere soggettivo. Per preservare tale multidimensionalità, è stato adottato un metodo di aggregazione delle dimensioni del benessere basato sull'ordinamento parziale. Questo contributo confronta i risultati della metodologia di Ordinamento Parziale con i risultati di altri metodi di aggregazione.*

Key words: Subjective Well-being indicators, Decent work, Partial Ordering Methodology.

1 Introduction – Field of Study

The principles of Decent work presently inspire the struggle against poverty, and the promotion of an equitable, inclusive and sustainable development (ILO, 2017). They lie on three main levels: universal rights, job quality, and subjective well-being in relation to work. Macro and micro social conditions influence the relationships between work and subjective well-being.

The present work originates from a study on the relationships between labour status and subjective well-being. It analyses data from the European Union Statistics on Income and Living Conditions 2013, with respect to the Italian dataset. Eurostat (2013) adopted, in fact, in that edition of Eu-SILC an ad-hoc module on subjective well-being. The module is inspired by the Guidelines on Measuring Subjective Well-being (OECD, 2013). The Guidelines, summarising main literature in term of subjective well-being indicators, identified three main dimensions of subjective well-being: cognitive, affective and eudaimonic.

1.1 The three main dimensions of Subjective Well-being

The cognitive dimension is revealed with questions about the satisfaction for life as a whole, based on assessment scales, usually of seven or eleven values. The basis of Life Satisfaction assessment is the comparison between an individual's past conditions, their ambitions and their performance in contrast to other people. (Michalos, 1985). Satisfaction of life as a whole is commonly considered the synthesis of the entire domain of satisfaction.

Emotional status refers to affects, which can be positive (trust, joy, happiness) or negative (worry, fear). In the well-being evaluation, most of the tools adopted to measure the emotional status come from mental health measurements scales. The five questions of SF-36 questionnaire (those concerning mental health status) or the WHO-5 are the most widespread. Usually questions on emotional status refers to the last 4 weeks. Studies demonstrated that, in the short term, positive and negative affects lay on a logical continuum, while in the long term, they may result as independent dimensions (Diener, 1984). To measure emotional status, Eu-SILC module applied the five Items on Mental Health of the SF-36 Questionnaire.

The third relevant dimension of well-being is Eudaimonia. The word refers to the concept used by Aristotle. Even if translated as happiness, it does not belong to the hedonic related terms. It denotes a sense of purpose, corresponding to a good. psychological functionality that goes beyond conscious evaluation or emotional feeling; it mostly regards self-realization, termed flourishing (Diener et al, 2009, Huppert and So, 2013). The relationships between working conditions and subjective well-being appear conceptually connected with this last dimension. The Eu-SILC module adopted as proxy of eudaimonic well-being a sole question on Meaning of life.

OECD in above mentioned Guidelines recognizes the relevance of the three dimensions, which refers to non-elementary concepts. It also suggests to choose indicators able to represent the multidimensionality of the concepts. The patterns of analysis have to respect this multidimensionality, and the choice of data processing methods has to conserve the informative potential (Maggino, 2015)

2 Data analysis

Eu-SILC is a sample statistical survey, which EU Member States have conducted since 2004, according to EC Regulation n.1177/2003. It allows cross-sectional and longitudinal comparison within and between countries.

The well-being ad-hoc module (2013) consists of 22 subjective items: nine questions on satisfaction (0-11 scale); one question on meaning of life (0-11); five questions on emotional status (a five-step scale); four questions on trust (0-11); two questions on personal relationships (binary variables); one question on physical security perception (a five-step scale).

Eurostat (2015) formulated its analysis comparing national aggregate data. Furthermore, according to specific knowledge needs, Eurostat chose to consider Life Satisfaction as the main dimension of subjective well-being. The results confirmed the common assumptions which recognize a relationship between labour status and subjective well-being.

The present analysis is applied to the Italian micro-data, because the micro-level analysis can offer further research perspectives. As the well-being module did not allow proxy answers, the response rate was below 70% for Italy, with a non-respondent share of 72.7% between people under-26. We chose to select individuals between 26 and 65 years (less than 16 thousands of respondents).

3 Some results

In a previous work (Conigliaro, 2018) we explored the micro-data, analysing the relationships between Self Defined Labour Status (SDLS) and three dimensions of subjective well-being. The three dimensions considered were Life Satisfaction (0-10), Meaning of Life (0-10), and the composite dimension of Emotional Status (ES).

The most delicate choice concerned the processing of the five items revealing ES. The five items involved belong to the mental health trait questions of the SF-36 Questionnaire. To calculate the mental health score, the authors of the questionnaire calculate the sum of the five scores (Ware et al., 1993). In the Eurostat report the so called "Mental well-being" index is instead computed by averaging the five-scale scores, recoding them into a range 0-100 (Eurostat, 2016). There are many other aggregating methods (e.g. Diener et al, SPANE, 2009); scholars consider feasible to aggregate in a sole measure values from different items revealing emotions.

In order to minimize intervention on data, we calculated the average value between items, based on a five-step scale, and carried the results into a five-step scale. The following Table 1 reports the correlation between ES and its components.

Table 1. CORRELATION BETWEEN EMOTIONS' ITEMS AND CALCULATED EMOTIONAL STATUS

Kendall's Tau-b	Nervous	Down in the dumps	Depressed	Peaceful	Happy
Emotional status	.686**	.684**	.726**	.689**	.636**

** Correlation is significant at the 0.01 level (2-tailed). Sig. 0.000. Number of records 15,281

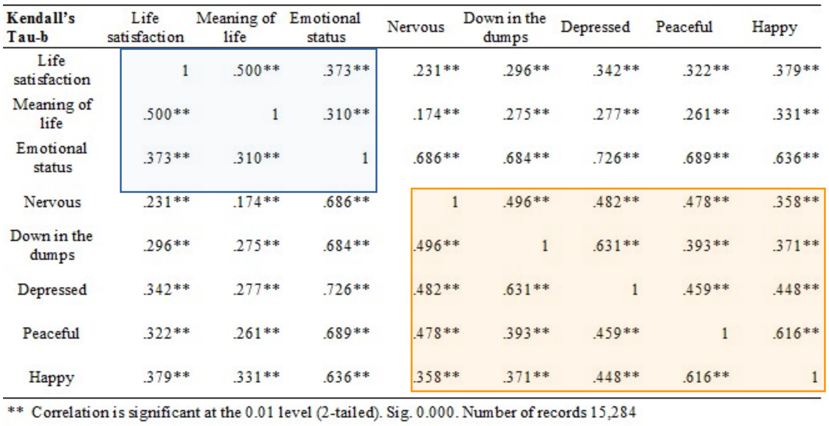
We compared the distribution of respondents according to the three dimensions of subjective well-being between different Labour status. The item that we adopted concerned the self-defined labour status (SDLS). It was expressed in 10 modalities.

We observed that:

- Permanently disabled, inactive, and unemployed have lower levels of Life Satisfaction (LS) and ES
- Meaning of Life (MoL) values are always higher for all respondents
- Each dimension of SWB has a different relationship with SDLS. Labour Status, in fact, seems to influence ES rather than MoL
- The relationship between SDLS and SWB assumes several nuances; these arise from the conjunction of some discriminants, e.g. gender, age, education, which influence that relationship. For example, MoL is lower for women when they are unemployed, but differences form employees are less evident when compared to that registered for men.

We also observed that in the whole sample (regardless SDLS) there is a low correlation between the three dimensions of SWB (Table 2.

Table 2- CORRELATION BETWEEN LIFE SATISFACTION, MEANING OF LIFE, EMOTIONAL STATUS AND FIVE AFFECTS



Kendall's Tau-b	Life satisfaction	Meaning of life	Emotional status	Nervous	Down in the dumps	Depressed	Peaceful	Happy
Life satisfaction	1	.500**	.373**	.231**	.296**	.342**	.322**	.379**
Meaning of life	.500**	1	.310**	.174**	.275**	.277**	.261**	.331**
Emotional status	.373**	.310**	1	.686**	.684**	.726**	.689**	.636**
Nervous	.231**	.174**	.686**	1	.496**	.482**	.478**	.358**
Down in the dumps	.296**	.275**	.684**	.496**	1	.631**	.393**	.371**
Depressed	.342**	.277**	.726**	.482**	.631**	1	.459**	.448**
Peaceful	.322**	.261**	.689**	.478**	.393**	.459**	1	.616**
Happy	.379**	.331**	.636**	.358**	.371**	.448**	.616**	1

** Correlation is significant at the 0.01 level (2-tailed). Sig. 0.000. Number of records 15,284

Observing the distribution of respondents within the three dimensions' space defined by the three variable of SWB (recoded in five-step scale), we registered that just the 26.3% of respondents declare the same level of subjective well-being for all dimensions. We can easily reclassify another 52.5% of respondents in an ordinal way, while for 21.2% of respondent any classification may be arbitrary. These questions could not arise in calculating average values. Applying a logical ordering of cases, we had been able to allocate more or less 75% of respondents, and to compare the distributions according to SDLS. But it was an empirical way to classify respondents. We concluded that, to preserve the multidimensionality of SWB it would be necessary to adopt a more punctual aggregating method based on logical ordering of respondents.

4 Partial ordinal methodology and first results

In a second work (Conigliaro, Alaimo, 2017) we applied a Partial Order methodology to aggregate multiple dimensions of well-being. This methodology allows, in fact, to deal with ordered variables referred to multidimensional information concerning complex phenomena.

A partially ordered set (Poset) is a set X equipped with a partial order relation, that is a binary relation satisfying these properties (Fattore et al 2015):

Reflexivity $p \leq p \ \ \forall \ p \in X$
Antisymmetry $p \leq q$ and $q \leq p$, then $p=q$, p,q □ X
Transitivity $p \leq q$ and $q \leq r$, then $p \leq q$, p,q, r □ X

In this case we recalculated the synthetic index of Emotional Status applying Parsec, which is a package developed in R by A. Arcagni and M. Fattore (2014).

We saw that the correlation between the five items concerning affects, were not so high. This result and other observations on conjoint distribution of levels of emotions, showed that there where a significant number of respondents which are both happy and sad, serene and nervous. Also the correlation between each affect and happiness was not so high, so that it could be not appropriate to adopt the happiness indicator as a proxy of emotional status, as Eurostat did (Eurostat, 2015).

The methodology adopted in those analysis was the simplest form of POSET, that compared the five dimensions of emotional status expressed in five-steps order. Data processing defined a set of 11 co-graduate levels of respondents.

In this way we assumed this as a new variable of 11 dimension that we can analyse at micro level.

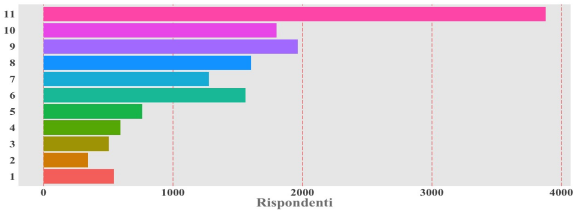


Figure 1: Number of Respondents According to ES. Aggregate Measure

Then we compared the level of ES according to SDLS.

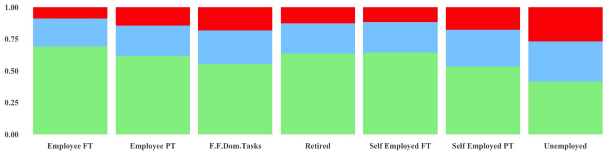


Figure 2: Level of ES According to SDLS (*percentage*)

Eventually, we calculated a well-being level applying the Partial ordering methodology to the three dimensions of subjective well-being.

The procedure adopted leads to interesting results. The Partial ordering methodology, in fact, does not hide differences between respondents that choose different response mixes.

5 Further steps of the study

The POSET methodology allows to bring ahead the analysis, defining a chain of possible answers organized in an ordinal way. It is furthermore possible to calculate the probability of error in attributing each respondent to one or another of the level of this chain.

Nevertheless, before applying higher sophisticated tools, we repute necessary to carry out a more punctual exam of measures resulting in applying this methodology.

So we intend to compare the distribution of level of subjective well-being according to different aggregative methods. We chose: 1) the average value; 2) the counting value; 3) the partial ordered value. Just to simplify the computational procedure, we decide to apply the test to a smaller number of respondents chosen with a quote sampling procedure between unemployed and full time employee.

References

1. Arcagni, A., and Fattore, M.: PARSEC: An R package for poset-based evaluation of multidimensional poverty. In: R. Bruggemann, L. Carlsen, & J. Wittmann (Eds.), MultiIndicator systems and modelling in partial order. Berlin: Springer (2014)
2. Conigliaro, P.: Labour Status and Subjective Well-being. A Micro-level Analysis and a Multidimensional Approach to Well-being. Working Papers Series of PhD Course in Applied Social Sciences (2018)
3. Conigliaro, P., and L. Alaimo: Labour Status and Subjective Well-being A Micro-level Analysis and a Multidimensional Approach to Well-being (step II applying a Partial ordering methodology). Firenze. Contribute to the AIQUAV Conference (2017)
4. Diener, Ed, and Robert A. Emmons: "The Independence of Positive and Negative Affects." In: Journal of Personality and Social Psychology 47(5): 1105-1117 (1984)
5. Diener, Ed, D. Wirtz, R. Biswas-Diener, W. Tov, C. Kim-Prieto, D. Choi, and S. Oishi "New: Measures of Well-being." In: Assessing Well-Being - The Collected Works of Ed Diener, pp.247-266. Springer International Publishing Agency (2009)
6. Eurostat: Eu-Silc Module on Wellbeing - Assessment of the Implementation. http://ec.europa.eu/eurostat/documents/1012329/1012401/2013+Module+assessment.pdf (2013)
7. Eurostat: Quality of life, Fact and Views. Luxembourg: Publications Office of the European Union (2015)
8. Eurostat: Analytical Report on Subjective Well-being. Luxembourg: Publications Office of the European Union (2016)
9. Fattore, M., F. Maggino, and A. Arcagni: "Exploiting Ordinal Data for Subjective Well-being Evaluation." In: Statistic in Transition. The measurement of Subjective Well-being in Survey Research. 16(3): 400-428 (2015)
10. Huppert, F. A., and T.T.C. So: "Flourishing Across Europe: Application of a New Conceptual Framework for Defining Well-being". Social Indicator Research 110(3): 837-861 (2013)
11. ILO: Decent Work and the 2030 Agenda for Sustainable Development. Geneva: ILO (2017)
12. Maggino, F. "Subjective Wellbeing and Subjective Aspects of Wellbeing: Methodology and Theory." In: Rivista internazionale di scienze sociali 128(1): 89-121 (2015)
13. Michalos, A. C: "Multiple Discrepancy Theory." In: Social Indicator Research 16(4):347-413 (1985)
14. OECD: Guidelines on Measuring Subjective Well-being. Paris: OECD Publishing (2013)
15. Ware, J. E., K. K. Snow, M. Kosinski, and B. Gandek: SF-36 Health Survey – Manual and Interpretation Guide. Boston: Nimrod Press (1993)

A note on objective Bayes analysis for graphical vector autoregressive models

Lucia Paci and Guido Consonni

Abstract Vector Autoregressive (VAR) models are widely used to estimate and forecast multivariate time series. However, the large number of parameters of VAR models can lead to unstable inference and inaccurate forecasts, particularly with many variables. For this reason, restrictions supported by the data are usually required. We propose an objective Bayes approach based on graphical VAR models for learning contemporaneous dependencies as well as dynamic interactions among variables. We show that, if the covariance matrix at each time is Markov with respect to the same decomposable graph, then the likelihood of a graphical VAR can be factorized as an ordinary decomposable graphical model. Additionally, using a fractional Bayes factor approach, we are able to obtain the marginal likelihood in closed form and perform Bayes graphical model selection with limited computational burden.

Key words: Bayesian model selection, decomposable graphical model, fractional Bayes factor, multivariate time series

1 Introduction

Vectore Autoregressive (VAR) models represent the workhorse models for estimating and forecasting multiple time series and widely applied in many fields such as macroeconomics, environmental sciences, neuroscience and genomic.VAR models are very flexible and allow to account for both contemporaneous dependencies among variables as well as their evolution over time. However, the large number of parameters of the VAR model usually leads to unstable inference and inaccurate

Lucia Paci
Università Cattolica del Sacro Cuore, Milano, e-mail: lucia.paci@unicatt.it

Guido Consonni
Università Cattolica del Sacro Cuore, Milano, e-mail: guido.consonni@unicatt.it

forecasts, particularly when dealing with many variables. This suggests to introduce parsimonious models.

Several solutions have been proposed in the literature. For instance, the Bayesian stochastic search variable selection approach, introduced by [9], has been extensively applied to select restrictions in VAR models. As an alternative, graphical modeling can be employed for the identification of the VAR model [6, 1, 2].

Following the latter track, we propose an objective Bayes approach for learning contemporaneous dependencies and dynamic interactions among variables under a graphical VAR model. We argue that, if the covariance matrix at each time is Markov with respect to the same decomposable graph, then the likelihood of a graphical VAR can be factorized as an ordinary decomposable graphical model. Additionally, using a fractional Bayes factor methodology, we are able to obtain the marginal likelihood in closed form and perform Bayes graphical model selection with limited computational burden.

2 Vector Autoregressive Model

Let $\mathbf{y}_t$ be a $(q \times 1)$ vector of observations collected at time t, $t = 1, \dots, T$. The reduced form of a stable VAR of order k, VAR(k), is given by

$$\mathbf{y}_t = \sum_{i=1}^{k} \mathbf{B}_i \mathbf{y}_{t-i} + \varepsilon_t, \qquad t = 1, \dots, T, \tag{1}$$

where $\mathbf{B}_i$ are $(q \times q)$ matrices of coefficients or lag matrices, determining the dynamics of the system and ε_t is a $(q \times 1)$ dimensional white noise process, that is $\varepsilon_t \mid \Sigma \sim N_q(\mathbf{0}, \Sigma)$, independently over time. Clearly, the observations depend linearly on the previous k observation vectors, where k is assumed to be known. Exogenous variables can be added to the model, leading to straightforward modifications of the results shown here. For simplicity, the intercept is also omitted in the following.

Let $\mathbf{z}_t = (\mathbf{y}'_{t-1}, \dots, \mathbf{y}'_{t-k})'$ denote the $(kq \times 1)$ vector of lagged observations at time t and $\mathbf{B} = \left(\mathbf{B}'_1, \dots, \mathbf{B}'_k\right)'$ be the $(kq \times q)$ obtained by matrix stacking the coefficients. Hence, equation (1) can be written as

$$\mathbf{y}_t = \mathbf{B}'\mathbf{z}_t + \varepsilon_t. \tag{2}$$

For given initial values $\mathbf{Y}_0 = (\mathbf{y}'_0, \mathbf{y}'_{-1}, \dots, \mathbf{y}'_{-k+1})'$, which we assume throughout to be available, the (conditional) likelihood of VAR(k) in (1) is written in the form

$$f\left(\mathbf{y}_1, \dots, \mathbf{y}_T \mid \mathbf{B}, \Sigma\right) = \prod_{t=1}^{T} f\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{B}, \Sigma\right), \tag{3}$$

where the conditional distribution $f\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{B}, \Sigma\right)$ in (3) is the multivariate normal distribution $\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{B}, \Sigma \sim \mathcal{N}_q\left(\mathbf{B}'\mathbf{z}_t, \Sigma\right)$. Let $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_T)'$ be the $(T \times q)$ matrix collecting all observations and $\mathbf{Z}$ be the $(T \times kq)$ matrix containing all the lagged variables, i.e., $\mathbf{Z} = (\mathbf{z}_t, \ldots, \mathbf{z}_T)'$. Equation (1) can be rewritten in matrix form as

$$\mathbf{Y} = \mathbf{ZB} + \mathbf{E}, \tag{4}$$

where $\mathbf{E} = (\varepsilon_1, \ldots, \varepsilon_T)'$ is the $(T \times q)$ matrix of errors following a Matrix Normal distribution with zero mean, row identity matrix $\mathbf{I}_T$ and column (or cross) covariance Σ, that is, $\mathbf{E} \mid \Sigma \sim N_{T,q}(\mathbf{0}, \mathbf{I}_T, \Sigma)$. Therefore, we can write the likelihood of VAR(k) as

$$f(\mathbf{Y} \mid \mathbf{B}, \Sigma) = (2\pi)^{-\frac{Tq}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}\left(\left(\mathbf{B}-\hat{\mathbf{B}}\right)' \mathbf{Z}'\mathbf{Z}\left(\mathbf{B}-\hat{\mathbf{B}}\right) + \hat{\mathbf{E}}'\hat{\mathbf{E}}\right)\right]\right\} \tag{5}$$

where $\operatorname{tr}(\cdot)$ is the trace operator, $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}}$ and $\hat{\mathbf{B}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$ is the OLS estimator of the coefficient matrix, requiring $T \geq kq$. In other words, the likelihood of the VAR model can be expressed as the likelihood of a multivariate regression model where the predictors are the lagged variables.

3 Fractional Bayes inference

To complete the Bayesian specification of the VAR(k) model in (1), a prior distribution for model parameters $\mathbf{B}$ and Σ is required. In this work, we employ an objective approach for model selection based on the Fractional Bayes Factor (FBF), see [11]. The idea of the FBF is to train a noninformative prior using a small fractional power b of the likelihood function, converting the noninformative prior into a proper prior that is then used to compute the marginal likelihood using the complementary fraction power $1-b$ of the likelihood.

Starting with a improper prior $p^D(\mathbf{B}, \Sigma) \propto |\Sigma|^{-(a_D+q+1)/2}$ and setting $b = T_0/T$, we can show that the fractional prior for VAR(k) is the Matrix Normal-Inverse Wishart $\mathscr{MNIW}(\underline{\mathbf{B}}, \underline{\mathbf{C}}, \nu, \underline{\mathbf{R}})$, where $\underline{\mathbf{B}} = \hat{\mathbf{B}}$, $\underline{\mathbf{C}} = T/T_0\,(\mathbf{Z}'\mathbf{Z})^{-1}$, $\nu = a_D - kq + T_0$ and $\underline{\mathbf{R}} = T_0/T\;\hat{\mathbf{E}}'\hat{\mathbf{E}}$. Hence, the fractional prior density is given by

$$p^F(\mathbf{B}, \Sigma) = K\left(\underline{\mathbf{C}}, \underline{\mathbf{R}}, \nu\right) |\Sigma|^{-\left(\frac{a_D+T_0+q+1}{2}\right)} \exp\left\{-\frac{T_0}{2T} tr\left[\Sigma^{-1}\left(\left(\mathbf{B}-\hat{\mathbf{B}}\right)'\left(\mathbf{Z}'\mathbf{Z}\right)\left(\mathbf{B}-\hat{\mathbf{B}}\right) + \hat{\mathbf{E}}'\hat{\mathbf{E}}\right)\right]\right\}, \tag{6}$$

where

$$K\left(\underline{\mathbf{C}}, \underline{\mathbf{R}}, \nu\right) = (2\pi)^{-kq^2/2} |\underline{\mathbf{C}}|^{-q/2} |\underline{\mathbf{R}}/2|^{\nu/2} \Gamma_q(\nu/2)^{-1} \tag{7}$$

is the normalizing constant with $\Gamma_q(\nu/2)$ the q-dimensional gamma function evaluated at $\nu/2$. Prior (6) is proper under two conditions: i) $a_D + T_0 - kq + 1 > q$ so that $\nu > q-1$; ii) $T - kq > q - 1$ so that $\hat{\mathbf{E}}'\hat{\mathbf{E}}$ is (almost surely) positive define. The first condition becomes $T_0 > q + kq - 1$ if $a_D = 0$, or $T_0 > kq$ if $a_D = q-1$, i.e., a larger T_0 is needed in the case $a_D = 0$. Since b has to be minimal, [5] recommend to set $a_D = q-1$ and $T_0 = kq+1$ such that $\nu = q$. The second condition simplifies to $T > q+kq-1$ that looks realistic when dealing with long time series.

Combining prior (6) with likelihood (5) we obtain a posterior distribution that is a Matrix Normal-Inverse Wishart with updated parameters, i.e., $\mathscr{M}\mathscr{N}\mathscr{I}\mathscr{W}(\overline{\mathbf{B}}, \overline{\mathbf{C}}, \overline{\nu}, \overline{\mathbf{R}})$, where $\overline{\mathbf{B}} = \hat{\mathbf{B}}$, $\overline{\mathbf{C}} = (\mathbf{Z}'\mathbf{Z})^{-1}$, $\overline{\nu} = a_D - kq + T$ and $\overline{\mathbf{R}} = \hat{\mathbf{E}}'\hat{\mathbf{E}}$.

Because of conjugacy, the fractional marginal likelihood of VAR(k) is available in closed form and can be obtained, up to a multiplicative factor, as the ratio of the prior and posterior normalizing constants, leading to

$$m^F(\mathbf{Y}) = \pi^{-\frac{(T-T_0)q}{2}} \left(\frac{T_0}{T}\right)^{\frac{(a_D+T_0)q}{2}} \left|\hat{\mathbf{E}}'\hat{\mathbf{E}}\right|^{-\frac{T-T_0}{2}} \frac{\Gamma_q\left((a_D-kq+T)/2\right)}{\Gamma_q\left((a_D-kq+T_0)/2\right)}. \tag{8}$$

Let $\mathbf{Y}_J$ be the $T \times |J|$ submatrix which contains selected columns of data matrix $\mathbf{Y}$ belonging to a subset J of cardinality $|J|$ of the full set of q variables. Using the result presented in [5], we can obtain the fractional marginal likelihood $m^F(\mathbf{Y}_J)$ based on the submatrix $\mathbf{Y}_J$ by making the following substitutions in (8):

$$q \to |J|, \qquad a_D \to a_D - |\overline{J}|, \qquad \hat{\mathbf{E}} \to \hat{\mathbf{E}}_J = \mathbf{Y}_J - \mathbf{Z}\hat{\mathbf{B}}_J, \tag{9}$$

where $\overline{J}$ denotes the complementary set of J and $\hat{\mathbf{B}}_J$ is the $kq \times |J|$ submatrix of $\hat{\mathbf{B}}$ whose column contain the OLS estimates of the regression coefficients for the selected responses. To ensure positive definiteness of $\hat{\mathbf{E}}_J'\hat{\mathbf{E}}_J$, the condition $|J| < T - kq + 1$ must be satisfied, when setting $a_D = q-1$ and $T_0 = kq+1$.

4 Graphical VAR

[7] introduced the class of time series chain graphs (TSCG). More specifically, let $Y = \{Y_t(a), t \in \mathbb{Z}, a = 1, \ldots, q\}$ be a q-variate stationary stochastic process and $V = \{1, 2, \ldots, q\}$ be the set of indexes. Let $G = (V_{TS}, E)$, be a graph with $V_{TS} = V \times \mathbb{Z}$ and edge set E, whose edges have at most lag k and which is invariant under translation. If $\mathbf{B}_i(b,a)$ is the (b,a)-element of matrix $\mathbf{B}_i$ in (1) and $\Omega(a,b)$ is the (a,b)-entry of precision matrix $\Omega = \Sigma^{-1}$, then the VAR model with the following constraints on the parameters

$$\begin{array}{lcll} (a,t-i) \;\to\; (b,t) & \in E \;\Leftrightarrow\; \mathbf{B}_i(b,a) \neq 0 & & i = 1, \ldots, k \\ (a,t) \;—\; (b,t) & \in E \;\Leftrightarrow\; \Omega(a,b) \neq 0 & & t = 1, \ldots, T \end{array}$$

represents a VAR(k, G) model. Thus, a nonzero element in $\mathbf{B}$ corresponds to a directed edge in the graph reflecting the recursive structure of the time series while undirected edges specify contemporaneous interactions among variables, that is a covariance selection model. Hence, our goal is making inference on graph G.

4.1 Covariance selection model

Let $G^u = (V, E^u)$ be the undirected graph representing the contemporaneous dependencies at any time t, and assume that Σ is Markov with respect to G^u. For the graph-theory terminology in this section we refer the reader to [10]. Also, we confine our analysis to the class of decomposable graphs for all time points. Recall that G^u is decomposable when all cycles in G^u admit a chord, that is an edge joining two non-consecutive vertices of the cycle. Let $\mathscr{C}$ and $\mathscr{S}$ denote the set of cliques and separators of the graph G^u, respectively. Then, we can show that the likelihood of a graphical VAR(k,G) factorizes as following:

$$f(\mathbf{Y} \mid \mathbf{B}, \Sigma, G^u) = \frac{\prod_{C\in\mathscr{C}} f\left(\mathbf{Y}_C \mid \mathbf{B}_C, \Sigma_{CC}\right)}{\prod_{s\in\mathscr{S}} f\left(\mathbf{Y}_S \mid \mathbf{B}_S, \Sigma_{SS}\right)}, \tag{10}$$

where $\mathbf{B}_C$ and $\mathbf{B}_S$ are the $kq \times |C|$ and $kq \times |S|$ matrices whose columns contain the coefficients of the selected responses $\mathbf{Y}_C$ and $\mathbf{Y}_S$, respectively.

If $\mathbf{B}$ is unconstrained and Σ is in $M^+(G^u)$, the set of all symmetric positive-definite matrices having elements in Σ^{-1} set to zero for all $(a,b) \notin E^u$, then a natural noninformative prior on $(\mathbf{B}, \Sigma \mid G^u)$ is

$$p^D(\mathbf{B}, \Sigma \mid G^u) \propto \frac{\prod_{c\in\mathscr{C}} |\Sigma|^{-|C|}}{\prod_{s\in\mathscr{S}} |\Sigma|^{-|S|}}, \tag{11}$$

which is a limiting form of an Hyper-Inverse Wishart distribution [8]. Training prior (11) with a fraction $b = T_0/T$ of the likelihood, the fractional prior for a VAR(k,G) becomes a Matrix Normal Hyper-inverse Wishart distribution [3], $\mathscr{MNHIW}(\underline{\mathbf{B}}, \underline{\mathbf{C}}, \delta, \underline{\mathbf{R}})$, where $\underline{\mathbf{B}}$,$\underline{\mathbf{C}}$, $\underline{\mathbf{R}}$ are defined as above and $\delta = T_0 - kq$. Therefore, the prior density is

$$p^F(\mathbf{B}, \Sigma \mid G^u) = H(\underline{\mathbf{C}}, \underline{\mathbf{R}}, \delta) \\ \times \frac{\prod_{c\in\mathscr{C}} |\Sigma_{CC}|^{-(|C|+T_0/2)} \exp\left\{-\frac{T_0}{2T} \operatorname{tr}\left[\Sigma_{CC}\left(\left(\mathbf{B}_C - \hat{\mathbf{B}}_C\right)' \mathbf{Z}'\mathbf{Z}\left(\mathbf{B}_C - \hat{\mathbf{B}}_C\right) + \hat{\mathbf{E}}_C'\hat{\mathbf{E}}_C\right)\right]\right\}}{\prod_{s\in\mathscr{S}} |\Sigma_{SS}|^{-(|S|+T_0/2)} \exp\left\{-\frac{T_0}{2T} \operatorname{tr}\left[\Sigma_{SS}\left(\left(\mathbf{B}_S - \hat{\mathbf{B}}_S\right)' \mathbf{Z}'\mathbf{Z}\left(\mathbf{B}_S - \hat{\mathbf{B}}_S\right) + \hat{\mathbf{E}}_S'\hat{\mathbf{E}}_S\right)\right]\right\}}, \tag{12}$$

where the normalizing constant is

$$H(\underline{\mathbf{C}}, \underline{\mathbf{R}}, \delta) = \frac{\prod_{c\in\mathscr{C}} (2\pi)^{-|C|kq/2} |\underline{\mathbf{C}}|^{-|C|/2} |\underline{\mathbf{R}}_{CC}/2|^{(\delta+|C|-1)/2} \Gamma_{|C|}\left((\delta+|C|-1)/2\right)^{-1}}{\prod_{s\in\mathscr{S}} (2\pi)^{-|S|kq/2} |\underline{\mathbf{C}}|^{-|S|/2} |\underline{\mathbf{R}}_{SS}/2|^{(\delta+|S|-1)/2} \Gamma_{|S|}\left((\delta+|S|-1)/2\right)^{-1}}.$$

In other words, a Markovian structure is assumed a priori for the lag coefficients that follows the structure of the likelihood. Thus, using prior (12), the fractional marginal likelihood has a closed form obtained, again, as the ratio of the prior and posterior normalizing constants. Equivalently, we can write

$$m^F(\mathbf{Y} \mid G^u) = \frac{\prod_{c \in \mathscr{C}} m^F(\mathbf{Y}_C)}{\prod_{s \in \mathscr{S}} m^F(\mathbf{Y}_S)}, \tag{13}$$

where, following [4], $m^F(\mathbf{Y}_C)$ and $m^F(\mathbf{Y}_S)$ are computed by means of (9) with $J = c$ and $J = s$, respectively, when setting $a_D = q - 1$. Again, formula (9) provides a valid marginal likelihood if $T > |C| + kq - 1$, for each $c \in \mathscr{C}$.

As a final remark here, we stress that the joint likelihood of a graphical VAR(k,G) factorizes as an ordinary decomposable graph model, even though the decomposable structure is assumed conditionally at each time step. As a result, the closed form of the marginal likelihood allows to perform Bayes graphical model selection of VAR(k,G) models with easy computation.

A simulation study (not presented here for brevity) shows the capability of the approach to recover the underlying graph according to different scenarios (sample size, number of variables, number of lags, lag matrix). Future work will explore the possibility to build a joint prior model that simultaneously accounts for restrictions both on lag coefficients and covariance matrix.

References

[1] F. Abegaz and E. Wit. Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, 14:586, 2013.

[2] D. F. Ahelegbey, M. Billio, and R. Casarin. Bayesian graphical models for structural vector autoregressive processes. *J Appl Econom*, 31:357–386, 2016.

[3] C. M. Carvalho and M. West. Dynamic matrix-variate graphical models. *Bayesian Anal*, 2:69–97, 2007.

[4] G. Consonni and L. La Rocca. Objective Bayes factors for Gaussian directed acyclic graphical models. *Scand J Stat*, 39:743–756, 2012.

[5] G. Consonni, L. La Rocca, and S. Peluso. Objective Bayes covariate-adjusted sparse graphical model selection. *Scand J Stat*, 44:741–764, 2017.

[6] J. Corander and M. Villani. A Bayesian approach to modelling graphical vector autoregressions. *J Time Ser Anal*, 27:141–156, 2006.

[7] R. Dahlhaus and M. Eichler. Causality and graphical models in time series analysis. In P. Green, N. Hjort, and S. Richardson, editors, *Highly structured stochastic systems*, pages 115–137. University Press, Oxford, 2003.

[8] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann Stat*, 21:1272–1317, 1993.

[9] E. I. George, D. Sun, and S. Ni. Bayesian stochastic search for VAR model restrictions. *J Econometrics*, 142:553–580, 2008.

[10] S. L. Lauritzen. *Graphical models*. Oxford University Press, Oxford, 1996.

[11] A. O'Hagan. Fractional Bayes factors for model comparison. *J R Stat Soc B*, 57:99–138, 1995.

Bayesian Population Size Estimation with A Single Sample

Stima della numerosità di una popolazione utilizzando un solo campione

Pierfrancesco Alaimo Di Loro and Luca Tardella

Abstract The estimation of the size of a finite population is a problem encountered in a variety of applications. One standard statistical approach relies on *mark-recapture* sampling, which may require high costs and annoyance to the population of interest. These considerations have motivated the search for alternative sampling strategies that allow to estimate the size of a population from a single capture. Hettiarachchige [4] proposes a method that is viable when the population is made of only two generations: a group of generators and one of generated units. We investigate Bayesian methods alternative to the frequentist estimators used in [4]. Preliminary results give evidence of competing performance of the Bayesian approach, which in some cases sensibly outperforms the frequentist alternatives.

Abstract La stima della numerosità di una popolazione è un problema comune a vari ambiti di applicazione. Le procedure di stima sono solitamente basate sul noto metodo *cattura-ricattura*, il quale comporta elevati costi e disturbo della popolazione. Tali considerazioni hanno stimolato la ricerca di tecniche che permettano di ottenere un stima utilizzando un unico campione. Hettiarachchige [4] propone un metodo applicabile nel caso in cui la popolazione sia composta di due sole generazioni: un gruppo di unità generatrici ed uno di unità generate. L'obiettivo del nostro lavoro è quello di ottenere un'estensione Bayesiana dell'originale modello frequentista. Risultati preliminari evidenziano accuratezza degli stimatori Bayesiani sensibilmente migliore rispetto alle alternative frequentiste.

Key words: population size, capture-recapture, mark-recapture, single sample, Bayesian inference, moment estimator, MLE.

Pierfrancesco Alaimo Di Loro
University of Rome "La Sapienza", Statistical Science Department, Piazzale Aldo Moro 5, 00185 Roma RM, e-mail: pierfrancesco.alaimodiloro@uniroma1.it

Luca Tardella
University of Rome "La Sapienza", Statistical Science Department, Piazzale Aldo Moro 5, 00185 Roma RM, e-mail: luca.tardella@uniroma1.it

1 Introduction

The problem of estimating the size or any other demographic parameter of a population of interest for which there is no complete enumeration or reference list is common to a variety of applications: ecology (e.g. natural and wildlife populations), reliability, epidemiology, social sciences. However, most of the literature regarding this matter has been developed in the statistical ecology field, where *capture-recapture* methods have been the ruling paradigm for the whole second half of the 20th century.

The modern foundation of these methods was laid in [2] and [6] and they are all based on the pioneering *mark-recapture* sampling technique which originated the well-known *Lincoln-Petersen* estimator. The most basic version consists of taking a random sample of size n_1 from the population and mark the captured individuals. They are then returned to the population and, at a later occasion, a second sample of size n_2 is taken. The previously applied tags allow to recognize if and how many of the captured individuals were already been sampled at the previous occasion. If m of them already have a tag, then the *Petersen* estimator is: $\hat{N}_P = (n_1 n_2)/m$.

The biggest issue with the application of such methods is that they require the population to be sampled at least twice. The necessity of at least one further capture occasion leads to increasing costs and, furthermore, can cause an ever-increasing annoyance to the population of interest. The latter can alter its natural equilibrium, leading to a change of conditions from one capture to the other. This may introduce bias in the estimates when those changes are not taken into proper account and requires behavioural adjustments on the basic model [3] . Moreover, there are a lot of situations in which the captured individuals cannot be returned to the population, making the procedure impractical. These considerations have motivated the search for procedures that allow to estimate the size of a population in alternative ways.

In the last decades genetic data have become increasingly important in ecology and conservation biology and their use in estimating the population size have been considered [5]. The underlying idea is that the degree of biological relationship between a sample of individuals from the population provides information about the population itself and DNA profiles can be used to detect the degree of relatedness between individuals. [7] exploit this idea in analogy to the traditional capture-recapture method and argued in favor of a *single sample* version of the Petersen estimator. An individual is marked by its presence in the sample, and "recaptured" if the sample contains one or more close relatives. In practice, it allows to generalize from "recapture of self" to "recapture of closely-related kin" ([1]), where the detection of kinship is based on the idea that an individual share more alleles with a parent than with a biologically unrelated individual.

However, as it has already been underlined in [7], this approach is sensibly more complex than the ordinary *capture-recapture* method. The effectiveness of this method depends on hypotheses on the population that are hardly matched in real life and relies on accurate estimates of the kinship coefficients between individuals. The complexity of the method has been reduced by [4], who put himself in a

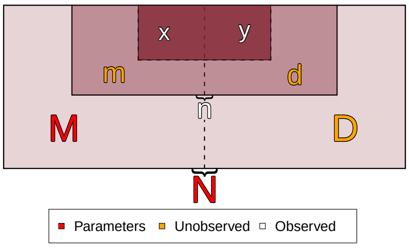


Fig. 1 Schematic visualization of the structure of population and of the quantities involved in the sample. Different colors are used to distinguish between parameters, latent variables and observed quantities.

simpler framework where kinship misclassification plays a slightly less important role. The framework of study will be more extensively investigated in Section 2.

Our contribution to [4]'s work is to introduce a suitable way of implementing Bayesian methods (Section 3) alternative to the frequentist estimators used by the original author. A comparative analysis of these estimators is provided at the end of the section. We finally provide an outline of some promising developments and extensions that may improve on the precision of the proposed estimates in Section 4.

2 General Framework

Let us introduce formally the assumptions of the model. The population structure considered in [4] is composed of only two generations: the individuals from the first generation are denoted as *mothers* and the individual from the second as *daughters*. For the purpose of this paper our main interest will be in estimating the size of the population of the *mothers*.

The two generations are assumed mutually exclusive and collectively exhaustive, and the population to be closed. A sample is taken and perfect identifiability of mother-daughter couples is assumed. The appropriateness of this assumption is discussed in [4].

In practice, we are dealing with a random sample from a population of N individuals, where M are "mothers" and D are "daughters" ($N = M + D$). The captured individuals will be in part mothers and in part daughters. We are able to recognize

	Latent		**Observables**
		$n = m + d$	all units in the sample
m	all mothers in the sample	x	all *identified* mothers in the sample
$(d_i)_{i=1}^{m}$	daughters for each mother in the sample	$(y_i)_{i=1}^{x}$	*id.* daughters for each *id.* mother in the sample
$d = \sum_{i=1}^{m} d_i$	all daughters in the sample	$y = \sum_{i=1}^{x} y_i$	all *identified* daughters in the sample

Table 1 Quantities of interests involved in the single sample

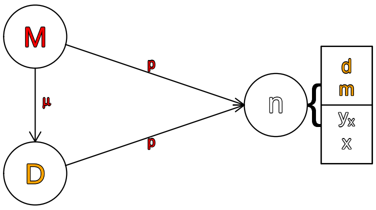


Fig. 2 Graphical visualization of the generating process of the data. Parameters in *red*, latent variables in *orange* and observables in *white*.

as daughters only those daughters whose mothers have been captured and viceversa. The relevant quantities involved in the single sample are visualized in Figure 1 and listed in Table 1 with appropriate notation:

Probabilistic Model. It is assumed that each mother has generated, at a previous time, a certain number of daughters D_i according to a $Pois(\mu)$. The total number of daughters in the population is then $D = \sum_{i=1}^{M} D_i \sim Pois(M\mu)$. The parameter μ is constrained to the set $[1, +\infty)$ by [4] for reasons related to the existence of the moment estimator. Our Bayesian approach, theoretically, does not require such an assumption but we will stick to this constraint to ensure a fair comparison. Furthermore, each individual is supposed to be captured independently with equal probability p.

The total number of mothers in the population M is the parameter of interest, while μ and p are just nuisance parameters. The situation is graphically reported in Figure 2.

An explicit form of the *marginal* likelihood can be obtained using a conditioning argument:

$$P(n,x,(y_i)_{i=1}^{x}|M,\mu,p) = \sum_{m=x}^{M\wedge(n-y)} P(n,x,(y_i)_{i=1}^{x}|m,M,\mu,p)P(m|M,p). \qquad (1)$$

$P(m|M,p)$ is the probability to capture independently m mothers given that there are M mothers in the population and they are captured with probability p[1], which is:

$$P(m|M,p) = Bin(m|M,p) = \binom{M}{m} p^m (1-p)^{M-m}.$$

The joint density of $(n,x,(y_i)_{i=1}^{x})$ conditioned on (m,M,μ,p), and hence the likelihood of the model, can be shown to be equal to:

$$P\left(n,x,(y_i)_{i=1}^{x}|\, m,\, M,\mu,p\right) = \binom{m}{x} \frac{e^{-M\mu p}(\mu p)^y}{\prod_{i=1}^{x} y_i!} \frac{((M-m)\mu p)^{(n-m-y)}}{(n-m-d)!}.$$

[1] This probability obviously does not depend on the mean number of daughters μ per mother.

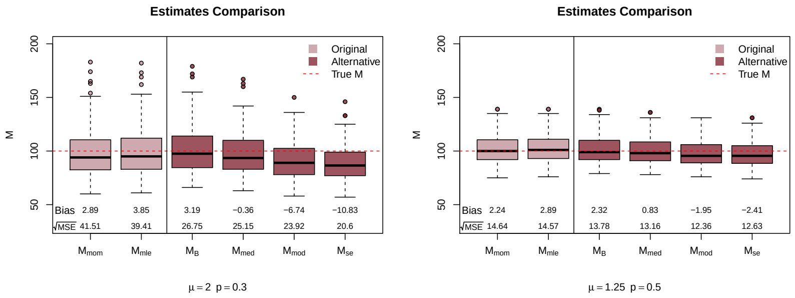


Fig. 3 Box-plots of the resulting estimates obtained for the 100 different simulation set for each estimator. From left to right: *Moment* and *Maximum Likelihood* estimators, posterior *Mean*, *Median*, *Mode* and *PMSE minimizer*

3 The Bayesian Extension

Some undesirable peculiarity of the likelihood such as presence of non-unique solutions and computational problems led [4] to discard the maximum likelihood estimator. The author decided to resort to a moment-based estimator in order to overcome these problems. However, the method of moments is still affected by well-known boundary solution instability.

We propose an estimation procedure based on a Bayesian approach, which should be able to regularize the model likelihood without incurring in the moment estimator deficiencies. Furthermore, Bayesian methods allow to include prior information on the demographic parameters of the population whenever such information is available and can eventually replace in a sensible way the *ad-hoc* constraint on μ.

Independent priors have been assigned to the parameters M, μ and p, so that the posterior distribution can be written as:

$$\pi\left(M,\mu,p|n,x,(y_i)_{i=1}^{x}\right) \propto \mathscr{L}(M,\mu,p;n,x,(y_i)_{i=1}^{x})\pi(M)\pi(\mu)\pi(p)$$

The forms chosen for the priors of each parameter are:

- low-informative $TruncatedGamma(0.05, 0.025)$ and $Beta(0.001, 0.001)$ priors for μ and p;
- inverse prior $\pi(M) \propto \frac{1}{M^a}$ for M, with $a = 0, 1, 2$.

Posterior samples from the joint posterior of the parameters have been obtained via *Metropolis-Within-Gibbs* algorithm, where a Gibbs-style update is performed for M and a bi-variate Normal random walk M-H is performed for (μ, p). The considered Bayesian estimators are: the posterior mean $\hat{M}_B$, the posterior median $\hat{M}_m$ and the approximated PMSE minimizer[2] $\hat{M}_{se}$.

[2] The *Posterior Mean Squared Error* minimizer is the value that minimizes the MSE with respect to the posterior distribution of the quantity of interest: $\text{argmin}_{a\in\mathscr{M}} \sum_{M\in\mathscr{M}}(M-a)^2\pi(M|\cdot)$, where M is the quantity of interest and $\mathscr{M}$ its domain.

Preliminary Results. A simulation study has been carried out in order to verify the effectiveness of the proposed Bayesian estimators. For different configurations of μ and p, with fixed $M = 100$ mothers, the simulation of mother-daughter sampling have been replicated for $S = 100$ times, producing $S = 100$ realizations of all the alternative estimators. The comparative performance is assessed in terms of *Mean Squared Error*. Our preliminary results are exposed in Figure 3 and give evidence of competing performance of the Bayesian approach which, especially in some configurations, sensibly outperforms the frequentist alternatives.

4 Concluding remarks and further Developments

We have proposed a Bayesian framework for the estimation of the population size in presence of a single sample. This technique relies on the pairing of mothers and daughters in the sample through the use of genetic markers on the line of [4]. We have shown that the Bayesian framework allows to reduce the error of the classical estimators up to 50% in specific parameter settings. Indeed, we are currently workin on many other improvements and extensions to the proposed Bayesian methodology:

1. formal derivation of non-informative priors and principled informative priors possibly removing the unnatural *ad-hoc* constraint $\mu > 1$;
2. inclusion of other kind of kinships and/or other covariates in order to reduce the variability of the unidentified part of the sample;
3. relaxation of restrictive model assumptions like the identical capture probability and the perfect identification.

References

1. Bravington, M.V. and Skaug, H.J. and Eric, C.: Close-kin mark-recapture. Statistical Science 31 (2), 259 – 274 (2016), Institute of Mathematical Statistics
2. Cormack, R.M.: Estimates of survival from the sighting of marked animals. Biometrika 51 (3/4), 429 - 438 (1964), JSTOR
3. Fegatelli, D.A. and Tardella, L.: Improved inference on capture recapture models with behavioural effects. Statistical Methods & Applications 22 (1), 45 - 66 (2013), Springer
4. Hettiarachchige, C.K.H.: Inference from single occasion capture experiments using genetic markers. PhD Thesis (2016)
5. Schwartz, M.K. and Tallmon, D.A. and Luikart, G.: Review of DNA-based census and effective population size estimators. Animal Conservation forum 1 (4), 293 - 299 (1998), Cambridge University Press
6. Seber, G.A.F.: A note on the multiple-recapture census. Biometrika 52 (1/2), 249 - 259, (1965), JSTOR
7. Skaug, H.J.: Allele-sharing methods for estimation of population size. Biometrics 57 (3), 750 – 756 (2001), JSTOR

Classification of the Aneurisk65 dataset using PCA for partially observed functional data

Classificazione del dataset Aneurisk65 utilizzando la PCA per dati funzionali parzialmente osservati

Marco Stefanucci, Laura Sangalli and Pierpaolo Brutti

Abstract When functional data are observed over a domain that is subject-specific, most of the techniques for functional data analysis are invalidated. Recently, new methods able to handle this situation were developed and in particular we focus on well-known functional PCA. With the aim of classifying the Aneurisk65 dataset, we apply a few possible methods and we show that carrying out the analysis over the full domain, where at least one of the functional data is observed, may not be the optimal choice. This is also confirmed in a simulation study, where the best interval for classification lies between the common domain and the full domain.

Abstract *Ogniqualvolta dei dati funzionali vengono osservati su un dominio dipendente dal soggetto considerato, non è più possibile utilizzare la gran parte delle tecniche per l'analisi di dati funzionali. Recentemente sono stati sviluppati nuovi metodi in grado di affrontare questa situazione e noi tratteremo la ben nota ACP funzionale. Con lo scopo di classificare il dataset Aneurisk65, abbiamo applicato diversi metodi e mostreremo che eseguire l'analisi sul dominio completo, dove è osservato almeno uno dei dati funzionali, può non essere la scelta ottimale. Ciò è confermato anche da uno studio di simulazione dove il migliore intervallo per la classificazione giace tra il dominio comune e il dominio completo.*

Key words: Functional Data, Partially Observed Data, Classification, Functional PCA

Marco Stefanucci
Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma e-mail: marco.stefanucci@uniroma1.it

Laura Sangalli
Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano e-mail: laura.sangalli@polimi.it

Pierpaolo Brutti
Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma e-mail: pierpaolo.brutti@uniroma1.it

In the last 20 years, functional data analysis emerged as one of the fastest growing fields of modern statistics [1]. Functional data can be viewed as realizations of a functional objects, and in general are collected as discrete and noisy observations. The domain where the data are observed is usually common to all the statistical units, while the observation grid may vary across units. When the domain is unit-dependent, standard procedures are invalidated and a practical solution is to restrict the analysis to the intersection of the domains, where all the curves are observed. However, including intervals where the data are only partially observed may result in a more powerful statistical analysis and in the last decade some authors developed methodology handling this particular situation.

The Aneurisk65 dataset is an important example of partially observed functional data. The data consist in the profiles of radius and curvature of the internal carotid artery of 65 subjects suspected to be affected by cerebral aneurysms [2] and are displayed in figure 1. The domain where all the curves are available is highlighted in light-gray. Outside this domain individual observations are progressively lost when moving towards the full domain. In this application it is relevant to investigate whether the morphology of the internal carotid artery influences aneurysms pathogenesis and this can be done considering that the data can be divided in two groups - displayed in orange and blue in the figure - depending on the presence and location of the cerebral aneurysms.

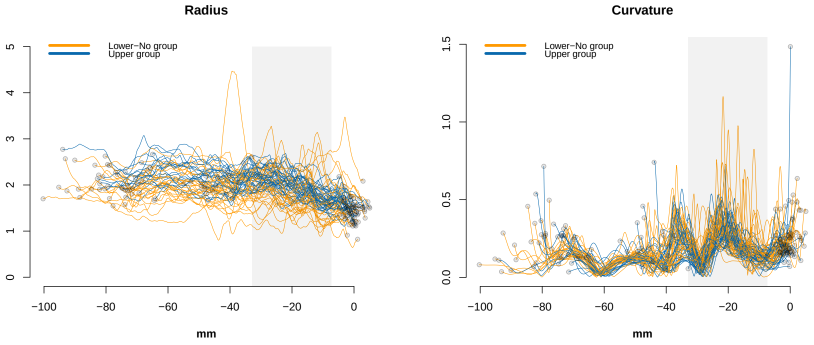


Fig. 1 The Aneurisk dataset.

In [2] a discriminant analysis based on the scores of the principal components of the radius and curvature profiles is presented, restricting the attention to the portion of the domain common across subjects. Our contribution is to improve the discrimination results by considering also portions of the domain where not all data are observed. In doing this, we need a methodology for functional principal components analysis in the case of partially observed data. Here we consider some proposals: the first one [3] is based on a spline reconstruction of the eigenfunctions, the second [4] is a method that estimates the principal component scores using conditional expectations, the third [5] is a generalization of the power method for extracting eigenvectors of a given matrix and the last one [6] is a fully functional approach in which the missing part of each score is predicted via best linear approximation of its conditional expectation.

We then use these methods to perform a discrimination based on the scores of the first K principal components. We show that carry out the analysis on the largest possible domain may not be the best choice for classification purposes. In fact, we suggest to explore different intervals, ranging from the common domain to the full domain. More specifically, we divide the domain where the data are partially observed in L portions, and we consider a collection of progressively larger domains I_l for $l \in \{0,\dots,L\}$, with $I_{l-1} \subset I_l$, where I_0 is the common domain and I_L is the full domain. We select the optimal number of principal components and the optimal domain extension I_l via cross-validation. For all the methods, the best interval for classification is between the common domain and the full domain and such an interval is the same for all the approaches, see figure 2. In this application, the approach [5] achieves the best result, corresponding to a 9 misclassified subjects, thus outperforming the result in [2]. Note that the application of the methodologies for partially observed data on the full domain does not lead to any improvement in the discrimination.

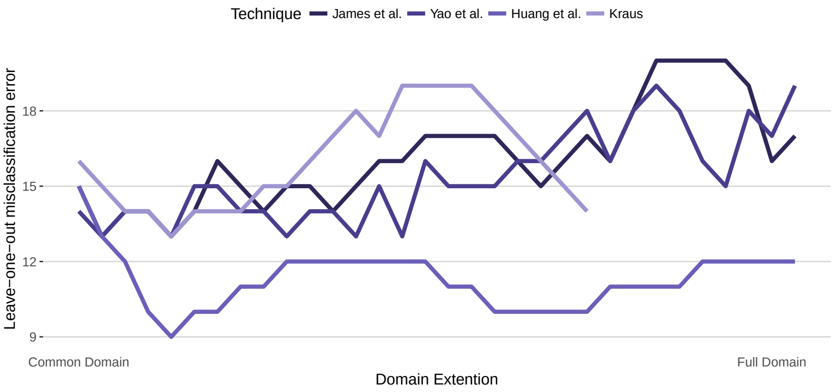


Fig. 2 Performances on the Aneurisk dataset.

In order to understand better the behaviour of the domain extension technique, we carry out simulations. We generate functional data over the interval $I_L = [0,1]$. We then completely retain the data generated over the interval $I_0 = [1/3, 2/3]$, while we censor them over the intervals $I_{\mathtt{left}} = [0,1/3]$ and $I_{\mathtt{right}} = [2/3,1]$, by sampling the starting point of each functional datum uniformly over $I_{\mathtt{left}}$, and its ending point uniformly over $I_{\mathtt{right}}$. The data are generated from a cubic B-splines basis with 16 internal knots, corresponding to a total of 20 bases. We generate two groups of 50 units each with the only difference being the mean function, see figure 3. The 100

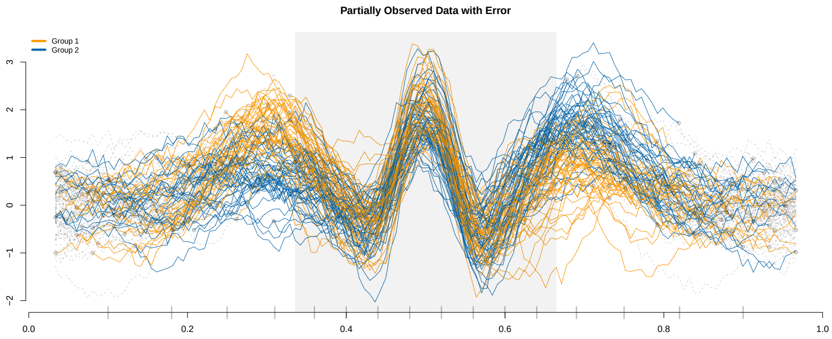


Fig. 3 Simulated data.

generated curves are evaluated on a regular grid of 150 points in $[0,1]$, contaminated with gaussian noise and finally classified over a set of progressively larger domains. This simulation is repeated 50 times.

In order to compare the performances of the four methods, we also apply standard PCA to the fully observed data. The results are presented in figure 4. The figure displays the boxplots of the leave-one-out misclassification error, for various tech-

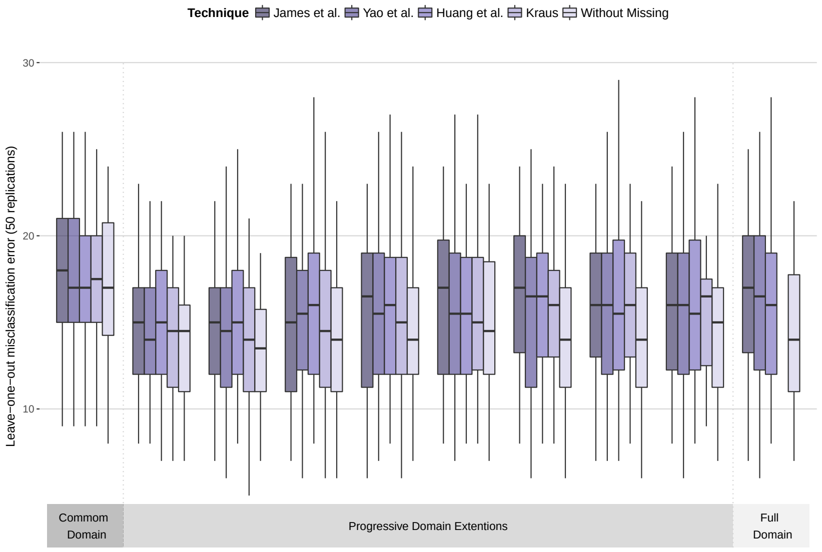


Fig. 4 Simulation results.

niques, for various domain extensions. For all methods, the misclassification error decreases when we start extending the domain with respect to the common domain, but then progressively increases as we approach the full domain. None of the methods outperforms the other. In conclusion, extending the domain with respect to the common domain improves the discrimination between the two groups; on the other hand, larger domain extensions, and in particular the full domain, do not lead to the best discrimination results.

References

1. Ramsay, J. and Silverman, B.W.: Functional Data Analysis. Springer (2005)
2. Sangalli, L.M., Secchi, P., Vantini, S. and Veneziani, A.: A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. Journal of the American Statistical Association **104**, 37-48 (2009)
3. James, G.M. and Hastie, T.J.: Principal component models for sparse functional data. Biometrika **87**, 587-602 (2000)
4. Yao, F., Müller,H.G. and Wang, J.L.:Functional Data Analysis for Sparse Longitudinal Data. Journal of the American Statistical Association **100**, 577-590 (2005)
5. Huang, J.Z., Shen, H., and Buja, A. : Functional principal components analysis via penalized rank one approximation. Electronic Journal of Statistics **2**, 678–695 (2008)
6. Kraus, D.: Components and completion of partially observed functional data. Journal of the Royal Statistical Society: Series B **77**, 777–801 (2015)

Deep Learning to the Test: an Application to Traffic Data Streams

Metodologie di apprendimento approfondito applicate a dati di traffico stradale

Nina Deliu and Pierpaolo Brutti

Sommario Deep learning is a broad class of machine learning techniques based on learning data representation through multiple levels of abstraction. It has been successfully applied in several areas of research, but very few literature addressed the problem of traffic flow forecasting. Thus, driven by the belief that deep learning algorithms may capture the sharp traffic data non-linearities, we aimed to develop a deep architecture, namely a feed-forward neural network, and evaluate its performances in predicting short-term traffic streams. We illustrate our methodology, consisting in a predictors selection step and a subsequent training step, using traffic speed data of the Grande Raccordo Anulare road of Rome for the month of June 2016.

Sommario *Gli algoritmi di apprendimento approfondito costituiscono una vasta classe di algoritmi machine learning basati sulla rappresentazione dei dati tramite molteplici livelli di astrazione. Essi sono stati applicati con successo in diverse aree di ricerca, tuttavia solo una piccola parte della letteratura deep learning si è interessata al problema della previsione di dati di traffico. Considerando la capacità di questi algoritmi di catturare non linearità presenti all'interno dei dati, ci siamo proposti di sviluppare un'architettura deep learning per prevedere il traffico a breve termine. Illustriamo la nostra metodologia, consistente in una fase di selezione dei predittori e una di apprendimento della rete, considerando una dataset di dati di velocità di veicoli sulla rete autostradale Grande Raccordo Anulare di Roma.*

Nina Deliu
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale A. Moro, 5 00185, Roma, Italy e-mail: nina.deliu@uniroma1.it

Pierpaolo Brutti
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale A. Moro, 5 00185, Roma, Italy e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction

Deep learning is a rather wide class of machine learning techniques based on learning data representation through multiple levels of abstraction. These methods have dramatically improved the state of the art in computer vision, natural language processing, speech recognition, object detection and many other domains such as drug discovery and genomics. However, according to the complexity of data, the prediction problem may be more or less difficult to perform. For instance, when aiming to model a time series, one should take into account the sequence dependence among the input values and handle the underlying temporal dynamics. While some areas relying on time series are well exploited and widely modelled with deep neural networks (e.g., speech recognition or financial applications [1]), some others are still poorly addressed, especially in the Italian research literature. One of these, is represented by road traffic modeling, which is a challenging and increasingly relevant field of research, as real-time forecasting may give travelers the ability to choose better routes and authorities the ability to manage the transportation system. Driven by the belief that deep learning algorithms might be able to capture nonlinear spatio-temporal effects in recurrent and non-recurrent traffic patterns, we aimed to develop a deep architecture and to asses if it may outperform standard modelling approaches in predicting short-term traffic streams.

2 Methods and Data

Deep learning algorithms, also known as deep artificial neural networks (ANNs) are statistical models directly inspired by biology, more precisely by the neural network of the human brain [2]. An ANN is organized as a graph, whose nodes or units are structured in multiple layers and interconnected by links to propagate activation from the initial to final units. Each link has a weight that determines the relative strength and sign of the connection, and each unit applies an activation function to the weighted sum of all incoming activations. The quintessential and most common deep learning architecture is the feed-forward neural network (FFNN), and it can be viewed as a directed acyclic graph [3].

The model: feed-forward neural network

Feed-forward neural networks are represented as a composition of many different non-linear functions, with the characteristic of flowing information hierarchically from the initial inputs X to all the neurons of the network through intermediate computations.

Formally, a deep feed-forward architecture can be described as follows. Let's suppose our network is composed by L hidden layers indexed as $l = 1, \dots, L$. Let $Z^{(l)}$ denote the vector of inputs into a generic layer $l+1$, $Z^{(0)}$ the vector of original inputs X and $Z^{(L+1)}$ the output vector of predictions $\hat{Y}(X)$. Using a matrix-vectorial notation, the model is given by

$$\begin{aligned} X &= Z^{(0)} \\ Z^{(1)} &= f^{(1)}\left(W^{(1)}Z^{(0)} + b^{(1)}\right) \\ \dots & \\ Z^{(L)} &= f^{(L)}\left(W^{(L)}Z^{(L-1)} + b^{(L)}\right) \\ \hat{Y}(X) &= Z^{(L+1)} = f^{(L+1)}\left(W^{(L+1)}Z^{(L)} + b^{(L+1)}\right). \end{aligned}$$

where $f^{(1)},\dots,f^{(L+1)}$ are the non-linear activation functions, $W^{(1)},\dots,W^{(L+1)}$ the weight matrices and $b^{(1)},\dots,b^{(L+1)}$ the bias or activation levels.
In summary, a feed-forward algorithm applies $L+1$ non-linear transformations to the input data X. Each transformation takes as input the output of the previous transformation, and gives a new abstract representation of the data.

Dataset

This study is based on a dataset of vehicle traffic streams obtained from QMap S.r.l. company, specialized in Intelligent Transport System and Info-mobility. All data were acquired by Global Positioning System (GPS) probes, which typically contain information on latitude, longitude, altitude, heading, speed and, most importantly, precision. This technology offers a low capital cost, a low installation cost, and a low data collection cost combined with a high location accuracy.
More precisely, data analyzed in this study consist in vehicle traffic speeds concerning the Grande Raccordo Anulare (GRA) road, that is a toll-free, ring-shaped orbital motorway, with 68.2 kilometres of circumference that encircles Rome. It represents one of the highest traffic volume Italian freeways, exceeding 160,000 vehicles per day (58 millions per year) [4]. In our dataset, the GRA road is partitioned in overall 279 locations (500 meters segments). For each location, it was recorded the respective traffic speed for the whole month of June 2016. With a length of 30 days and with observations available every 3-minutes interval, for this month were collected overall 14,400 observations for each GRA location.

3 Deep Learning for Traffic forecasting

Traffic speed data are represented as an ordered sequence of elements indexed by a regular (e.g., every day or ever year) time interval t and by a spatial index s. The goal in vehicle traffic forecasting is to develop a predictor of the quantity of interest. Suppose that for a certain location s we have observations up to time t, which we denote with $x_s^t := (x_{1,s},\dots,x_{t,s})$, and we want to predict the traffic speed for that location at a future time $t+t'$, i.e., $X_{t+t',s}$. Considering S the overall number of locations, the input matrix X^t of all the available observations is given by

$$X^t = \begin{bmatrix} x_{1,1} & \dots & x_{t-k,s} & \dots & x_{1,S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{t,1} & \dots & x_{t,s} & \dots & x_{t,S} \end{bmatrix}$$

The time point $t = 1$ is referred to the starting point of data collection or the first time for which we have information concerning the variable. However, it may be often reasonable to select an appropriate sub-interval time period and use a subset of observations for prediction, for example taking as initial time point $t-k$ and use $(x_{t-k,s},\dots,x_{t,s})$ to predict $X_{t+t',s}$. In this case, our deep learning predictor of $X_{t+t',s}$ is given by

$$\begin{aligned} X^t_{t+t',s} &= \hat{Y}(x^t_s) = \hat{Y}(x_{t-k,s},\dots,x_{t,s}) = \\ &= f^{(L+1)} \circ \dots \circ f^{(1)} \left(W^{(1)} x^t_s + b^{(1)} \right) = \\ &= f^{(L+1)} \circ \dots \circ f^{(1)} \left(Z^{(1),t}_s \right), \end{aligned}$$

where $f^{(1)},\dots,f^{(L+1)}$ and L represent the hyperparameters of the deep artificial neural network.

Our prior assumption is that to predict traffic at a given location we might need to use only recent measurements. In addition, is computationally prohibitive and somehow unreasonable to use data from every single road segment to forecast the speed of that location. The problem of selecting an optimal initial time point $t-k$ and the relevant locations, thus an optimal input matrix, is equivalent to perform a predictor selection task to find the spatio-temporal relations in the data. A predictor selection problem requires an algorithm to find a sparse model [5].

Predictor Selection

Our predictors selection procedure involves two widely used models in the time series framework, with the first one being an Autoregressive Integrated Moving Average (ARIMA) model [6, 7]. By taking into account the temporal dependencies in the data, it allows us to estimate the optimal number of temporal lags to be used. The estimated number of lags k is then used to identify the spatial relation as well, by developing a sparse Vector Autoregressive Model (sVAR). We thus consider the problem of finding a sparse matrix A in the following model:

$$X^t_{t+t'} = Ax^t + \varepsilon_t, \;\; \varepsilon_t \sim N(0,V),$$

where A is a matrix of size $S \times Sk$, with S the number of locations and k the number of previous measurements or lags.

Our predictors selected as the result of finding the sparse linear model are then used as input values at the first layer to build the deep learning model.

The training process

The training problem of our supervised deep learning model consists in finding the set of learning parameters $(\hat{W},\hat{b})$ that optimally predicts the outcome of interest. To do this, we use a training set $D = \{Y_i, X_i\}_{i=1}^{N}$ of input-output pairs to train the model by solving an optimization problem as the following

$$\arg\,\min_{W,b} \frac{1}{N} \sum_{i=1}^{N} ||Y_i - \hat{Y}^{W,b}(X_i)||_2^2. \tag{1}$$

When we have a non-linear model, the most loss functions cannot be optimized in a closed form [8] and are required iterative numerical optimization procedures, such as gradient descent [9]. Gradient descent algorithm tells us how to learn the the parameters we're interested in, but it supposes the ability to compute the gradient of the loss function, i.e., all the partial derivatives. A good algorithm employed to this purpose is Back-propagation [10]. It demonstrated to behave really better and faster than other earlier methods, becoming the workhorse of learning in neural networks. The feed-forward deep neural network was implemented with the `h2o` package in `R`. It was evaluated on both in-sample and out-of-sample metrics (i.e., mean squared error and R^2) based on a random split of 70%, 10% and 20% for the training, validation and test set, respectively.

4 Results and Conclusions

Before testing our deep learning model, we performed a Grid Search to find the optimal network (number of layers L, number of units N_l for each layer l, the activation functions f, and the regularization lasso λ_1 and ridge λ_2 penalty parameters). Based on $N = 10^4$ Monte Carlo samples, we obtained the following neural network

$$\begin{aligned} f &= \texttt{ReLU} \\ (\lambda_1, \lambda_2) &= (1\mathrm{e}-5, 0) \\ \texttt{structure} &= (100, 50, 100, 50). \end{aligned}$$

In Table 1 we show the training and test error of our deep learning model compared with a general linear model (GLM). Both models are based on the same input sets, in a first step we consider the previous 10 time measurements and in a second step we only take into account the optimal spatio-temporal structure estimated with the ARIMA and sVAR models.

We noticed a non uniform model performance throughout the day, based on recurrent and non recurrent patterns of the proposed time series. However, compared to general linear models, the deep learning models showed much better results in terms of both in sample and out of sample outcomes, particularly with the optimal number of lags and the selected two adjacent locations.

Further significant improvements are likely to be obtained with other classes of deep learning, e.g., recurrent neural networks [11] or long short term memory networks [12], and future analyses should be carried out to evaluate their performances.

Tabella 1 In Sample and Out of Sample metrics according to a general linear model (GLM) and a deep learning model (DLM). 10 and 2 indicates 10 and 2 previous time measurements. All the values are based on a Monte Carlo sample.

	MSE		R^2	
Model	In Sample	Out of Sample	In Sample	Out of Sample
GLM10	36.48	37.95	0.89	0.89
DLM10	28.36	36.53	0.92	0.89
GLM2	37.47	37.57	0.89	0.89
DLM2	31.54	33.96	0.91	0.90

Riferimenti bibliografici

1. J. B. Heaton, N. G. Polson, and J. H. Witte, “Deep learning for finance: deep portfolios,” *Applied Stochastic Models in Business and Industry*, 2016. asmb.2209.
2. W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
3. Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 5 2015.
4. ANAS, “ANAS, Grande Raccordo Anulare di Roma,” 2017. [Online; accessed 28-nov-2017].
5. N. Polson and V. Sokolov, “Deep Learning Predictors for Traffic Flows,” *ArXiv e-prints*, Apr. 2016.
6. G. J. G.E.P. Box, *Time Series Analysis, Forecasting and Control*. San Francisco, CA: Holden-Day, revised edition ed., 1970.
7. G. P. Zhang, “Time series forecasting using a hybrid arima and neural network model.,” *Neurocomputing*, vol. 50, pp. 159–175, 2003.
8. I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning.” Book in preparation for MIT Press, 2016.
9. L. A. Cauchy, “Methode generale pour la resolution des systemes d’equations simultanees,” *Compte Rendu a l’Academie des Sciences*, vol. 25, pp. 536–538,, 1847.
10. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” in *Neurocomputing: Foundations of Research* (J. A. Anderson and E. Rosenfeld, eds.), pp. 696–699, Cambridge, MA, USA: MIT Press, 1988.
11. A. Graves and J. Schmidhuber, “Offline handwriting recognition with multidimensional recurrent neural networks,” in *Advances in Neural Information Processing Systems 21* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 545–552, Curran Associates, Inc., 2009.
12. F. Gers, “Long short-term memory in recurrent neural networks,” 2001.

Estimating the number of unseen species under heavy tails

Sulla stima del numero di nuove specie nell'ipotesi di code pesanti

Marco Battiston, Federico Camerlenghi, Emanuele Dolera and Stefano Favaro

Abstract Species sampling is a popular subject in several scientific disciplines. Assuming to be provided with an initial sample of size n, a crucial issue is the estimation of the number of new species that will be observed in an additional sample of size λn, being $\lambda > 0$. The case $\lambda < 1$ has been successfully tackled in [6] and [7], but the most interesting situation $\lambda \geq 1$ has been addressed only recently in [11]. We will show that the solution of [11] is unsatisfying when the species' proportions have regularly varying heavy tails. Under this assumption, we provide another estimator for the number of new species and we empirically show its performance.

Abstract *Il campionamento di specie è di particolare interesse in molti contesti. Avendo a disposizione un campione di ampiezza n, un problema particolarmente rilevante consiste nella stima del numero di nuove specie che verranno osservate nelle prossime* λn *osservazioni, con* $\lambda > 0$. *Per il caso* $\lambda < 1$ *il problema fu risolto in [6, 7], mentre il caso* $\lambda \geq 1$ *è stato affrontato solo di recente in [11]. Mostreremo che la soluzione proposta in [11] non è soddisfacente quando le porzioni delle varie specie hanno code pesanti. Sotto opportune ipotesi sulle code delle porzioni, proporremo uno stimatore per il numero di nuove specie e mostreremo le sue proprietà attraverso alcune simulazioni.*

Marco Battiston
Department of Statistics, University of Oxford, 24-29 St Giles', OX1 3LB Oxford, UKe-mail: marco.battiston@stats.ox.ac.uk

Federico Camerlenghi
Department of Economics, Management and Statistics, University of Milano–Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy e-mail: federico.camerlenghi@unimib.it

Emanuele Dolera
Department of Mathematics, University of Pavia, via Ferrata 5, 27100 Pavia, Italy e-mail: emanuele.dolera@unipv.it

Stefano Favaro
Department of Economics and Statistics, University of Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy e-mail: stefano.favaro@unito.it

Key words: Good-Toulmin type estimators, regular variation, species estimation, two parameter Poisson-Dirichlet prior

1 Introduction

Consider a generic population of individuals $(X_i)_{i\geq 1}$ belonging to different species $(X_i^*)_{i\geq 1}$ with unknown proportions $(p_i)_{i\geq 1}$. Given an initial sample of size n, say $(X_1,\dots,X_n)$, from the population of interest, a crucial problem is the estimation of hitherto unseen species that will be observed in an additional sample of size λn, being $\lambda > 0$. More precisely, denoted by $N_{n,i}$ the frequency of the i-th species in the sample, one is typically interested to estimate the quantity

$$U_{\lambda n} := \sum_{i\geq 1} \mathbb{1}_{\{N_{n,i}=0\}} \mathbb{1}_{\{N_{\lambda n,i}>0\}},$$

i.e. the number of new species that will be observed in an additional sample $(X_{n+1},\dots,X_{\lambda n-n})$ of size λn.

The first solution of this problem have been suggested in the seminal contributions of [6] and [7]. To fix the notation, we denote by $M_{n,r}$ the number of species with frequency r in $(X_1,\dots,X_n)$, for any $1\leq r\leq n$, and by $m_{n,r}$ the corresponding value in the observed sample. Besides K_n represents the number of distinct species in $(X_1,\dots,X_n)$, and k_n the observed value. The following estimator for $U_{\lambda n}$

$$\hat{U}_{\lambda n} := -\sum_{j\geq 1} (-\lambda)^j m_{n,j} \tag{1}$$

has been provided in [7]. Such an estimator works very well whenever $\lambda < 1$, but it is useless for $\lambda \geq 1$, due to the exponential growth of the coefficients $(-\lambda)^j$. In order to predict $U_{\lambda n}$ for $\lambda > 1$, [1, 7] suggested to use the so–called Euler's transformation, which converts the series in (1) into another one having the same sum, but featuring a faster convergence of its partial sums. However, no theoretical guarantees for the resulting estimator have been established until the work by [11]. They have been able to define a general estimator $U_{\lambda n}$ for the case $\lambda > 1$, which amounts to be

$$\hat{U}^L_{\lambda n} := -\sum_{j\geq 1} (-\lambda)^j \mathbb{P}[L\geq j] m_{n,j}, \tag{2}$$

where L is a random variable whose tail probability compensates for the growth of $(-\lambda)^j$. If L is the Binomial random variable with parameter $(k,(1+\lambda)^{-1})$ then (2) coincides with the Euler-smoothed estimator of [1], with k being the truncation level of (2).

In order to illustrate the performance of (2), we consider a population of 10^6 species whose proportions p_i's are masses of the Zipf distribution, i.e. $p_i \propto i^{-s}$ for some $s > 0$. The parameter s controls the tail of the distribution, to large values of s corresponds heavy tails distributions. Figure 1 shows the estimator (2) for different choices of s, i.e. from left to right and top to bottom $s = 0.6, 0.8, 1.0, 1.2, 1.4, 1.6$. All experiments are averaged over 100 iterations. The true value is shown in black, and estimated values are colored according to the three choices of the distribution of L considered in Table 1 of [11]: i) a Poisson distribution with parameter $(2\lambda)^{-1}\log_e(n(\lambda+1)^2/(\lambda-1))$; ii) a Binomial distribution with parameter $(2^{-1}\log_2(n\lambda^2/(\lambda-1)),(\lambda+1)^{-1})$; iii) a Binomial distribution with parameter $(2^{-1}\log_3(n\lambda^2/(\lambda-1)),2(\lambda+2)^{-1})$. Shaded bands correspond to one standard deviation. Figure 1 highlights how the tail behavior of the p_i's affects the experimental performance of the estimator $\hat{U}^L_{\lambda n}$: the heavier the tail of $(p_i)_{i\geq 1}$, or rather the lower the species discovery rate, the worse the performance of $\hat{U}^L_{\lambda n}$. The underestimation phenomenon thus suggests that the methods proposed by [1] and then by [11] are not useful for heavy-tailed p_i's. Indeed those methods rely on analytic considerations aimed at improving the rate of convergence of the estimator (1), without acting on the species compositions $(p_i)_{i\geq 1}$. Heavy-tailed species proportions is a common setting in several application areas (see, e.g., [14, 15]), hence the definition of an estimator for $U_{\lambda n}$ under the assumption of heavy-tailed proportions p_i's is a problem of paramount importance.

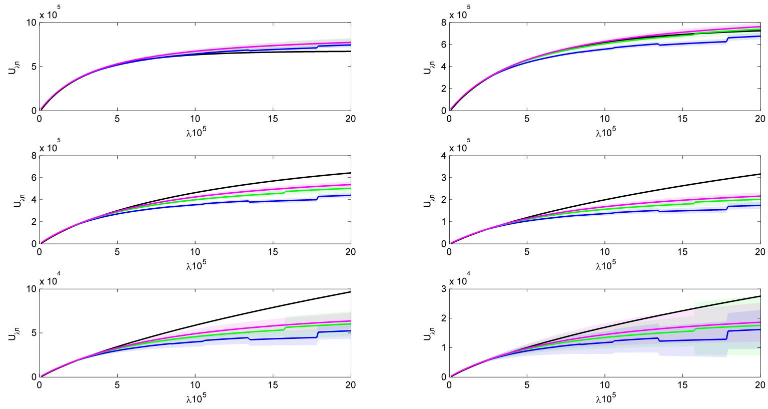


Fig. 1 Estimator of $U_{\lambda n}$ in six Zipf scenarios. The true value is drawn in black, the estimated values are colored in blue (L being the Poisson distribution), green (L being the binomial distribution with success probability $1/(\lambda+1)$) and magenta (L being the binomial distribution with success probability $2/(\lambda+2)$). The shaded bands correspond to one standard deviation.

In the present paper, we introduce the estimator of $U_{\lambda n}$ under heavy–tailed proportions p_i's, showing that it has an opposite behaviour with respect to that highlighted in Figure 1, namely the estimations improve as the parameter s of the Zipf distribution increases. In Section 3 we briefly discuss how to choose the best estimator of $U_{\lambda n}$ among those presented here in relation to the problem at the hand.

Finally we hint possible connections with the Bayesian nonparametric approach, which merit further investigation.

2 Good-Toulmin estimators under regular variation

In the previous section we have seen that the higher the tail of $(p_i)_{i\geq 1}$ (i.e. the higher the parameter s of the Zipf law), the worse the underestimation of $\hat{U}^L_{\lambda n}$. In order to define a suitable estimator for large values of s, we impose a specific assumption on the tails of $(p_i)_{i\geq 1}$, more precisely we resort to the theory of regular variation [8]. In the sequel we will use the notation $f \sim g$ to mean $f/g \to 1$, besides define $\nu(\mathrm{d}x) := \sum_{i\geq 1} \delta_{p_i}(\mathrm{d}x)$ and the measure $\overline{\nu}(x) := \nu[x,1]$. We will say that $(p_i)_{i\geq 1}$ is regularly varying with regular variation index $\alpha \in (0,1)$ if $\overline{\nu}(x) \sim x^{-\alpha}\ell(1/x)$ as $x \downarrow 0$, where $\ell(t)$ is a slowly varying function, that is $\ell(ct)/\ell(t) \to 1$ as $t \to +\infty$ for all $c > 0$. Karlin [8] has proven that in such a context

i) $K_n \overset{\text{a.s.}}{\sim} \mathbb{E}[K_n] \sim \Gamma(1-\alpha)n^{\alpha}\ell(n)$,
ii) $M_{n,r} \overset{\text{a.s.}}{\sim} \mathbb{E}[M_{n,r}] \sim \frac{\alpha\Gamma(r-\alpha)}{r!}n^{\alpha}\ell(n)$

where $\Gamma(\cdot)$ represents the Gamma function. The regular variation index is not known *a priori* and needs to be estimated from the data. This issue can be easily addressed taking the ratio of the number of species with frequency one and the total number of species, namely $\hat{\alpha} := \frac{M_{n,1}}{K_n}$ is a (strongly) consistent estimator of α. For additional details on regular variation refer to [8] and [5].

We now define an estimator for $U_{\lambda n}$, when $\lambda > 1$ and the sequence $(p_i)_{i\geq 1}$ has regularly varying heavy tails. In order to to this, we consider (2) when L is a Binomial random variable with parameters $(2^{-1}\log_2(n\lambda^2/(\lambda-1)), (\lambda+1)^{-1})$, and we tune this estimator under the hypothesis of regular variation, thus obtaining

$$\hat{U}^L_{\lambda n}(\alpha) := k_n \sum_{z=1}^{k_L} \binom{z+\alpha-1}{z} \left(\frac{\lambda}{\lambda+1}\right)^z, \tag{3}$$

where k_L is the truncation level, i.e. $k_L := 2^{-1}\log_2(n\lambda^2/(\lambda-1))$. Note that $k_L \to +\infty$ as $n \to +\infty$.

Finally it is worth noticing that in (3) the regular variation index α is unknown, hence, in order to use the estimator $\hat{U}^L_{\lambda n}(\alpha)$, one should replace α with the corresponding consistent estimator $\hat{\alpha} = \frac{M_{n,1}}{K_n}$.

We now consider the same Zipf's scenarios presented in Section 1 to illustrate the performance of (3). Figure 2 shows $\hat{U}^L_{\lambda n}(\hat{\alpha})$ for different choices for the parameter s of the Zipf distribution, i.e. from left to right and top to bottom $s = 0.6, 0.8, 1.0, 1.2, 1.4, 1.6$. All experiments are averaged over 100 iterations. The true value is shown in black, the estimated value in red, and the shaded band corresponds to one standard deviation. By a comparison between Figure 1 and Figure 2, one immediately realizes that $\hat{U}^L_{\lambda n}(\hat{\alpha})$ has an opposite behaviour with respect to

$\hat{U}^L_{\lambda n}$. That is, the heavier the tail of $(p_i)_{i\geq 1}$, or rather the lower the species discovery rate, the better the performance of $\hat{U}^L_{\lambda n}(\hat{\alpha})$.

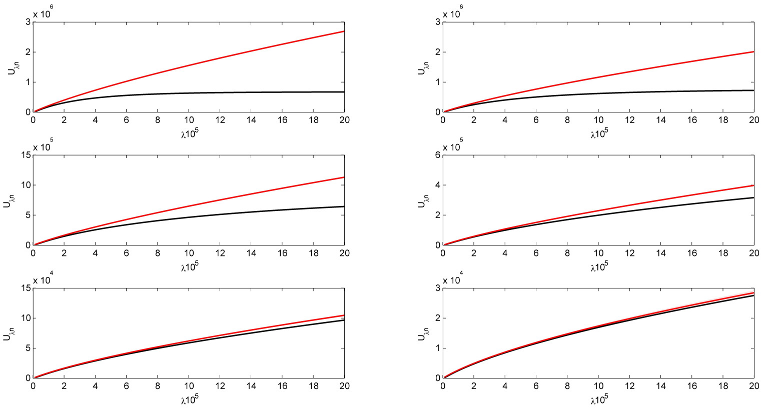


Fig. 2 Estimator of $U_{\lambda n}$ in six Zipf scenarios. The true value is drawn in black, the estimated value in red. The shaded bands correspond to one standard deviation.

3 Discussion

In this paper we focused on the estimation of the number of unseen species that will be observed in a future sample of size λn. The performance of the estimators presented here has been assessed empirically for the ubiquitous Zipf distribution with parameter s. We have shown that the estimator (2) proposed by [11] is useful when $s \leq 1$, but it radically worsens when $s > 1$. For this reason, in Section 2, we have tuned such an estimator under the assumption of regularly varying heavy tails p_i's. In Figure 2, we have empirically shown that $\hat{U}^L_{\lambda n}(\hat{\alpha})$ performs very well when $s > 1$, but not when $s \leq 1$, featuring an opposite behaviour with respect to $\hat{U}^L_{\lambda n}$. In real applications the parameter s is not given and one has to decide whether to employ either $\hat{U}^L_{\lambda n}$ or $\hat{U}^L_{\lambda n}(\hat{\alpha})$. In order to face this issue one can find an estimate of the parameter s by means of linear regression as suggested in [10], thus using $\hat{U}^L_{\lambda n}$ if the resulting estimator of s is less than 1, $\hat{U}^L_{\lambda n}(\hat{\alpha})$ otherwise.

An interesting open problem which merits further investigation is the connection between the estimator of the number of unseen species $\hat{U}^L_{\lambda n}(\alpha)$ presented here and the Bayesian nonparametric estimator derived in [2] and [3]. The Bayesian viewpoint needs the specification of a prior distribution for the species proportions $(p_i)_{i\geq 1}$, namely one needs to choose a prior distribution for the random probability measure $\tilde{p} = \sum_{i\geq 1} p_i \delta_{X_i^*}$. In such a context $(X_1,\dots,X_n)$ is a sample coming from an exchangeable sequence of observations driven by $\tilde{p}$, i.e.

$$\begin{aligned} X_i \,|\, \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \qquad i=1,\dots,n, \\ \tilde{p} &\sim \mathscr{P}, \end{aligned} \tag{4}$$

where $\mathscr{P}$ is the distribution of $\tilde{p}$. A common choice for $\mathscr{P}$ is the law of the two parameter Poisson-Dirichlet process, which was introduced in [12] and further investigated in [13]. The sequence $(p_i)_{i\geq 1}$ is such that $p_1 = v_1$ and $p_i = v_i \prod_{1\leq j\leq i-1}(1-v_j)$, for any $i \geq 2$, where the v_j's are independent random variables, each v_j is distributed according to a Beta distribution with parameter $(1-\alpha, \theta + j\alpha)$, for $\alpha \in (0,1)$ and $\theta > -\alpha$. In [4], the authors have proven that the celebrated Good-Turing estimator of the discovery probability is asymptotically equivalent, for a large sample size, to the Bayesian nonparametric one under the assumption of a two parameter Poisson-Dirichlet prior. Analogously, we would like to asses whether $U_{\lambda n}$ is asymptotically equivalent to the regularly varying nonparametric estimator $\hat{U}^L_{\lambda n}(\alpha)$ for specific choices of $\mathscr{P}$, as the sample size n increases.

References

1. Efron, B., Thisted, R.: Estimating the number of unseen species: How many words did Shakespeare know? Biometrika **63**, 435–447 (1976)
2. Favaro, S., Lijoi, A., Mena, R.H., Prünster, I.: Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. J. Roy. Statist. Soc. Ser. B **71**, 993–1008 (2009)
3. Favaro, S., Lijoi, A., Prünster, I.: A new estimator of the discovery probability. Biometrics **68**, 1188–1196 (2012)
4. Favaro, S., Nipoti, B., Teh, Y.W.: Rediscovery of GoodTuring estimators via Bayesian nonparametrics. Biometrics **72**, 136–145 (2016)
5. Gnedin, A., Hansen, B., Pitman, J.: Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. Probab. Surv. **4**, 146–171 (2007)
6. Good, I.J.: The population frequencies of species and the estimation of population parameters. Biometrika **40**, 237–264 (1953)
7. Good, I.J., Toulmin, G.H.: The number of new species, and the increase in population coverage, when a sample is increased. Biometrika **43**, 45–63 (1956)
8. Karlin, S.: Central limit theorems for certain infinite urn schemes. J. Math. Mech. **17**, 373–401 (1967)
9. Lijoi, A., Mena, R.H., Prünster, I.: Bayesian nonparametric estimation of the probability of discovering new species. Biometrika **94**, 769–786 (2007)
10. Motwani, S., Vassilvitskii, S.: Distinct value estimators for power law distributions. In Proceedings of the Workshop on Analytic Algorithms and Combinatorics (2006)
11. Orlitsky, A., Suresh, A.T., Wu, Y.: Optimal prediction of the number of unseen species. Proc. Natl. Acad. Sci. USA **47**, 13283–13288 (2016)
12. Perman, M., Pitman, J., Yor, M.: Size-biased sampling of Poisson point processes and excursions. Probab. Theory Related Fields **92**, 21–39 (1992)
13. Pitman, J., Yor, M.: The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**, 855–900 (1997)
14. Sampson, G.: Empirical linguistic. Bloomsbury Academic (2002)
15. Thompson, W.K.: Sampling rare or elusive species. Island Press (2004)

How to measure cybersecurity risk

Misurare il cyber rischio

Silvia Facchinetti, Paolo Giudici and Silvia Angela Osmetti

Abstract In the last years there have been a scholars increasing interest in cybersecurity risk measurement. This paper proposes a methodology to measure cyber risk, using ordinal data, to prioritise appropriate interventions. The method relies on the construction of a Criticality index: a new measure of risk based on the cumulative probabilities of the ordinal variable that represents the level of severity for different risk event. Its properties are derived and compared with alternative measures employed in operational risk measurement. we apply our proposal to real data of a telecommunication service company. The proposed measure is found to be quite effective to rank cyber risk types and, therefore, allow selective preventive actions.
Abstract *Negli ultimi anni si è registrato un interesse crescente da parte degli studiosi al problema del cyber rischio e alla sua misurazione. Dato che in tale contesto i dati sono spesso di natura ordinale, nel presente lavoro proponiamo un indice per stimare il cyber rischio utilizzando dati qualitativi riguardanti il livello di severità dei cyber attacchi. Noi analizziamo le caratteristiche dell'indice, e lo confrontiamo con misure alternative utilizzate nella valutazione del rischio operativo. La nostra proposta viene applicata ai dati di una compagnia di telecomunicazioni, riguardanti la severità dei cyber attacchi subiti dai clienti della compagnia. L'indice risulta efficace per classificare le diverse linee di business in base alla loro rischiosità e quindi per consentire tempestive azioni preventive.*

Key words: criticality index, risk measure, ordinal variables, cyber attacks

Silvia Facchinetti
Department of Statistical science, Università Cattolica del Sacro Cuore, Milano, e-mail: silvia.facchinetti@unicatt.it

Paolo Giudici
Department of Economics and Management, University of Pavia, Pavia e-mail: paolo.giudici@unipv.it

Silvia Angela Osmetti
Department of Statistical science, Università Cattolica del Sacro Cuore, Milano, e-mail: silvia.osmetti@unicatt.it

1 Introduction

In the last years the number of cyber attacks in information technology (IT) systems is surging. Therefore, cybersecurity has become more of a concern for businesses. Among operational risks caused by IT systems, cyber risks are gaining increasing importance, due to technological advancements and to the globalisation of financial activities (see e.g. [2],[3]). Therefore the cybersecurity risk measurement is an increasing interest for scholars (see e.g. [1],[6]).

We propose a methodology to measure cyber risks, starting from ordinal random variables, that represent the levels of severity for different risk events (expressed on ordered categories, such as ”high”, ”medium” or ”low” risk), in different business lines. In particular we propose, for different business lines, a cybersecurity risk index that is based on the relative frequencies of the severity levels. As a result, we obtain an ordinal measure of risk which will be used to compare different events and business lines, producing an ordering among risks useful to prioritise intervention in process controls.

Besides the theoretical proposal, we will present empirical evidences on the performance of our index, using a real data set, that concerns cyber risk measurement in a telecommunication company.

2 Proposal

Let $X \sim \{x_k, p_k; k = 1, 2, \ldots, K\}$ be a categorical random variable with ordered categories x_k and probabilities $p_k = P(x_k)$, that represents a severity variable, with decreasing levels, $k = 1, 2, \ldots, K$.

We define a *Criticality Index* for the categorical random variable X with the following expression:

$$I = \frac{1}{K-1} \sum_{k=1}^{K-1} (K-k) p_k = \frac{\sum_{k=1}^{K} F_k - 1}{K-1}, \tag{1}$$

where $F_k = \sum_{l=1}^{k} p_l$ are the values of the cumulative distribution function of the ordinal variable X, for $k = 1, 2, \ldots, K$.

It is a natural measure of risk for ordinal variables, with values in $[0,1]$. It thus provides a risk measure easy to interpret and suitable for the comparison of the risk level between different risk events and/or business lines.

We propose to estimate cybersecurity risk in each business line/event type combination by the sample version of the *Criticality Index*, obtained by replacing the probabilities p_k with their estimators $\hat{p}_k = r_j/n$ for $k = 1, 2, \ldots, K-1$:

$$\hat{I} = \frac{1}{K-1} \sum_{k=1}^{K-1} (K-k) \frac{r_k}{n} = \frac{\sum_{k=1}^{K} \tilde{F}_k - 1}{K-1}, \tag{2}$$

where

$$\tilde{F}_k = \sum_{l=1}^{k} \frac{r_l}{n} \text{ for } k = 1,2,\dots,K,$$

is the empirical cumulative distribution function, $r_l = \sharp(\tilde{x}_i \equiv x_l)$ is the number of observations in the sample equal to the category x_l, with $r_l \in \mathbf{N}$ and $\sum_{l=1}^{K} r_l = n$ (n is the total number of risk events observed for the j-th business line).

It is possible to demonstrate that the *Criticality Index* estimator is asymptotically normally distributed for $n >= 30$. Moreover, $\hat{I}$ is an unbiased and consistent estimator for I [4].

3 Application

We apply our proposal to real data provided by a telecommunication company that installs telephone exchange systems and offers post-installation technical assistance for upgrading and problem resolution in different event types, that include, in particular, Network communications. The service is offered to a wide range of customers, that are grouped in several business lines.

The main research problem is to estimate for each business line, a measure of cybersecurity risk based on ordinal data, collected by the customer care center, describing the level of severity of cyber attack suffered by customers.

The data are reported in Table 1. Each row shows a business line and each column reports how many times a cyber problem in Network communication has been reported, for levels of severity equal to high (H), medium (M) and low (L).

Table 1 Data for Network communication event type

Business Line	H	M	L
Banking	23	128	4
Computers	3	26	0
Cooperatives	13	93	2
Defence	34	149	7
Health	43	222	8
Hotels	10	108	3
Industry	13	94	7

In Table 2, column 2 and 3, we report the *Criticality Index* estimates and its associated asymptotic confidence interval.

We show that Defence is the business line with the highest level of risk, followed by Health and Banking. Therefore, a mitigation intervention to prevent cyber risk should prioritise the Defence business line, and the customers in that business line.

To evaluate the robustness of our results we compare them with what could be obtained with the approach proposed by [5] in the context of operational risks. We apply their Stochastic Dominance Index (SDI) to our data and their suggested

Table 2 *I* and SDI risk measure estimates and their respective confidence intervals (CI)

Business line	$\hat{I}$	CI	SDI	Bayesian CI
Banking	0.561	0.517-0.606	0.708	0.685-0.728
Computers	0.552	0.473-0.630	0.701	0.654-0.734
Cooperatives	0.551	0.503-0.599	0.701	0.676-0.723
Defence	0.571	0.527-0.616	0.714	0.692-0.735
Health	0.564	0.529-0.599	0.709	0.692-0.726
Hotels	0.529	0.488-0.570	0.686	0.665-0.706
Industry	0.526	0.472-0.580	0.684	0.657-0.711

Bayesian procedure to derive a confidence interval, by means of a Gibbs Sampling algorithm with R=10000 interactions. What obtained is reported in Table 2, column 4 and 5. Note that since $n \simeq 30$ for all the business lines in Table 2, is not strictly necessary to apply a Bayesian approach to derive confidence intervals of our index, we can apply the asymptotic normality. Bayesian confidence intervals may instead be useful for ”rare” problem business lines, for which an asymptotic confidence interval is not possible.

By compare the results of Table 2, we observe, obviously, a different values for the two indices since they are calculated in different way: SDI is based on the observed frequencies, whereas our index is based on the observed relative frequencies of the severity variables conditional on the total n of that business line. Moreover, we observe that the SDI values are always higher than the results obtained with our approach. By comparing the results for different BLs, we show that our index and the SDI produce a consistent ranking, indicating similar priorities of intervention. This confirm that our index could be suitable to measure the level of risk, to compare the level of risk for many BLs or even types and that it can be considered as a synthetic priority of intervention indicator.

References

1. Afful-Dadzie, A., and Allen, T.T.: Data-Driven Cyber-Vulnerability Maintenance Policies. Journal of Quality Technology, **46**, 234-250 (2017).
2. Cebula, J.J., and Young, L.R. (2010). A Taxonomy of Operational Cyber Security Risks, Technical Note CMU/SEI-2010-TN-028, Software Engineering Institute, Carnegie Mellon University, 1-34.
3. Edgar, T.W., and Manz, D.O. (2017). Research Methods for Cyber Security, Elsevier.
4. Facchinetti, S., and Osmetti, S.A.: A risk index for ordinal variables and its statistical properties: a priority of intervention indicator in quality control framework. Quality and Reliability Engeneering International, **34**, 265275 (2018).
5. Figini, S., and Giudici, P.: Measuring risk with ordinal variables. Journal of Operational Risk, **8**, 35-43 (2013).
6. Hubbard, D.W., and Seiersen, R.: How to Measure Anything in Cybersecurity Risk. Wiley, New York (2016).

Implementation of an innovative technique to improve Sauvignon Blanc wine quality

Applicazione di una tecnica innovativa per migliorare la qualità di vini Sauvignon Blanc

Filippa Bono[1], Pietro Catania[1] and Mariangela Vallone[1]

Abstract The purpose of the study was to compare two different pressing systems of Sauvignon Blanc grapes using an innovative wine press manufactured by Puleo Srl Company (Marsala, Italy). Grape pressing is a very important step in the winemaking process as it may promote the presence and/or absence of enzyme processes on the must, leading to the creation of different products in terms of chemical composition from the same grapes. Chemical composition of must firstly and wine after, obtained from the two pressing mode, was analysed in first instance with PCA method.

Results are encouraging and open up new research prospective with the aim of applying innovative techniques to improve the quality of the final product.

Abstract *Scopo del lavoro è valutare due diversi sistemi di pressatura delle uve Sauvignon Blanc ottenute impiegando una macchina innovativa prodotta dalla ditta Puleo di Marsala. La pressatura dell'uva è un passo molto importante nel processo di vinificazione in quanto può favorire la presenza e/o l'assenza di processi enzimatici sul mosto, portando alla creazione di diversi prodotti in termini di composizione chimica a partire dalle stesse uve. Le componenti chimiche del mosto prima e del vino dopo, ottenuti con due differenti modalità di pressatura, sono stati confrontati con il metodo PCA.*

I risultati sono incoraggianti per nuove prospettive di ricerca con l'obiettivo di applicare sistemi di pressatura innovativi a migliorare la qualità del prodotto finale

Key words: winemaking, quality, PCA

[1] Filippa Bono, Department of Economic Business and Statistical Sciences, email: filippa.bono@unipa.it; Pietro Catania, Department of Agricultural, Food and Forest Sciences, email: pietro.catania@unipa.it; Mariangela Vallone, Department of Agricultural, Food and Forest Sciences, email: mariangela.vallone@unipa.it.

1 Introduction

Oxygen plays a crucial role in the winemaking process and it can influence the composition and quality of the must and wine. Oxygen can influence the composition and quality of wine drastically, either positively or negatively. Oxygen exposure occurs naturally during mechanical harvesting, crushing and pressing. In must, during alcoholic fermentation and during ageing of white and red wines oxygen has different effect [1]. The use of sulphur dioxide as an anti-oxidant dates back to the early 18th century and the protection of wine from unwanted oxidative spoilage has been recognized [2]. There is very little scientific research on the effect of oxygen use during fermentation on wine composition or sensory proprieties. Furthermore these effects change with grapes cultivar. Smith et al. [3] studied the oxygen influence in fermentation, colour, tannin and sensory of shiraz wines. Recently, Boselli et. al. [4] observed the effect of Nitrogen gas on three white grapes varieties (Chardonnay, Grechetto and Orvieto) obtaining the strongest protective effect of nitrogen on phenolics in Chardonnay and Grechetto musts. They showed that Nitrogen gas is therefore particularly recommended not only in positive pressure, but also for vacuum-pressing of white grapes containing high levels of catechin or gallic acid due to early harvest or peculiar varietal composition. Interest in wine press innovation suggest the opportunity to carry out real-time experiments to assess the effect that different types of wine presses may have on the must and, consequently, on the quality of the wine. In our study we consider two different wine pressing modes applied on Sauvignon Blanc grapes using a modern press by Puleo s.r.l (Marsala, Italy). Aim of the study is to identify if the two wine pressing modes influence must and wine characteristics.

The statistical method to compare results is the Principal Component Analysis (PCA).

2 Data and Methods

Grapes pressing represents a very important phase in winemaking since it can promote the presence and/or absence of enzymatic processes endogenous to the juice itself, leading to the creation of different products in terms of chemical composition, although starting from the same grapes.

The most important elements influencing the quality of wine in grapes pressing are: press type; oxygen in the tank; level of pressing; juice extraction time. In the analysis, Sauvignon Blanc grape variety is considered. Grapes were manual harvested in the third decade of August 2016, the must composition was observed in three different times during winemaking (29th August, 1th September, 3th September). The analyses on wine were performed at the end of fermentation.

Pneumatic discontinuous press is actually the most used machine in quality wine making.
In our study the "Vortex System" pneumatic press developed by Puleo Srl was used. The “Vortex system” puts the idea of pressing under inert gas with nitrogen recovery towards a new and more advanced perspective. The idea of using inert gas in pressure during working cycle, guarantees a continuous draining action highly greater than any other kind of pressing, with excellent results in product quality/yelds. A further advantage is the immediate extraction of the must from the cylinder to the storage tank thanks to the “Vortex” created inside the press. During grapes processing, the inert gas is continuously filtered to be purified by any pollution due to unwanted parts that can be carried by the gas itself. This wine press has a close tank and two operating modes: the traditional Air Pressing mode (AP) and the Nitrogen Pressing mode (NP).
The AP mode is a traditional pneumatic press with a close tank. Grape must is extracted from the inner draining channels and comes out from nozzles to an external collection tray. In the NP mode, grapes contact with air is minimized, lesser than any pneumatic press with close tank.
Analytical determination on must and wine were performed by Foss Integrator WineScan™, (FOSS Italia S.p.A.). The must and wine determinations were alcohol [%/vol], density [g/l], sugar [g/l], pH, total acidity [g/l], volatile acidity [g/l], malic acid [g/l], citric acid [g/l], tartaric acid [g/l], potassium [g/l], polyphenols [mg/l], ashes[g/l], APA (readily assimilable nitrogen) [g/l], gluconic acid [g/l], methanol [g/l], CO2, absorbance at 280, absorbance at 325, catechins [mg/l]. The observations are in triplicate respectively in 29th August, 1th September, 3th September.
The PCA method is used to compare the must and the wine characteristics obtained in the different press modes.

3 Results

Table 1 shows descriptive statistics in must composition obtained with the two pressing modes, AP and NP. These differences in must composition determine the quality of the final wine. Must obtained in the NP pressing mode had a lower level of alcohol, pH, total acidity, methanol and CO_2 than AP, while the levels of APA, potassium, sugars, malic and citric acids and polyphenols were particularly higher than AP.
In particular, the higher level of acids obtained in NP mode is favourable to the development of ester profile that determines the aromatic component of wines.
The presence of polyphenols is very important as antioxidant and preservatives. The NP pressing mode doubles polyphenols than the AP pressing mode.
PCA analysis was performed to evaluate changes in must’s components obtained with the two pressing modes.

Table 1. Descriptive statistics of must composition in AP and NP modes

	AP				*NP*			
Variable	***Mean***	***Std. Dev.***	***Min***	***Max***	***Mean***	***Std. Dev.***	***Min***	***Max***
Alcohol	10.84	1.51	8.90	12.55	10.35	1.88	7.89	12.12
Density [g/l]	1.01	0.02	0.97	1.04	1.01	0.02	0.98	1.04
Sugar [g/l]	39.58	22.85	15.85	69.00	44.14	29.27	15.98	82.49
pH	3.41	0.05	3.33	3.51	3.38	0.04	3.33	3.44
Volatile Acidity [g/l]	0.21	0.01	0.18	0.22	0.18	0.01	0.16	0.20
Malic acid [g/l]	2.66	0.15	2.50	2.99	3.52	0.13	3.35	3.73
Citric acid [g/l]	0.29	0.04	0.24	0.36	0.31	0.04	0.26	0.37
Tartaric acid [g/l]	2.57	0.15	2.35	2.78	2.47	0.16	2.20	2.66
Potassium [g/l]	988.63	52.54	926.37	1073.00	1153.22	28.74	1107.60	1189.99
Polyphenols [mg/l]	269.92	32.15	221.19	321.16	468.70	56.88	380.00	528.53
Ashes [g/l]	2.38	0.10	2.24	2.56	2.64	0.11	2.50	2.80
APA [g/l]	28.18	17.50	6.52	50.00	80.18	18.10	58.68	103.90
Gluconic acid [g/l]	2.54	0.13	2.34	2.72	2.56	0.16	2.30	2.80
Methanol [g/l]	0.45	0.51	0.10	1.15	0.14	0.02	0.12	0.19
CO_2	1418.73	156.31	1236.00	1638.56	1251.59	261.60	948.00	1615.00

Table 2. shows eigenvalues of PCA performed on must data. Following the eigenvalue>1 criterion, three components are extracted that overall explain 82% of the total variance of must components. The score plot gives us a feeling for the similarities and differences between the pressing mode in component loading.

Table 2. Eingenvalues of the must Principal Component

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	5.583	0.322	0.372	0.372
Comp2	5.261	3.824	0.351	0.723
Comp3	1.438	0.649	0.096	0.819

The correlation between elements of must and the first principal component extracted suggests us to name it as *fermentation* component due to its positive correlation with sugars, citric acid, methanol and CO_2 and negative correlation with alcohol and APA. The second component measures *antioxidant and preservative* characteristics given its correlation with malic acid, polyphenols and ashes and negative correlation with volatile acidity. The third component can measure the *must character* just its correlation with tartaric acid and pH. The Graph 2 shows the score plot that highlights a net separation between musts obtained with the two pressing modes, AP and NP, in the combination of the three components extracted.

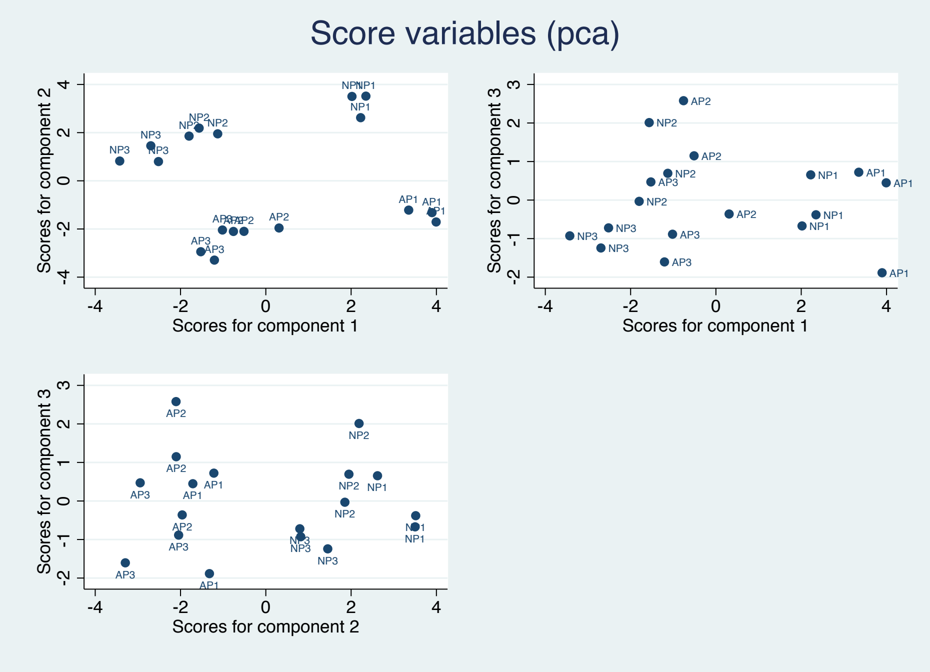


Figure 1: Score plot in must

Starting from up in the right of the graph, AP1 (must in the first time of observation) is opposite to NP3. The NP pressing mode has higher antioxidant and preservative characteristics associated with lower fermentation component than AP pressing mode. Down on the left we note also a net separation between second and third components that measure the character of must and its stability.

The component 3 has a higher correlation both in AP and NP when values are observed in time 2 even if with a net separation between the data of the two pressing modes. Component 2 has the highest correlation at time 1, i.e when grape juice is extracted.

The PCA on wine data shows two principal components that explain on the whole the 92% of the total variability of wine in the elements observed. In particular, the first component is expressive of the innovative press system showing a net separation of the wines obtained in the two pressing modes (Fig. 2). T test shows significant differences between the two pressing modes in all the elements except absorbance at 520 nm, absorbance at 620 nm, catechins, acetaldehyde (data not shown). Density, total extract, pH, total acidity, malic acid, citric acid, ashes, polyphenols, calcium, copper and sulphates in NP are higher than AP. Alcohol, sugars, volatile acidity, potassium, glycerine, absorbance at 420 nm, methanol and CO_2 observed in NP are significantly lower than AP.

Concluding, the application of NP pressing mode produced a wine with higher qualitative characteristics than the wine obtained with the AP pressing mode.

Starting from the results obtained by the PCA method, future studies will focus on the possibility of making an "ideal" wine through the estimation of a quality indicator based on the must characteristics. This indicator could also be useful to support the role of oenologists in the different phases of winemaking.

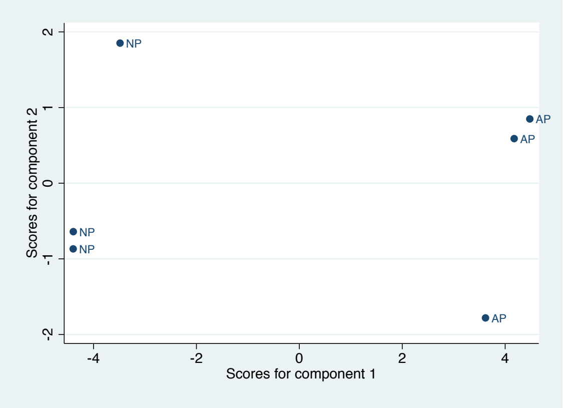


Figure 2: score plot of wine determinations in NP and AP mode

References

[1] Du Toit, W., J., Marais, J., Pretorius, I.S., and du Toit, M.: Oxygen in must and wine: A review. S. Afr. J. Enol. Vitic., Vol.27, No 1, 2006

[2] Ribéreau-Gayon, P., Dubourdieu, D., Doneche, B. & Lonvaud, A., Handbook of Enology, Volume 1: The microbiology of winemaking and vinifications.. Ed. Ribéreau-Gayon P. Wiley, Chichester, England (2000)

[3] Smith, P., Day, P., Schmidt, S., McRae, J., Bindon, K., Kassara, S., Schulkin, A., Kolouchova R., Wilkes, E., Henderich M., Johnson, D.: Exploring oxygen's influence. Wine & Viticulture J, November December 2014

[4] Boselli, E., Di Lecce, G., Alberti, F., Frega, N.G.: Nitrogen gas affects the quality and phenolic profile of must obtained from vacuum-pressed white grapes. LWT- Food Sci Technol. 43(10):1494-1500 (2013)

Investigating the effect of drugs consumption on survival outcome of Heart Failure patients using joint models: a case study based on regional administrative data.

Indagine sull'effetto del consumo di farmaci sulla probabilità di sopravvivenza di pazienti con Scompenso Cardiaco mediante l'utilizzo di modelli congiunti: un caso studio basato su dati amministrativi regionali.

Marta Spreafico, Francesca Gasperoni, Francesca Ieva

Abstract In this work, we propose an innovative approach for investigating the effect of drugs consumption on survival outcomes of patients affected by Hearth Failure (HF), a widespread chronic heart disease. In order to achieve this goal, we consider Joint Models approach [7] on administrative dataset of Lombardia Region. In this database several information is collected about patients' pharmacological history, which can be used to recover time-dependent data concerning drug assumptions over time. Through the application of this data, we are able to study the influence of longitudinal processes given by pharmacological treatments consumptions on patients' survival outcomes.

Abstract *In questo lavoro proponiamo un approccio innovativo per indagare l'effetto del consumo di farmaci sulla probabilità di sopravvivenza dei pazienti affetti da Scompenso Cardiaco. A tal fine, consideriamo l'approccio dei modelli congiunti [7] applicandolo al dataset amministrativo di Regione Lombardia. Questo database raccoglie diverse informazioni sulla storia farmacologica dei pazienti, le quali possono essere utilizzate per risalire a variabili tempo-dipendenti rigurdanti il consumo del farmaco nel tempo. Attraverso l'uso di questi dati, siamo in grado di studiare l'influenza del processo longitudinale dato dai trattamenti farmacologici sugli esiti di sopravvivenza dei pazienti.*

Marta Spreafico
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: marta1.spreafico@mail.polimi.it

Francesca Gasperoni
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: francesca.gasperoni@polimi.it

Francesca Ieva
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: francesca.ieva@polimi.it

Key words: Heart Failure, drug consumption, adherence, administrative data, time-varying covariate, joint models.

1 Introduction

Heart Failure (HF) is a chronic cardiac disease, widespread all over the world especially among people over 65 years. About 80,000 new cases per year are recorded [5]. In pharmacoepidemiology the concept of adherence, which generally refers to whether a patient takes a prescribed medication according to schedule [1], is a key factor in effective disease management of many chronic conditions. Drug Utilization Research (DUR) is the branch of pharmacoepidemiology that deals with the use of drugs and it has the goal of facilitating the rational use of drugs in patients populations. For these purposes administrative data allow to measure the effective drug utilization given the limitation of not being able to assert if the patient is currently consuming the dispensed drug. In fact, in order to evaluate the use of a drug we need a statistical measure of consumption. Among possible measures introduced in [9], one of the most used is the Defined Daily Dose (DDD), which is defined as *"the assumed average maintenance dose per day for a drug used for its main indication in adults"*. DDDs can be recovered using the Anatomical Therapeutic Chemical (ATC) classification system introduced by the World Health Organization (WHO) in 1976. ATC code allows to identify phramacological classes and includes information about drug's DDD and routes of administration. DDD is a unit of measure and it does not necessarily correspond to the Prescribed Daily Dose (PDD) by a doctor. In addition to evaluate the assumed drug quantity, we want also to establish if the drug is taken continuously during all the follow up period. There exist lots of different adherence measures. According to [4], we use Proportion of Days Covered (PDC), that is:

$$\text{PDC} = \frac{\text{number of distinct coverage days}}{\text{number of days in the observation period}} \in [0,1] \tag{1}$$

Finally, using PDC, we are able to determine adherent (PDC $\geq$ 0.80) and non-adherent patients or we can categorize them in four levels of PDC, which are $[0;0.25)$, $[0.25;0.5)$, $[0.5;0.75)$ and $[0.75;1]$.

2 The dataset

In the Lombardia Region dataset patients hospitalized for HF from 2000 to 2012 are considered, as described in [6]. For our work, we use a representative sample composed by 1,333,954 events related to 4,872 patients with their first HF hospitalization between 2006-2012.

Each patient, identified by its unique anonymous ID code, is followed from the

starting date (i.e. discharge from first HF hospitalization) until death or censoring. Administrative censoring date is December 31st, 2012. For each patient, age, gender, a list of comorbidities [2] and procedures he/she underwent are recorded.
Moreover, each record in the dataset is related to an event, which can be a hospitalization or a pharmacological prescription. In the first case, the dates of admission and discharge, together with the lenght of stay in hospital are given. In the second one, ATC codes, dates of prescription and coverage days are provided. We focus our work on five pharmacological classes: ACE-Inhibitors (ACE), Angiotensin Receptor Blockers (ARB), Beta-Blocking agents (BB), Anti-Aldosterone agents (AA) and Diuretics (DIU).
Finally, for each type of drug, we calculate patients' PDC and adherence level, setting an observation period of 365 days (one year), as done in [4].

3 Time-Varying covariates

When dealing with longitudinal and/or survival data, time-dependent covariates are often of interest. Since in classical survival models adherence is usually considered as a binary fixed variable, we are interested in representing pharmacological information as a time-varying covariate, which is a more realistic representation. In particular, we want to consider treatments as time-dependent internal covariates, since they are modified according to the development of the illness.
In order to do that, we compute a time-dependent variable which indicates the total days covered by the type of drug up to time t. We set an observation period of 365 days and we consider only distinct days, which means that, in case of overlapping periods between two prescriptions, we consider the first event entirely and only the days of the second one not covered by the first. Furthermore, we hypothesize that all the prescribed types of drug are assumed by patients during the whole period of hospitalization.
Each patient could potentially have five different curves, one for each pharmacological class (ACE, ARB, BB, AA and DIU) depending on which drugs he/she assumes. An example of these types of curves is given in Fig. 1.
In our analysis all types of drug lead to similar results, so in Section 4 we report only those based on ACE to avoid repetitions.

4 Joint Model of survival and drug assumption for HF patients

In 2010 Rizopoulos proposed a Joint Model (JM) for dealing with internal time-dependent covariates [7] and wrote the associated R package `JM` [8]. We use this approach in order to investigate how patients' time-to-event outcome are influenced by longitudinal data (pharmacological treatment curves).
Let $y_i(t)$ denote the value for the longitudinal outcome at time point t for the i-th

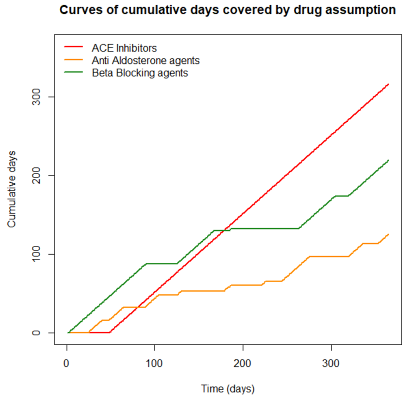


Fig. 1 Examples of time-varying curves of cumulative days covered by drug assumption for a random patient which assumes ACE (red), BB (green) and AA (orange).

subject. $y_i(t)$ is not actually observed at all time points, but only at the very specific occasions t_{ij} so the observed longitudinal data consist of the measurements $y_{ij} = \{y_i(t_{ij}), j = 1,...,n_i\}$. Let $m_i(t)$ denote the true and unobserved value of the longitudinal outcome at time t. Rizopolous supposes that there is a linear relationship between $y_i(t)$ and $m_i(t)$:

$$y_i(t) = m_i(t) + \varepsilon_i(t) = \widetilde{X}_i^T(t)\gamma + Z_i^T(t)b_i + \varepsilon_i(t) \tag{2}$$

where γ is the vector of the unknown fixed effects parameters, b_i is the vector of random effects, $\widetilde{X}_i(t)$ and $Z_i(t)$ denote row vectors of the design matrices for the fixed and random effects and $\varepsilon_i(t)$ is the normally distributed measurement error term with $\varepsilon_i(t) \sim \mathcal{N}(0,\sigma^2)$. Moreover, the random effects b_i are assumed independent of $\varepsilon_i(t)$ and normally distributed with $b_i \sim \mathcal{N}(0,D)$.
To quantify the effect of $m_i(t)$ on the risk for an event, Rizopoulos introduces the following *relative risk model*:

$$\lambda_i(t|\mathcal{M}_i(t), X_i) = \lambda_0(t)\exp\{X_i^T\beta + \alpha m_i(t)\} \tag{3}$$

where $\mathcal{M}_i(t) = \{m_i(u), 0 \le u < t\}$ denotes the history of the true unobserved longitudinal process up to time point t, $\lambda_0(\cdot)$ denotes the baseline risk function (unspecified or approximated using step functions or spline-based approaches), X_i is a vector of baseline covariates, β is the vector of regression coefficients and α is a parameter that quantifies the effect of the undelying longitudinal outcome to the survival risk for an event.
Finally, coeffients β and α have to be estimated through maximization of the log-likelihood, that is a computationally challenging task for which some numerical approximation methods are needed. In order to do that, we take advantage of the algorithms implemented in `JM` package (see [8] for details).

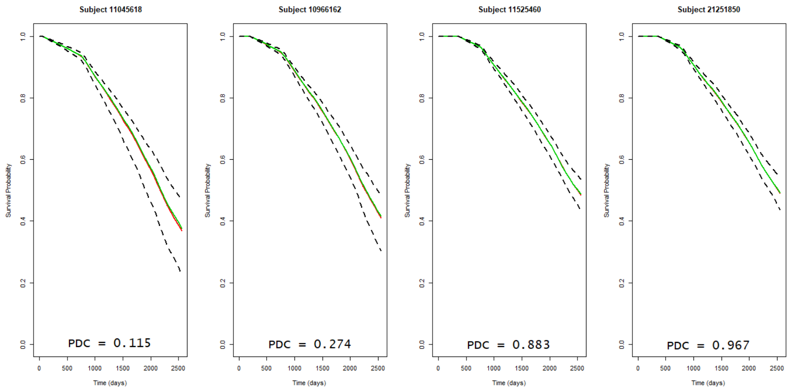


Fig. 2 Survival probability plots for female patients with 80 years old, one hospitalization and two comorbidities. From the left panel patients have a PDC of 0.115, 0.274, 0.833 and 0.967.

In our analysis the longitudinal process $m_i(t)$ is given by the square root of the value of cumulative days covered by drug assumption, whereas the survival process is adjusted according to age, gender, hospitalizations and comorbidities covariates. Moreover, we use a piecewise-constant baseline risk function and the Gauss-Hermite integration rule to approximate integrals in log-likelihood.

It results that all the covariates are significant at 5%, except for gender. In particular, being younger corresponds to a higher survival probability, whereas having a higher number of hospitalizations or of initial comorbidities corresponds to a lower survival risk, as it might be expected. We observe that the higher the value of final PDC, the higher the survival. Furthermore, having a lower PDC leads to larger confidence intervals over time, as shown in Fig. 2, so the uncertainty about the prediction of the survival outcome increases.

5 Conclusion

Modelling the drug assumption process as time-varying covariates in a joint model setting is a promising tool for exploring the effects of pharmacological treatments on survival. For example, it allows us to confirm some pharmacoepidemiological intuition as the fact that medication nonadherence is commonly associated with adverse health conditions [4] in a more suitable way.

Some improvements may be included into the model proposed in (2) in order to provide a more proper modelling of the functional covariate. Moreover, a lot of work is needed in order to include simultaneously all the treatments in a not trivial way. Limitations of administrative data should be overcome through suitable integration of administrative data with clinical registries, as proposed in [3].

References

1. Andrade, S.E., Kahler, K.H., Frech F., Chan, F.A.: Methods for evaluation of medication adherence and persistence using automated databases. In: Pharmacoepidemiology and Drug Safety, **15**, 565574 (2006).
2. Gagne, J.J., Glynn, R.J., Avorn, J., Levin, R., Schneeweiss, S. : A combined comorbidity score predicted mortality in elderly patients better than existing scores. In: Journal of Clinical Epidemiology, **64**(7), 749-59 (2011).
3. Gasperoni, F., Ieva, F., Barbati, G., Scagnetto, A., Iorio, A., Sinagra, G., Di Lenarda, A.: Multi state modelling of Heart Failure care path: a population-based investigation from Italy. In: PlosOne **12**(6): e0179176, (2017).
4. Karve, S., Cleve, M.A., Helm M., Hudson, T.J., West, D.S., Martin, B.C.: Prospective Validation of Eight Different Adherence Measures for Use with Administrative Claims Data among Patients with Schizophreniavhe. In: Value In Health, **12**(6) (2009).
5. Maggioni A.P., Spandonaro, F.: Lo scompenso cardiaco acuto in italia. Giornale italiano di cardiologia, **15**(SUPPL.2 AL N 2), 3S4S (2014)
6. Mazzali, C. et al.: Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012. In: BMC Health Service Research, **16**(234) (2016).
7. Rizopolous, D.: JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. In: Journal of Statistical Software, **35**(9), 1-33 (2010).
8. Rizopolous, D.: Package ”JM”. (June, 2017).
   URL `https://cran.r-project.org/web/packages/JM/JM.pdf`
9. World Health Organization: Introduction to Drug Utilization Research. In: WHO Library Cataloguing-in-Publication Data, (2003).

Mapping the relation between University access test and student's university performance

Test d'accesso e performance universitaria

Vincenzo Giuseppe Genova, Antonella Plaia

Abstract The access test is the most used tool to evaluate the initial preparation of an italian student before enrolling at the university. Although the necessity to select the most deserving students seem unquestionably reasonable, we have to wonder if it appears to be a good predictor of university performance. In order to answer this question, the university careers of the students enrolled at two Degree Courses at the University of Palermo (academic year 2013/2014) were analysed. The very aim of this paper is to propose a graphical tool, the Student Efficiency Nomogram, that shows the access test scores together with the first-year performance of the student.

Abstract *Lo strumento più utilizzato in Italia per verificare la preparazione iniziale di uno studente universitario è il test d'accesso. Per quanto la necessità di selezionare gli studenti più meritevoli appaia indiscutibilmente ragionevole, bisogna chiedersi se il test d'accesso risulti essere un buon predittore della futura performance universitaria. Per rispondere a tale quesito sono state analizzate le carriere universitarie degli studenti di due corsi di laurea immatricolati presso l'Università degli Studi di Palermo nell'A.A. 2013/2014. In questo lavoro proponiamo uno strumento grafico, il Nomogramma d'efficienza studentesca, per mettere in relazione i punteggi conseguiti al test d'accesso con la performance dello studente alla fine del primo anno di studi universitari.*

Key words: Selective access test, first-year university career, Student Efficiency Nomogram

Antonella Plaia, Vincenzo Giuseppe Genova
Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Palermo, Viale delle Scienze, Edificio 19, 90128 Palermo, Italy, e-mail: antonella.plaia/vincenzogiuseppe.genova@unipa.it

1 Introduction

The Ministerial Decree n. 270 enters in force on 20^{th} October 2004, establishing that an adequate initial preparation is required among the minimum prerequisites to access a degree course. The most used tool to evaluate the initial preparation of a student is the access test, which is intended to select the most deserving students who will potentially continue their studies with success.[2]

The goal of this study is to understand if there exist a relationship between the access test and the student's performance. For this purpose, the analysis was carried out on the student population of two degree courses, say A and B, enrolled at the University of Palermo in the academic year 2013/2014.

2 Data and variables

The analysed dataset was provided by Sistema Informatico d'Ateneo. The dataset consisted of 504 records in 10 variables, where each record identify a statistical unit, i.e. a student enrolled on one of the two degree courses during the academic year 2013/2014. The dataset was made up of information about the student's university career and information about the score at the access test, where the latter were provided by one of the Companies in charge for the Access tests at the University of Palermo. The variables taken into account were the number of ECTSs earned the student at the end of the first academic year, the average mark at the end of the first year, and the scores for each area of knowledge. The areas of knowledge for Degree course A were Law and Economics; Italian; Logic and Philosophy; History. For degree course B they were Biology; Chemistry; Physics; Mathematics.

The analysis was focused on the first-year university career, as it has been shown that the first year is a good predictor of student performance. According to our goal, we decided to calculate the number of years expected to obtain the degree as the ratio between the number of ECTSs to get the degree (180-300 in Italy) and the number of ECTSs actually earned by the *i-th* student at the end of the first year.[1]. Actually, this is a rough prevision of the time to get the degree, but this is not the aim of this paper.

3 Student Efficiency Nomogram

According to our goal, we decided to propose a graphical method which can simultaneously evaluate different student performance indicators. The graphical method shows simultaneously: the student's average mark; the number of ECTSs earned at the end of the first year; the number of years expected to obtain degree; and the access test score.

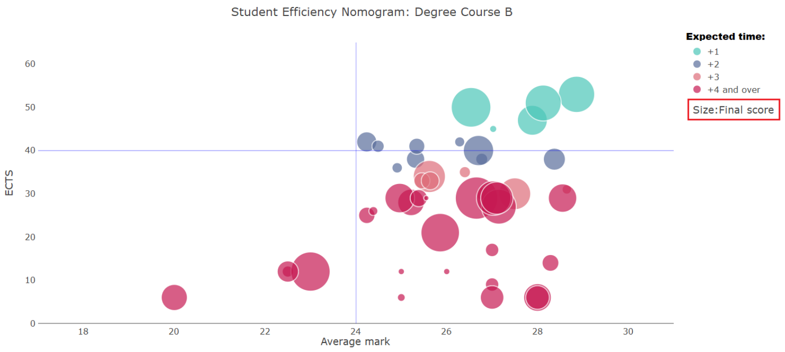


Fig. 1 Student efficiency Nomogram: Degree Course B, final score

As shown in Figure 1, the Student Efficiency Nomogram can be represented in a Cartesian coordinate system. On the horizontal axis it shows the student's average mark and on the vertical axis it shows the number of credits acquired by the student at the end of the first year. The origin of the Cartesian axes is fixed at the point of coordinates (24, 40), this choice was made in accordance with the MIUR directives. Indeed, according to the Ministry point of view, a student earning more than forty ECTSs at the end of the year is considered an efficient student.[3]

Looking at average mark, it has been decided to divide the range of values into two parts. Where, under 24, we identify a student below average and vice versa a student above average. In addition, the pair of points ($ECTS_i$, $Average\ mark_i$) identifies the number of ECTSs and the average mark of the *i-th* student at the end of the first academic year.

According to this graphical method, we have three possible scenarios: *i)* the student is highly efficient ($ECTS > 40, Average\ mark > 24$); *ii)* the student is inefficient ($ECTS < 40$, $Average\ mark < 24$); *iii)* the student is in an in between situation: efficient, but he/she doesn't show high marks or he/she is not efficient but with high mark ($ECTS > 40$, $Average\ mark < 24$ or $ECTS < 40$, $Average\ mark > 24$).

In addition, we were also interested in evaluating the information regarding the number of years expected to obtain degree and the access test score. In order to take into account this information in the graph, we assign to each observation a colour which identifies the time expected for the degree and we scale the point size proportionally to the scores gained at the access test. So, if on the one hand this tool allows to understand if the student is efficient or not, on the other hand it allows to verify if the access test was a good tool to evaluate students at the entrance. Indeed, if it was a good tool, you would expect to identify in the upper right quadrant, students with a high access test score, while in the lower left quadrant, students with a low score.

4 Results

From Figures 1 and 2 e) we can see that only 20% of the statistical units falls into the efficiency region. It is also necessary to point out that the size of the point, which expresses the total score obtained in the access test, seems to be randomly distributed among the regions of the graph. Moreover, with the exception of only 3 statistical units for Degree Course A, none of the observed students, according to the indicator used in this paper, will take is degree in time (3-5 years).

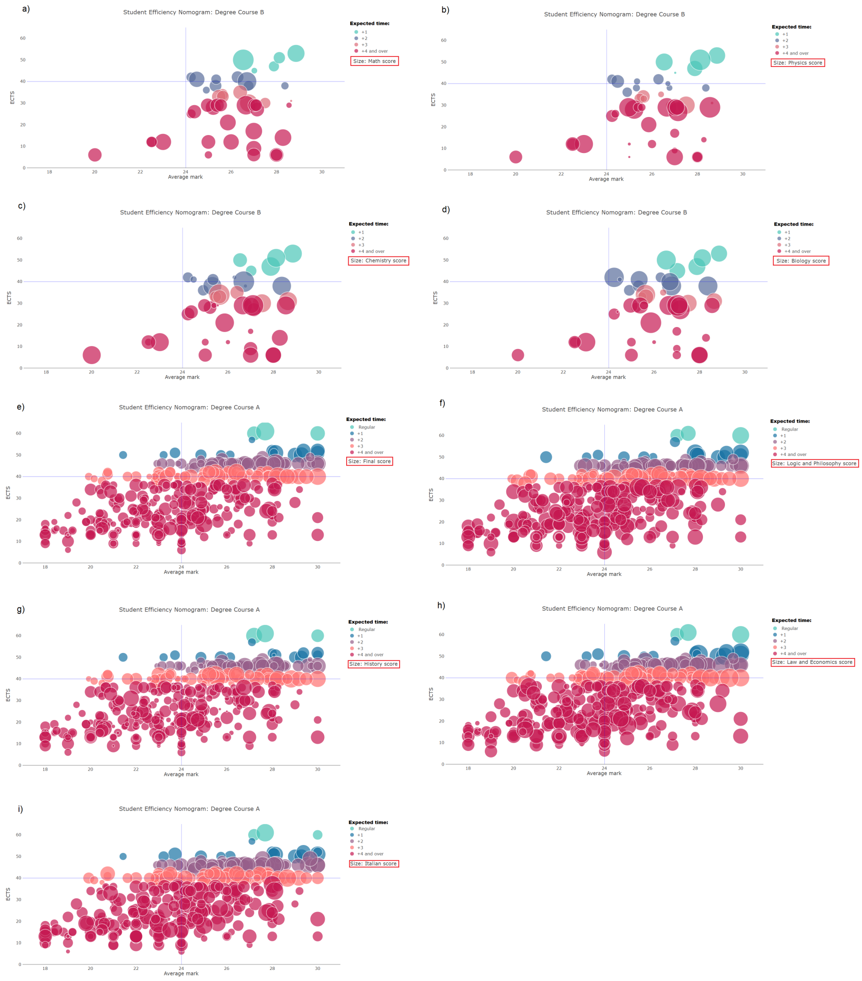


Fig. 2 Student efficiency Nomogram, Degree Courses A and B

We were interested in evaluating each areas of knowledge and in particular, we wanted to understand if at least one of the areas of knowledge of the access test was a good predictor of the student's career.

Figure 2 shows the Student Efficiency Nomograms of both the degree courses related to each areas of knowledge. We can notice that all areas of knowledge do not appear as good predictors of the university career performance. Anyway, Biology and History seem to have a greater discriminant effect, but not enough to call them predictors of student's performance.

To validate our graphical method a zero-one inflated (Course A) and a zero inflated (Course B) model were applied [4, 5], and in particular we have analyzed the relationship between the fraction of ECTSs and the set of explanatory variables for both the Degree Courses.

Table 1 Model coefficients of Degree Course A

μ **(fraction)**	**Estimate**	**S.e**.	***p-value***
Intercept	−2.70	0.39	0.000
Law and Economics score	0.45	0.29	0.12
High school mark	0.02	0.004	0.000
Classical high school	0.24	0.21	0.26
Teacher-training high school	−0.12	0.27	0.64
ITC high school	−0.36	0.29	0.21
Scientific high school	0.18	0.22	0.41
Male	−0.19	0.09	0.03
α **(zero-inflated)**	**Estimate**	**S.e.**	***p-value***
Intercept	5.40	1.66	0.001
History score	−1.99	1.25	0.11
Law and Economics score	−2.34	1.34	0.08
High school mark	−0.04	0.02	0.01
Classical high school	−2.53	0.62	0.000
Teacher-training high school	−0.68	0.67	0.30
ITC high school	−0.88	0.74	0.23
Scientific high school	−2.09	0.63	0.001
τ **(one-inflated)**	**Estimate**	**S.e.**	***p-value***
Intercept	−18.46	8.79	0.03
High school mark	0.15	0.09	0.09
ϕ **(precision)**	**Estimate**	**S.e.**	***p-value***
Intercept	−0.43	0.04	0.000

Looking at Table 1 (Degree Course A), we can notice that the History score, and the Law and Economics score reduce the probability of obtaining zero ECTSs at the end of the first year, but they haven't a significant effect on average (Table 1 - top section). Table 2 (Degree Course B) shows a reduction in probability of obtaining zero ECTSs due to the Physics score Table 2 - mid section), but we can notice only a significant positive effect on average (Table 2 - top section) due to the Biology score. The same analysis was carried out analyzing the relationship between the *Average mark* and the set of explanatory variables, whithout obtaining results significantly different from those in Tables 1 and 2.

Table 2 Model coefficients of Degree Course B

μ **(fraction)**	**Estimate**	**S.e.**	***p-value***
Intercept	-8.72	1.29	0.000
Physics score	1.06	0.66	0.11
Chemistry score	-2	0.89	0.03
Biology score	3.39	1.55	0.03
High school mark	0.05	0.007	0.000
Classical high school	1.13	0.32	0.001
Scientific high school	1.02	0.19	0.000
Male	0.76	0.17	0.000
Public school	1.13	0.62	0.07
α **(zero-inflated)**	**Estimate**	**S.e.**	***p-value***
Intercept	-1.48	5.53	0.78
Classical high school	-4.15	2.03	0.04
Scientific high school	-3.55	1.68	0.04
Physics score	-6.34	3.81	0.10
High school mark	0.06	0.06	0.32
ϕ **(precision)**	**Estimate**	**S.e.**	***p-value***
Intercept	20.37	0.02	0.000
Classical high school	-19.14	0.31	0.000
Scientific high school	-18.28	0.26	0.000

According to these results, we can conclude that our graphical tool is an easy way to analyze the student performance in relationship with access test.

References

1. Attanasio M., Boscaino G., Capursi V., Plaia A.: Can the students' career be helpful in predicting an increase in universities income?. In: Statistical Models for Data Analysis, P. Giudici, S. Ingrassia, M. Vichi (Ed.) pp. 9-16. Springer (2013)
2. Ministero dell'Istruzione dell'Università e della Ricerca:Decreto n.270, Modifiche al regolamento recante norme concernenti l'autonomia didattica degli atenei. In: Gazzetta Ufficiale n.266 (2004)
3. Ministero dell'Istruzione dell'Università e della Ricerca: Decreto autovalutazione, valutazione, accreditamento iniziale e periodico delle sedi e dei Corsi di Studio. In: Gazzetta Ufficiale, registro decreti prot. n. 0000987, Italia (2012)
4. Ospina, R., Ferrari, S.: Inflated beta distributions. Statistical Papers, 51, 111-126. (2010)
5. Ospina, R., Ferrari, S. L.: A general class of zero-or-one inflated beta regression models. In Computational Statistics and Data Analysis 56 (p. 1609-1623). Elsevier. (2012).

Multivariate analysis of marine litter abundance through Bayesian space-time models

Analisi multivariata dell'abbondanza di rifiuti marini attraverso modelli bayesiani spazio-temporali

C. Calculli, A. Pollice, L. Sion, and P. Maiorano

Abstract This work focuses on the analysis of abundance data for marine litter categories, collected during trawl surveys regularly conducted at local scale, in the Central Mediterranean. Here marine litter data are modeled in order to estimate the effects affecting the dynamics of litter assemblages at different spatio/temporal scales. A correlated response model with latent variables is proposed. This modeling approach is particularly suitable to infer potential environmental covariates while controlling for correlation between litter categories and providing a method for residual ordination. MCMC estimation is implemented within the Bayesian hierachical framework that allows to integrate environmental and anthropogenic processes into a single model.

Abstract *Questo lavoro riguarda l'analisi delle abbondanze di diverse categorie di rifiuti marini, censiti durante campagne di pesca a trascico condotte su scala locale nel Mediterraneo centrale. In questo ambito, i dati sui rifiuti sono modellati al fine di investigare i fattori che influiscono sulle dinamiche di aggregazione dei rifiuti su scala spazio/temporale. Viene presentato un modello con risposte correlate che rappresenta un approccio particolarmente idoneo per investigare l'influenza di covariate ambientali, controllare la correlazione tra categorie di rifiuti e fornire un metodo per l'ordinamento residuale. Stime MCMC sono implementate nel contesto bayesiano gerarchico che permette di integrare processi ambientali e antropogenici in un unico modello.*

Key words: Bayesian hierarchical modeling approach; Correlated response model; Latent variables; Marine Litter

Crescenza Calculli, Alessio Pollice
Department of Economics and Finance, University of Bari Aldo Moro, Largo Abbazia S. Scolastica, Bari, Italy, e-mail: crescenza.calculli@uniba.it, alessio.pollice@uniba.it

Letizia Sion, Porzia Maiorano
Department of Biology, University of Bari Aldo Moro, Via E. Orabona 4, Bari, Italy, e-mail: letizia.sion@uniba.it, porzia.maiorano@uniba.it

1 Introduction

Marine litter has recently become a recognized global and local ecological concern that might jeopardize the status of marine ecosystems. The debris quantity and its distribution on the Mediterranean seafloor are still not well-known although the effects on the sea bed and on the marine living communities are pretty clear [1]. Experimental bottom trawl surveys carried out in the Mediterranean basin in the last years represent a valuable source of information about wastes; litter typologies can be seen as special items caught by the trawl net together with marine species. While single-species distribution models have been commonly used to explain and predict the response of different taxa to environmental variation, the analysis at the community-level is still lacking. Some innovative approaches that explicitly acknowledge the multivariate nature of species assemblages were recently proposed [2, 4]. These approaches model the actual processes that determine the assemblage of community samples, taking into account for the various sources of correlation across species. In this study we analyze multivariate litter abundance data using a correlated response model in the spirit of [2]. This model merges univariate Generalized Linear Models with latent variables to account for the residual correlation across litter categories, *e.g.* due to environmental interactions or unaccounted covariates. Latent variables provide a method for "residual ordination". The whole implementation is performed in a hierarchical Bayesian framework.

2 Materials and Methods

2.1 Study area and data

Litter data are collected during experimental trawl surveys conducted from 2013 to 2017 in the North-Western Ionian Sea as a complementary (voluntary) activity of the international project MEDITS (MEDiterranean International Trawl Surveys). The North-Western Ionian is the deepest sea in the Mediterranean basin characterized by a complex geomorphology and the presence of important fisheries as well as main harbours. An increasing touristic activity is developing along the Ionian coasts, thus the sea bottoms are here exposed to a strong increase in anthropogenic impact.The same 70 depth-stratified hauls are carried out between 10 and 800 m in depth every year, summing to 350 hauls in 5 years. Wastes caught during the trawl surveys are classified in 8 categories: plastic, rubber, metal, glass/ceramic, cloth/natual fibres, processed wood, paper/cardboard, other/unspecified. The number of collected items for each litter category was scaled to the swept surface unit (1 km^2), thus obtaining density indices (N/km^2) for each litter category and survey at every haul location. Litter density is a semi-continuous zero-inflated non-negative variable. Preliminarily, to investigate factors influencing the density of litter categories, we consider only the seafloor depth as environmental covariate.

2.2 Statistical framework

To investigate assemblage of litter in terms of density and composition, a mixture model where latent variables are included alongside the measured covariates is proposed. In particular, densities of litter categories are jointly modeled as semi-continuous zero-inflated multivariate responses assuming the Tweedie distribution model. The mean density μ_{ij} of the j-th litter category at the i-th haul is specified by the following mixture model:

$$g(\mu_{ij}) = \alpha_1(t_i) + \alpha_2(s_i) + \beta_{0j} + \sum_{k=1}^{p} \beta_{jk} X_{ik} + z_i' \theta_j \quad i = 1, \dots, 350; \quad j = 1, \dots, 8 \quad (1)$$

where $g(\cdot)$ is the link function, $\alpha_{1,2}(\cdot)$ are effects adjusting for differences in site and time (year) on the overall litter density, β_{0j} is the litter type-specific intercept and β_{jk} is the type-specific regression coefficient of the k-th covariate (preliminarily, only the seafloor depth). Finally, $z_i = (z_{i1}, \dots, z_{iq})'$ is a q-dimensional vector of latent variables, while $\theta_j = (\theta_{j1}, \dots, \theta_{jq})$ are the corresponding litter type-specific loadings. Latent variables can be considered here as missing informative predictors for the multivariate response inducing residual correlation between litter categories. Independent weakly informative $N \sim (0, 10)$ priors were assumed for all site and time effects, type-specific intercepts, type-specific regression coefficients, latent variables and loadings. Uniform priors $U \sim (0, 30)$ are adopted for all dispersion and variance parameters in the model. Inferences for model in Eq. 1 were implemented by the `boral` package [2] that provides an interface between `R` and `JAGS` [5] for multi-species models with latent variables.

3 Results

The model in Eq.1 was fitted with 1 to 3 latent variables and with fixed or random site and time effects. All model estimates were obtained using 20,000 iterations, discarding the first 5,000. The Geweke diagnostic and the graphical inspection of trace plots provided clear evidence of the convergence of MCMC chains for all model parameters. As reported in Table 1, the best models in terms of lowest BIC consider random site/time effects. Results for the model with two latent variables, that also enable to draw a scatterplot of the ordinations as for distance-based techniques, suggest a positive correlation between plastic and glass litter due to depth (Figures 1A-B). Strong, positive residual correlations are observed: plastic is correlated with all other materials except for metal and other/unspecified wastes. Estimated spatial effects represented in Figure 1C, allow to identify some *hot-spots* assemblages for all litter categories. This work represents a starting point for the analysis of space/time structured multivariate litter data. Further developments include the selection among

Table 1: Values of the BIC for models with 1-3 latent variables (LVs) and fixed/random site and time effects

	site/time effects	
	fixed	random
1 LV	8731.59	8387.27
2 LVs	8777.11	8434.66
3 LVs	8833.30	8486.78

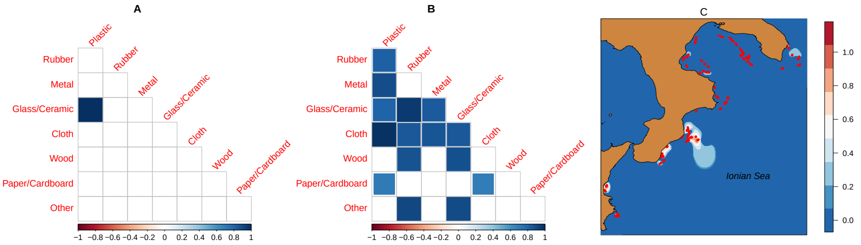


Fig. 1: (A) Correlations between litter categories due to seafloor depth; (B) Residual correlations (we report correlations with 95% credible intervals excluding zero); (C) Predictions of spatial effects of sites.

a larger number of environmental predictors by Boosted Regression Trees (BRT) as already suggested in [3] for ecological data.

Acknowledgements C. Calculli and A. Pollice were supported by the PRIN2015 project EphaStat - *Environmental processes and human activities: capturing their interactions via Statistical methods*, funded by the Italian Ministry of University and Research (MIUR).

References

1. Gall, S.C., Thompson,R.C.: The impact of debris on marine life. *Marine Pollution Bulletin* **92**, 170–179 (2015)
2. Hui, F. K. C.: boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R. *Methods in Ecology and Evolution* **7**, 744–750 (2016)
3. Leathwick, J. R., Elith, J., Francis, M., Hastie T., Taylor, P.: Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series* **321**, 267–281 (2006)
4. Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. and Abrego, N.: How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letter* **20**, 561–576 (2017)
5. Plummer, M. et al.: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* Vienna, Austria (2003)

Power Priors for Bayesian Analysis of Graphical Models of Conditional Independence in Three Way Contingency Tables

Power Prior per l'analisi bayesiana di modelli grafici di indipendenza condizionale in tabelle di contingenza a tre vie

Katerina Mantzouni, Claudia Tarantola and Ioannis Ntzoufras

Abstract In this paper we illustrate a comprehensive Bayesian analysis of graphical models of conditional independence, involving suitable choices of prior parameters, estimation, model determination, as well as the allied computational issues for three way contingency tables. Each conditional independence model corresponds to a particular factorization of the cell probabilities and a conjugate analysis based on Dirichlet prior can be performed. Unit information interpretation priors are used as a yardstick in order to identify and interpret the effect of any other prior distribution used. The posterior distributions of the graphical models parameters, are obtained using simple Markov chain Monte Carlo (MCMC) schemes. A real data application will be analytically presented in the poster.

Abstract *In questo lavoro sviluppiamo un'analisi Bayesiana completa di modelli grafici di indipendenza condizionale, inclusa la scelta di opportuni parametri iniziali, stima, scelta del modello, nonché i problemi computazionali collegati per le tabelle di contingenza a tre vie. Ogni modello di indipendenza condizionale corrisponde a una fattorizzazione particolare delle probabilit di cella e puó essere eseguita un'analisi coniugata utilizzando prior Dirichlet. Prior di tipo Unit information sono usate come termine di paragone per identificare e interpretare l'effetto delle altre prior proposte. Le distribuzioni a posteriori dei parametri dei modelli grafici, sono ottenute utilizzando semplici schemi Markov della catena Monte Carlo (MCMC). Nel poster verrà descritta un'applicazione a dati reali.*

Katerina Mantzouni
Department of Statistics, Athens University of Economics and Business, Athens, Greece e-mail: aikmantz@aueb.gr

Claudia Tarantola
Department of Economics and Management, University of Pavia, Italy e-mail: claudia.tarantola@unipv.it

Ioannis Ntzoufras
Department of Statistics, Athens University of Economics and Business, Athens, Greece e-mail: ntzoufras@aueb.gr

Key words: Conditional Independence Parametrization, Contingency Tables, Graphical Models, Power Prior

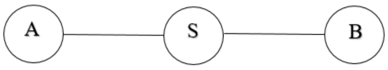
1 Graphical Models of Conditional Independence

The use of graphical models to describe association between categorical variables dates back to the work of Darroch *et al.* (1980), where graphical log-linear models were introduced. Graphical models turns out to be an efficient methodology for categorical data analysis and in this paper we focus of on graphical models of conditional independence. Conditional independence is important when modelling highly complex systems.

A undirected graph $G = (\mathscr{V}, E)$ is characterized by a vertex set $\mathscr{V}$ and an edge set E. For sets A, B and $S \subset \mathscr{V}$, A nd B are conditionally independent given S, whenever A and B are separated by S, all path in the graph connecting A and B pass through S. Discrete random variables X_v, $v \in \mathscr{V}$ defined by undirected graphs and the cliques of the graph correspond to the maximal terms in log-linear model.

Dawid and Lauritzen (1993) described the Bayesian framework for decomposable models. Figure 1 shows such a model which also embodies the assumption that A and B are independent given S.

Fig. 1 A decomposable undirected graphical model



The cliques of this graph are A, S and S, B and there is a single intersection S. The factorization form of the joint distribution given by:

$$P(ABS) = \frac{P(AS)P(SB)}{P(S)} = P(A|S)P(B|S)P(S) \tag{1}$$

In this paper we focus in decomposable models in three way contingency tales, for which the graph is chordal. These are the closed form log-linear models and their parameter can be estimated without the utilization of iterative methods.

2 Conjugate Priors

In the following we work using conjugate priors on the probability parameters and then calculate the corresponding log-linear parameters using Monte Carlo schemes. For the specification of the prior distribution on the probability parameter vector we initially consider a Dirichlet distribution with parameters $\alpha = \big(\alpha(i), i \in \mathscr{I}\ \big) = \big(\alpha_{abc}, a = 1, \ldots, |\mathscr{I}_A|, b = 1, \ldots, |\mathscr{I}_B|, c = 1, \ldots, |\mathscr{I}_C|\ \big)$ for the vector of the joint probabilities of the full table π. Hence, for the full table $\pi \sim \mathscr{D}i(\alpha)$ with prior density given by

$$
\begin{aligned}
f(\pi) &= \frac{\Gamma(\alpha)}{\prod\limits_{i\in\mathscr{I}} \Gamma\Big(\alpha(i)\Big)} \prod_{i\in\mathscr{I}} \pi(i)^{\alpha(i)-1} \\
&= \frac{\Gamma(\alpha)}{\prod\limits_{a=1}^{|\mathscr{I}_A|}\prod\limits_{b=1}^{|\mathscr{I}_B|}\prod\limits_{c=1}^{|\mathscr{I}_C|} \Gamma\Big(\alpha_{abc}\Big)} \prod_{a=1}^{|\mathscr{I}_A|}\prod_{b=1}^{|\mathscr{I}_B|}\prod_{c=1}^{|\mathscr{I}_C|} \pi_{abc}^{\big(\alpha_{abc}-1\big)} = f_{\mathscr{D}i}(\pi;\,\alpha)
\end{aligned}
\tag{2}
$$

where $f_{\mathscr{D}i}\big(\pi;\ \alpha\big)$ is the density function of the Dirichlet distribution evaluated at π with parameters α and $\alpha = \sum_{i\in\mathscr{I}} \alpha(i)$.

When no prior information is available then we usually set all $\alpha(i) = \frac{\alpha}{|\mathscr{I}|}$ resulting to

$$
E\big[\pi(i)\big] = \frac{1}{|\mathscr{I}|} \ \text{ and } \ V\big[\pi(i)\big] = \frac{|\mathscr{I}|-1}{|\mathscr{I}|^2(\alpha+1)}\ .
$$

Small values of α increase the variance of each cell probability parameter. Usual choices for α are the values $|\mathscr{I}|/2$ (Jeffrey's prior), $|\mathscr{I}|$ and 1 (corresponding to $\alpha(i)$ equal to $1/2$, 1 and $1/|\mathscr{I}|$ respectively); for details see Dellaportas and Forster (1999). The choice of this prior parameter value is of prominent importance for the model comparison due to the well known sensitivity of the posterior model odds and the Bartlett-Lindley paradox (Lindley, 1957, Bartlett, 1957). Here this effect is not so adverse, as for example in usual variable selection for generalized linear models, for two reasons. Firstly, even if we consider the limiting case where $\alpha(i) = \frac{\alpha}{|\mathscr{I}|}$ with $\alpha \to 0$, the variance is finite and equal to $(|\mathscr{I}|-1)/|\mathscr{I}|^2$. Secondly, the distributions of all models are constructed from a common distribution of the full model/table making the prior distributions 'compatible' across different models (Dawid and Lauritzen, 2000 and Roverato and Consonni, 2004).

3 Power Priors

In graphical model literature there is a debate about the use of conjugate priors based on Dirichlet distributions; see for example in Steck and Jaakkola (2002), Steck (2008) and Ueno (2008). The parameters of the Dirichlet prior should be carefully specified. In order to do this we adopt ideas based on the power prior approach of Ibrahim and Chen (2000) and Chen *et al.* (2000). We use their approach to advocate sensible values for the Dirichlet prior parameters on the full table and the corresponding induced values for the rest of the graphs.

3.1 Specification of Prior Parameters Using Imaginary Data

Let us consider imaginary set of data represented by the frequency table $n^* = (n^*(i), i \in \mathscr{I})$ of total sample size $N^* = \sum_{i \in \mathscr{I}} n^*(i)$ and a Dirichlet 'pre-prior' with all parameters equal to α_0. Then the unnormalized prior distribution can be obtained by the product of the likelihood of n^* raised to a power w multiplied by the 'pre-prior' distribution. Hence

$$\begin{aligned} f(\pi) &\propto f(n^*|\pi)^w \times f_{\mathscr{D}i}\big(\pi;\ \alpha(i) = \alpha_0, i \in \mathscr{I}\big) \\ &\propto \prod_{i \in \mathscr{I}} \pi(i)^{wn^*(i)+\alpha_0-1} \\ &= f_{\mathscr{D}i}\big(\pi;\ \alpha(i) = wn^*(i) + \alpha_0, i \in \mathscr{I}\big)\ . \end{aligned} \tag{3}$$

Using the above prior set up, we expect a priori to observe a total number of $wN^* + |\mathscr{I}|\alpha_0$ observations. The parameter w is used to specify the steepness of the prior distribution and the weight of belief on each prior observation. For $w = 1$ then each imaginary observation has the same weight as the actual observations. Values of $w < 1$ will give less weight to each imaginary observation while $w > 1$ will increase the weight of believe on the prior/imaginary data. Overall the prior will account for the $(wN^* + |\mathscr{I}|\alpha_0)/(wN^* + N + |\mathscr{I}|\alpha_0)$ of the total information used in the posterior distribution. Hence for $w = 1$, $N^* = N$ and $\alpha_0 \to 0$ then both the prior and data will account for 50% of the information used in the posterior.

For $w = 1/N^*$ then $\alpha(i) = p^*(i) + \alpha_0$ with $p^*(i) = n^*(i)/N^*$, the prior data n^* will account for information of one data point while the total weight of the prior will be equal to $(1 + |\mathscr{I}|\alpha_0)/(1 + N + |\mathscr{I}|\alpha_0)$. If we further set $\alpha_0 = 0$, then the prior distribution (3) will account for information equivalent to a single observation. This prior set-up will be referred in this paper as the unit information prior (UIP). When no information is available, then we may further consider the choice of equal cell frequencies $n^*(i) = n^*$ for the imaginary data in order to support the simplest possible model under consideration. Under this approach $N^* = n^* \times |\mathscr{I}|$ and $w = 1/N^* = \frac{1}{n^* \times |\mathscr{I}|}$ resulting to

$$\pi \sim \mathscr{D}i\Big(\alpha(i) = 1/|\mathscr{I}|, i \in \mathscr{I}\Big)\ .$$

The latter prior is equivalent to the one advocated by Perks (1947). It has the nice property that the prior on the marginal parameters does not depend on the size of the table. This property is retained for any prior distribution of type (3) with $w^* = 1/N^*$, $p^*(i) = 1/|\mathscr{I}|$ and $\alpha_0 \propto 1/|\mathscr{I}|$.

3.2 Comparison of Prior Set-ups

Since Perks' prior (with $\alpha(i) = 1/|\mathscr{I}|$) has a unit information interpretation, it can be used as a yardstick in order to identify and interpret the effect of any other prior distribution used. Prior distributions with $\alpha(i) < 1/|\mathscr{I}|$, or equivalently $\alpha < 1$, result in larger variance than the one imposed by our proposed unit information prior and hence they a posteriori supports more parsimonious models. On the contrary, prior distributions with $\alpha(i) > 1/|\mathscr{I}|$, or $\alpha > 1$, result in lower prior variance and hence they a posteriori support models with more complicated graph structure. So the variance ratio between a Dirichlet prior with $\alpha(i) = \alpha/|\mathscr{I}|$ and Perks prior is equal to

$$VR = \frac{V\big(\pi(i)\big|\ \alpha(i) = \frac{\alpha}{|\mathscr{I}|}\big)}{V\big(\pi(i)\big|\ \alpha(i) = |\mathscr{I}|^{-1}\big)} = \frac{2}{\alpha+1}\ .$$

In this parer we considered the comparison of the information from the following prior choices:

(i) the Jeffrey's prior with $\alpha(i) = 1/2$;
(ii) the Unit Expected Cells prior (UEC) with $\alpha(i) = 1$;
(iii) the Unit Information Prior (UIP) which is derived by a power prior with $\alpha(i) = p^*(i)$, $w = 1/N^*$ and $a_0 = 0$; where $p^*(i)$ is the sample proportion of cell i estimated from a set of imaginary data $n^*(i)$;
  (a) Perks' prior (UIP-Perks') with $\alpha(i) = 1/|\mathscr{I}|$ which is equivalent to UIP coming from a table of imaginary data with all cell frequencies equal to one;
  (b) the Unit Information Empirical Bayes Prior (UI-EBP), which is derived by UIP with $p^*(i)$ set equal to the sample proportions $p(i) = n(i)/N$.

It is observed that Jeffreys' prior variance is lower than the corresponding Perks' prior. The reduction is even greater for the Unit Expected Cell prior reaching. Finally, for the Empirical Bayes prior, based on the UIP approach, the prior variance for each $\pi(i)$ is equal to $V[\pi(i)] = \frac{1}{2}p(i)\big(1-p(i)\big)$. Hence it depends on the observed proportion and can vary from zero (if $p(i) = 0$ or 1) to $1/8$ if $p(i) = 1/2$. For values in the interval $(0.058, 0.942)$ the variance of the UI-EBP is higher than the corresponding UIP variance reaching its maximum when $p(i) = 1/2$ where it is 4.6 times the corresponding UIP prior variance. For $p(i) = 0.058$ or 0.942 the variances of the UIP and UI-EBP are equal while for the remaining values, UIP variance is higher.

3.3 Illustrative examples

We consider a data set presented by Healy (1988) regarding a study on the relationship between patient condition (more or less severe), assumption of antitoxin (yes or not) and survival status (survived or not); see Table 1.

Table 1 Antitoxin data

		Survival (S)	
Condition (C)	Antitoxin (A)	No	Yes
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

We compare posterior model probabilities under the four different prior setups, we compare the results obtained with our yardstick prior, the UIP-Perks' prior ($\alpha(i) = 1/|I|$), with those obtained using Jeffrey's ($\alpha(i) = 1/2$), Unit Expected Cell ($\alpha(i) = 1$), and unit information Empirical Bayes ($\alpha(i) = p(i)$) priors.Under all prior assumptions the maximum a posteriori model (MAP) is SC+A , assuming the conditional independence of Antitoxin from the remaining two variables.

References

1. Bartlett, M. S.: A comment on Lindley's statistical paradox. Biometrika **44**, pp. 533-534. (1957)
2. Chen, M.H., Ibrahim, J.G. and Shao, Q. M. Power prior distributions for generalized linear models. Journal of Statistical Planning and Inference **84**, pp. 121-137. (2000)
3. Darroch, J.N., Lauritzen, S.L., and Speed, T.P.: Markov Fields and Log-Linear Interaction Models for contingency Tables. Annals of Statistics **8**, pp. 522-539. (1980)
4. Dawid A.P. and Lauritzen S.L.: Hyper-Markov laws in the statistical analysis of decomposable graphical models. Annals of Statistics **21**, pp. 1272-1317.(1993)
5. Dawid, A.P. and Lauritzen, S.L.: Compatible prior distributions. In Bayesian Methods with Applications to Science Policy and Official Statistics. The sixth world meeting of the International Society for Bayesian Analysis (ed. E.I. George), pp. 109-118.(2000) `http://www.stat.cmu.edu/ISBA/index.html`.
6. Dellaportas, P. and Forster, J.J.: Markov Chain Monte Carlo Model determination for hierarchical and graphical log-linear Models. Biometrika **86**, pp. 615-634.(1999)
7. Dellaportas, P., Forster, J.J. and Ntoufras I.: On Bayesian model and variable selection using MCMC. Statistics and Computing **12**, pp. 27-36.(2002)
8. Healy, M.J.R. Glim: An Introduction, Claredon Press, Oxford, UK.(1988).
9. Ibrahim J.G. and Chen M. H.: Power Prior Distributions for Regression Models. Statistical Science **15**, pp. 46-60.(2000)
10. Lindley, D. V.: A statistical paradox. Biometrika, 44, 187192. (1957)
11. Perks, W.: Some observations on inverse probability including a new indifference rule. Journal of the institute of actuaries **73**, pp. 285-334. (1947)
12. Roverato A. and Consonni G.: Compatible Prior Distributions for DAG models. Journal of the Royal Statistical Society B **66**, pp. 47-61. (2004)
13. Steck, H. and Jaakkola, T.: On the Dirichlet Prior and Bayesian Regularization. NIPS, pp. 697-704. (2002)
14. Steck, H.: Learning the Bayesian Network Structure: Dirichlet Prior vs Data. Proceedings of the conference on Uncertainty in Artificial Intelligence, pp. 511-518. (2008)
15. Hamburger, C.: Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. Ann. Mat. Pura. Appl. **169**, pp. 321–354. (1995)
16. Ueno, M.: Learning Likelihood-Equivalence Bayesian Networks Using an Empirical Bayesian Approach. Behaviormetrika, 35 pp. 115-135.(2008)

Random Garden: a Supervised Learning Algorithm

Giardino Casuale: un Algoritmo di Apprendimento Supervisionato

Ivan Luciano Danesi, Valeria Danese, Nicolò Russo and Enrico Tonini

Abstract Classification and Regression Trees model and two other tree-based models are considered. These latter tree-based models are the Random Forest and the Random Garden, presented in this work. The feature selection impact on the different algorithms is investigated. The described procedures are applied to 18 Customer Relationship Management data sets constructed in Banking field. The goal is binary classification. Our results show that the best algorithm depends on data set characteristics, as dimensions, proportion of success events and the application of feature selection.

Abstract *Gli alberi di classificazione e di regressione e altri due modelli ad albero sono considerati. Questi ultimi sono la Foresta Casuale e il Giardino Casuale, presentato in questo lavoro. L'impatto della selezione delle variabili è investigato. Le procedure descritte sono applicate a 18 tabelle costruite per la gestione della relazione con i clienti in un contesto bancario. Lo scopo è una classificazione binaria. I nostri risultati mostrano come il migliore algoritmo dipenda dalle caratteristiche dei dati, come dimensioni, proporzione di successi e l'applicazione di procedure di selezione delle variabili.*

Key words: Classification and Regression Trees, Customer Relationship Management, Feature Selection, Random Garden, Random Forest, Tree Bagging.

Ivan Luciano Danesi, Valeria Danese and Nicolò Russo
Data & Analytics Data Science
UniCredit Business Integrated Solutions S.C.p.A
Via Livio Cambi 1, 20151 Milano, Italy, e-mail: ivanluciano.danesi@unicredit.eu,
valeria.danese@unicredit.eu, nicolo.russo@unicredit.eu

Enrico Tonini
Independent Author e-mail: enrico.tonini.stat@gmail.com

1 Introduction

In Big Data era, the Customer Relationship Management (CRM) has been receiving a great deal of attention. CRM can leverage on several data sources. During the years, clients' data have been collected in many forms. Today there is the technology for the storage of all of these data in the same place, as well as the tools for merging and using them.

In this work we apply tree-based classification models to different data sets of an important financial institution. Such data sets are constructed in CRM field. The goal of classification applications is to discriminate between success and failure events. The success events are defined as CRM relevant occurrences (*e.g.* claims or product purchases).

Firstly, the features useful for the modeling step are selected. During feature selection step, we consider either a supervised and an unsupervised approach.

Secondly, three tree-based models are applied. A Classification and Regression Tree (CART) model is used as benchmark for our evaluation. The *random forest* algorithm, the most diffused algorithm for bagging and ensembling trees, is then considered. Finally, the *Random Garden*, a CART bagging designed for high dimensional data, is applied.

The tree-based models are estimated on the different data sets, with and without feature selection.

In Section 2 the feature selection techniques and the algorithms are briefly outlined. The results are presented in Section 3 and discussed in Section 4.

2 Section Heading

2.1 Feature Selection

Usually the higher the number of features that can be collected, the more relevant is discriminating between valuable predictors (to be kept) and not (to be discarded). The consequences of including non-informative features in a model may be different, depending on the selected algorithm. Furthermore, in presence of a high number of features, some relevant ones could not be included in trees splitting process enough times to correctly determine the results, as pointed out in [5].

Feature selection procedures can be mainly divided in two categories: wrapper methods and filter methods ([4]). The former involves the use of predictive algorithms while the latter approach analyzes the features one by one and keeps only the ones satisfying a defined rule.

As wrapper methods, we consider the permutation test and as filter one the analysis of variance and correlations predictors. From now on, we refer them as our features selection (FS) procedures through which we generate the datasets.

2.2 *Considered Tree-Based Algorithm*

2.2.1 Classification and Regression Trees

Classification and Regression Trees (CART) are constructed to generate a response or a class Y using a set of inputs $X_1, \dots, X_p$ by means of binary splits (see [2] for details).

2.2.2 Random Forest

Random forests (RF) [1] improve predictive accuracy by generating a large number of random trees, then classifying a case using each tree in this new *forest*, and deciding a final predicted outcome by combining the results across all the trees.

More specifically, from the available data a number k of bootstrapped training samples is considered. On each of these k samples, a prediction tree is constructed considering, at each split, a random sample of $m < p$ predictors. This selection of predictors is performed in order to avoid that the strongest predictors would determine the top split in almost all the estimated trees. On the contrary, sampling at each split m predictors, the strong predictor is considered $(p-m)/p$ times. The result is a *forest* composed by *decorrelated* trees. This is done since averaging not-correlated quantities leads to a variance reduction, otherwise higher averaging high-correlated quantities. In random forests the trees are usually fully grown.

2.2.3 Random Garden

In this work, we introduce an algorithm for ensembling fully grown CART in order to generate a forest. The trees are different from each other by randomly selecting both the individuals (bagging) and the features, like random forest. More in detail, the algorithms is composed by the following steps.

1. The features are divided into two groups, based on their impact on the response variable. Features impact on the response variable is evaluated by using the p-value of a correlation test between features and the response variable. The choice of correlation test depends on variable type:

   - $F-$test for binary features;
   - χ^2-test for categorical features with more than 2 levels;
   - Wilcoxon test for numerical features.

   Features with p-value less or equal to critical value 0.05 are considered highly relevant, otherwise less relevant. The former and the latter group constitute the sample F_R and F_N of features respectively. Clearly $F_R \cup F_N = F$, where F is the complete set of features.
2. Each tree is constructed by

a. Sampling with a replacement of a number of records equal to the total amount of observations, thus considering on average 63.21 % different observations in each sample (see [2]).
b. Sampling *mtry* number of features, set by default equal to the square root of the dimension of F. This sample is constructed in a stratified way from F_R and F_N sets. From F_R are extracted q_R features, where q_R is the proportion of F_R dimension with respect to F dimension. Conversely, from F_N are extracted q_N=*mtry*-q_R features.
c. Growing a fully grown CART (not pruned).

3. Considering the mean of all CART outputs as predicted response variable.

The result is a forest of trees with a lower number of freedom degrees than Random Forest trees. Since the trees are constructed one by one and well-finished, we name this algorithm as Random Garden (RG).

The procedure described above for RG is different from RF mainly due to the random feature selection, since it is performed in a stratified way. An example of RF definition with stratified random feature selection is given by xRF algorithm introduced in [5]. In this study our aim is to stress the differentiation of the trees. As a matter of fact, we do not extract the features at each split as in xRF, but we sample features once for each tree during RG construction.

3 Application

3.1 The Data

The application is performed on 18 datasets collected in a financial institution [1]. Datasets are related to Bank customers and are constructed by merging different sources of data and this has been the first step of the analysis. The response variable is binary (0 or 1, for failure and success respectively). Data sets characteristics are in Table 1.

By looking at Table 1, we can identify three clusters of datasets with respect to Ncust and Pev characteristics. The clusters index is indicated in Table 1. In particular we observe three clusters.

Cluster 1. Population with higher number of customers and lower proportion of success events.
Cluster 2. Population with lower number of customers and lower proportion of success events.
Cluster 3. Population with lower number of customers and higher proportion of success events.

[1] More details about the datasets are not provided due to Legal and Compliance issues.

Table 1 Index number of data set (DS), number of features (Nfeat), number of customers (Ncust), number of events (Nev), proportion of success events in percentage (Pev) and relative cluster (Cluster) for the 18 data sets.

DS	Nfeat	Ncust	Nev	Pev	Cluster
1	311	576546	1983	0.344	1
2	256	158229	463	0.293	2
3	224	171134	412	0.241	2
4	295	94456	1381	1.462	2
5	317	100024	1193	1.193	2
6	264	95229	939	0.986	2
7	689	220598	910	0.413	2
8	376	227578	500	0.220	2
9	423	39620	4760	12.014	3
10	440	35377	706	1.996	2
11	382	12957	353	2.724	2
12	388	24892	2021	8.119	3
13	443	223276	25163	11.270	3
14	257	70257	2376	3.382	2
15	824	423929	12624	2.978	1
16	642	129791	2031	1.565	2
17	932	421887	8903	2.110	1
18	578	11801	515	4.364	2

3.2 Results

Datasets are divided in training and test sets. The training set is used for model estimation and is composed by the 30% of the observations, while the testing set is used for model performance evaluation and is composed by the remaining 70% of the observations. We apply Monte Carlo cross-validation. This procedure creates multiple splits of data into train and test sets and each split is randomly performed from the full dataset. The number of random splits is set equal to 50.

For every combination of dataset, Monte Carlo split and algorithm/feature selection configuration, the model is trained and tested for measuring the performance indicators. Overall, 5400 different models are trained. We choose Area Under ROC Curve (AUC) as performance metric due to its relevance in this application ([3]). As a matter of fact, we are interested in *ranking* the customers according to the model output. We report a summary for AUC values in Table 2.

4 Discussion

The approaches show good performances applied to the different datasets in almost all the trials. As expected, results can be very different among the datasets.

Algorithms performances, measured as AUC score, seems to be related to the cluster where each dataset belongs to. By way of example, in Clusters 1 and 2 which

Table 2 AUC values on average for the three dataset clusters for the different algorithm and FS configurations.

FS	Algorithm	Cluster 1	Cluster 2	Cluster 3
no	CART	0.7136	0.7568	0.8702
	RF	0.7977	0.7833	0.8827
	RG	0.8074	0.7995	0.8359
yes	CART	0.7470	0.7359	0.8589
	RF	0.7935	0.7690	0.8767
	RG	0.8026	0.7923	0.8403

are the ones exhibiting the lowest percentages of positive targets, RG algorithm performs better, RF one performs worse although much better than CART. On the other hand, on Cluster 3 datasets which are the ones with a smaller customer size and more balanced in terms of positive targets proportion, RF is the algorithm performing better, followed by CART and RG. Regarding the impact of the FS procedure, the more valuable improvement is observed only for CART.

Declaration of Interest

The views and opinions expressed in this paper are those of the authors only, and do not necessarily represent the views and opinions of UniCredit Business Integrated Solutions S.C.p.A. or any other organization. All the computations have been conducted on anonymized data on UniCredit servers by UniCredit employee. The results have been observed only in aggregated form.

References

1. Breiman, L.: Random forests. Machine Learning **45(1)**, 5–32 (2001)
2. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer series in statistics, Springer, Berlin (2001)
3. Hanley, J.A. and McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143(1)**, 29–36 (1982)
4. John, G., Kohavi, R., Pfleger, K.: Irrelevant Features and the Subset Selection Problem. Proceedings of the Eleventh International Conference on Machine Learning **129**, 121–129 (1994)
5. Nguyen, T.T., Huang, J.Z., Nguyen, T.T.: Unbiased Feature Selection in Learning Random Forests for High-Dimensional Data. The Scientific World Journal (2015) doi: 10.1155/2015/471371

Spatiotemporal Prevision for Emergency Medical System Events in Milan

Andrea Gilardi, Riccardo Borgoni, Andrea Pagliosa, Rodolfo Bonora

Abstract Organizing the emergency medical system in a big city is an extremely difficult task given the huge number of people that everyday pass through the city area. In this paper we employ a spatio-temporal process to model the emergency event occurrences in Milan. The proposed approach has been found effective in predicting events through the city area and computationally efficient despite the big amount of data to be processed.
Abstract *L'organizzazione di un servizio di emergenza sul territorio risulta essere un compito complesso nelle aree metropolitane come Milano dato l'enorme numero di persone che vi transitano quotidianamente. In questo lavoro si adotta un modello spazio temporale per rappresentare la dinamica delle chiamate di emergenza sul territorio del capologuo Lombardo. Il metodo adottato si é dimostrato essere efficace nel prevedere gli eventi sul territorio comunale nel periodo di tempo considerato e computazionalmente efficiente nonostante la consistente mole di dati da elaborare.*

Key words: emergency medical system, spatio-temporal point process

1 Introduction

Given the large number of people that everyday pass through the metropolitan area, the organization of the emergency medical system (EMS) in Milan is an extremely

Andrea Gilardi
University of Milano-Bicocca, e-mail: a.gilardi5@campus.unimib.it

Riccardo Borgoni
University of Milano-Bicocca e-mail: riccardo.borgoni@unimib.it

Andrea Pagliosa, Rodolfo Bonora
AREU (Azienda Regionale Emergenza Urgenza)

difficult task. Numerous studies have proposed different models for the optimal allocation of ambulances in the territory and each of these models is based on ad-hoc predictions for the future locations of emergency events. In this paper we implemented an algorithm that predicts the distribution of the ambulance interventions in Milan for every hour of the day and every place of the urban area.

All ambulance dispatches from 1st of January 2015 till 25th of September 2017 have been considered for a total of, approximately, 500’000 events. The spatial distribution of the events is reported in (Figure 1-(a)). The figure clearly points out an anomalous pick of event intensity in the north-west part of the map caused by the universal exposition hosted by Milan (EXPO) that took place in this area in 2015. All the events occurred in the EXPO area have been removed from the subsequent analysis.

This type of data is challenging for several reasons:

Sparsity: even if the dataset is extremely large, there are only 21 events per hour on average scattered on an area of about 180 km^2;

Computational challenges: the numerical estimation of a spatio-temporal model is particularly difficult considering the long training time of the algorithms;

Seasonality: the total number of events per hour exhibits both daily and weakly seasonality (Figure 1-(b)).

2 Spatio-Temporal Model

Using the approach suggested in [5], we modeled Milano’s ambulance demand on a continuous spatial domain $S \in \mathbb{R}^2$ and a discretized temporal domain of one-hour intervals $T = \{1, 2, \ldots, \}$. We assumed that ambulance demand follows a Non-homogeneous Poisson Process [1] with intensity function $\lambda_t(\mathbf{s})$ for each time period t. Furthermore, we decomposed this intensity function as

$$\lambda_t(\mathbf{s}) = \delta_t f_t(\mathbf{s}), \quad \mathbf{s} \in S \subseteq \mathbb{R}^2,\ t \in \mathbb{N} \tag{1}$$

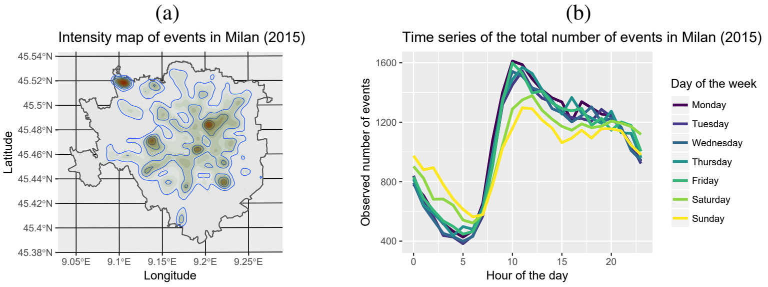


Fig. 1 Spatial distribution of the events in Milan during 2015 (a). Temporal distribution of the number of interventions per hour of the day (b).

where δ_t models the expected number of events during the period t in the region S and $f_t(\mathbf{s})$ represents the continuous spatial density of the ambulance demand at time t. A dynamic latent factor structure has been assumed for the temporal component and it has been estimated iteratively (Section 2.1). The spatial component has been estimated non-parametrically via a weighted kernel (Section 2.2).

2.1 The temporal component

Following the approach suggested in [3, 2], we assumed a dynamic latent factor model for the temporal component δ_t and estimated it using smoothing splines [4]. More specifically we supposed that it is possible to predict the mean value δ_t using a set of deterministic covariates, namely the hour of the day, the day of the week and the week of the year, that were included in the model by applying constraints on the factor loadings as explained below.

To avoid negative values we modelled the intra-day pattern on the log scale and we assumed that it can be approximated using a linear combination of a small number K of factors[1] i.e.

$$\log \delta_{ij} = L_{i1} f_{1j} + \ldots + L_{iK} f_{Kj}, \quad i = 1, \ldots, 365, \quad j = 1, \ldots, 24. \tag{2}$$

The model can be expressed in matrix form as $\log \Delta = LF'$. The matrix L is further partitioned as

$$\log \Delta = LF' = (H_1 B_1 + H_2 B_2) F' \tag{3}$$

where H_1 and H_2 are the incidence matrix that identifies the day of the week and the incidence matrix that identifies the week of the year respectively, whereas B_1 and B_2 are matrices of unknown coefficients with suitable dimensions.

Since both F and L are unknown matrices, we implemented the following algorithm [2] to estimate them:

1. using singular value decomposition we decomposed the matrix of the observed counts logarithms as UDV' and we obtained an initial estimate $L = UD$ and $F' = V'$;
2. considering F as known, we updated the estimate of L using a *gam* model with a Poisson response variable and the day of the week and the week of the year as covariates;
3. considering L as known, we updated the estimate of F using again a *gam* model with a Poisson response and the hour of the day as covariate;
4. we iterated 2 and 3 until convergence.

[1] The value of K was chosen by training the model on the data collected in 2015 data validating it using 2016 data.

2.2 The spatial component

The continuous spatial density function $f(\cdot)$ has been estimated at a future time period u using the spatio-temporal weighted kernel density estimator [5]

$$f_u(\mathbf{s}) = \frac{\sum_{t \in \mathscr{T}_{obs}} \omega(\mathbf{s}_{i,t}, u) K_{\mathbf{H}}(\mathbf{s} - \mathbf{s}_{t,i})}{\sum_{t \in \mathscr{T}_{obs}} \omega(\mathbf{s}_{i,t}, u)}, \qquad \mathbf{s} \in S \tag{4}$$

where $\mathscr{T}_{obs}$ represents a set of past periods, $\omega(\cdot)$ a weight function and $K_{\mathbf{H}}(\cdot)$ a multivariate gaussian Kernel with bandwidth matrix $\mathbf{H}$.

Figure 1-(b) shows that the observed number of events exhibits both daily and weakly seasonality. The aim of the weight function is to take advantage of these patterns and select those observations that are mostly influential in predicting the density function at a generic future time point u. Analysing the data it emerged that the strength of these patterns was different in different areas of Milan. For this reason we divided the municipality in $C = 9$ neighbourhoods and the weight function was estimated separately in each cell.

We assumed [5] that the prediction at a future time point u depends only upon the temporal lag between u and the event at time t. The impact of the event occurred at time t is weighted by the following weight function that measures how two observations located in a cell c are positively correlated:

$$\omega_c(u-t) = \rho_{1,c}^{u-t} + \rho_{2,c}^{u-t} \rho_{3,c}^{\sin^2\left(\frac{\pi(u-t)}{24}\right)} \rho_{4,c}^{\sin^2\left(\frac{\pi(u-t)}{168}\right)}, \quad c \in \{1, \ldots, C\} \tag{5}$$

where ρ_1 describes any short-term seasonality, ρ_3 and ρ_4 express daily and weakly seasonality respectively and ρ_2 is a discount factor. Since using a likelihood approach to estimate these coefficients is prohibitive due to the computational costs, we implemented the algorithm suggested in [5]. Hence $\rho_1, \ldots, \rho_4$ have been estimated by minimizing the following quantity:

$$\min_{\rho_{j,c}, j \in \{0, \ldots, 4,\}} \sum_{l=1}^{M} (A_c^+(l) - \rho_{0,c}\omega_c(l))^2 \tag{6}$$

where A_c^+ represents the positive part of the autocorrelation function of the proportion of events in the spatial cell c, M is the maximum lag considered in the autocorrelation function and $\rho_{0,c}$ is a normalizing constant. The minimization problem was worked out using the *optim* function of R [6].

3 Results and conclusions

We trained the algorithm described in section 2.1 using the data from 2015. The number of factors K was identified via an external cross validation based on the dataset collected in 2016. $K = 4$ is the optimal value suggested by the procedure. Finally the model has been tested using again the 2016 data. The results are displayed in Figure 2 where the scatter plot of the predicted counts versus the observed counts is reported. A boxplot of predictions is drawn at each observed frequency to improve the graph readability. Patterns somehow similar were found for the two years. Despite a great variability and some potential outlying values, the predictions seem to replicate reasonably well the observed data both for the train set (year 2015) and for the test set (year 2016).

The spatial dynamic has been estimated using the procedure detailed in Section 2.2. First we solved the minimization problem mentioned above to obtain an estimate of the coefficients for the weight function. Then we estimated the spatial density using Equation 4 for one particular day. To exemplify the procedure we considered two maps for the 26th of September 2017, one estimated at 02:00 AM and the other estimated at 2:00 PM. The results are displayed in Figure 3 where darker areas are associated to higher values of the spatial density. There are clear differences between the two plots: during daytime hours the events are mostly concentrated around popular places and crossing points such as the Duomo area or the Central Station areas whereas during the night interventions are more scattered over the whole territory of the municipality.

This study demonstrates how resorting to a spatio-temporal non homogeneous Poisson model is adequate to represent the temporal and the spatial patterns that are present in the EMS data of Milan. Some areas and some hours of the day are found more critical for emergency events and this is a preliminary step to support local authorities in optimally allocating ambulances and resources in the territory. From the statistical modelling perspective some further enhancements can be introduced to improve the diagnostic, for instance developing algorithms to simulate events from the estimated model and to create measures of the prediction performance. Moreover, alternative specifications for the weight function should be also considered. Finally, a comparison between the temporal and spatial model suggested in this paper and other approaches such as machine learning algorithms (for instance, boosting or random forest) can be also useful to identify the best modelling strategy to support authorities in the day by day activity.

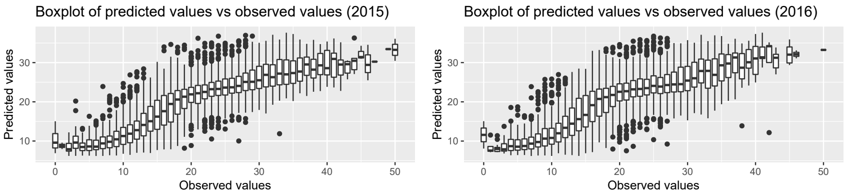


Fig. 2 In these two plots we display the predictive performance of the dynamic latent factor model for the training data (2015, left plot) and for the test data (2016, right plot). We can see that, despite high variance in the estimates, both plots show good predictive performances.

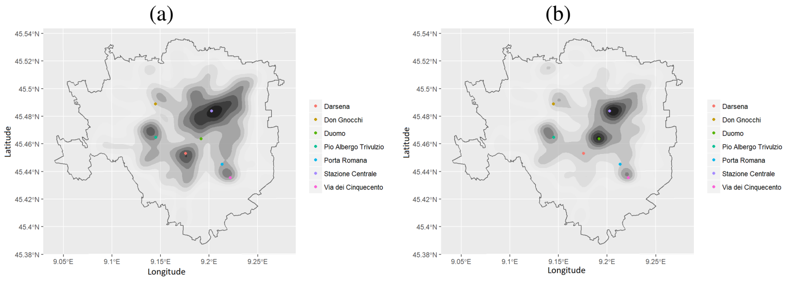


Fig. 3 Spatial density estimate for the 26th of September 2017 at 02:00 AM (a) and 02:00 PM (b). Some important places of the city are highlighted on both maps. Darker areas are associated to higher values of the spatial density.

References

1. Diggle, P.J.: Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Chapman and Hall/CRC, Boca Raton, Florida (2013)
2. Matteson, D.S., McLean, M.W., Woodard, D.B. and Henderson, S.G.: Forecasting emergency medical service call arrival rates. Ann. Applied Statistics 5, 1379-1406, (2011)
3. Shen, H. and Huang, J.: Forecasting Time Series of Inhomogeneous Poisson Processes with Application to Call Center Workforce Managament. Ann. Applied Statistics 2, 601-623, (2008)
4. Wood, S.N.: Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC, Boca Raton, Florida (2017)
5. Zhou, Z. and Matteson, D.S.: Predicting ambulance demand: A spatio-temporal kernel approach. ACM SIGKDD. ACM 2297-2303, (2015).
6. R Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria, (2017).

SPATIAL SEGREGATION OF IMMIGRANT HOUSEHOLDS IN MESSINA

Segregazione spaziale delle famiglie di immigrati nella città di Messina

Angelo Mazza and Massimo Mucciardi

Abstract We investigate spatial segregation of the most representative groups of foreign immigrants in the city of Messina.
We use spatial modeling with point level data to assess attraction among migrants' households of the same ethnicity while adjusting for spatial inhomogeneity in the cost of housing, job availability and other factors. This work is still in progress; first results show more dispersed settlements patterns for ethnicities more involved with household services than for those devoted to commerce and peddling.
Abstract *In questo lavoro vengono analizzati i modelli insediativi dei gruppi di immigrati più rappresentativi nella città di Messina. Attraverso l'impiego di modelli spaziali per dati puntuali misuriamo l'attrazione tra nuclei familiari della stessa etnia, al netto delle eterogeneità spaziali nel costo degli affiti e nella disponibilità di lavoro. Il lavoro è ancora in progress; i primi risultati mostrano modelli insediativi meno concentrati per quei gruppi etnici maggiormente coinvolti nei servizi alle famiglie rispetto a quelli maggiormente impegnati nel commercio e nella vendita ambulante.*

Keywords: Spatial Clustering, Voluntary Segregation, Migrant Households

Introduction

Ethnic residential segregation has long been investigated in the USA and South Africa; however, only recently this topic has become prominent in Europe as well, following from a debate that links segregation with the perceived failure of integrating immigrants into host societies (Kalandides and Vaiou, 2012, Andersson, 2013). A minority ethnic group is segregated when the arrangement of minority households departures from expectations based upon a random spatial allocation. Following Schelling (1971), we distinguish two sources of spatial segregation. The first is mostly economically induced, and it is mainly ascribed to within-city inhomogeneity in the price of residential property, and in the availability of jobs. The second is attraction, i.e., individuals prefer living in areas where their group is a majority or near-majority (Clark and Fossett, 2008). Moreover, newly arrived minority migrants may benefit from positive spillovers in settling close to their compatriots, and transnational social networks play an important role in channeling arriving migrants into specific neighborhoods.

Whereas economic induced segregation might explain some initial degree of segregation and raises questions of social equity, the Schelling (1971) model highlights the importance of individually motivated segregation and posits that even mild preferences for living with similar neighbors carry the potential to be strong determinants for residential segregation (Clark and Fossett 2008).

From a methodological point of view, traditional indices of spatial segregation rely on data aggregated by areal unit, typically census tracts (Mazza and Punzo, 2015; Mucciardi, Mazza and Altavilla, 2017); however these analyzes could suffer from the modifiable area unit problem (i.e., alternative zoning or a different scale may yield different results; see Openshaw, 1984 and Wong, 2009) and do not allow to distinguish between the two sources of segregation. Instead, in this paper, we use individual point data, and we apply the inhomogeneous K-function (Baddeley et al., 2000), which allows assessing spatial attraction among migrants while adjusting for spatial inhomogeneity (Mazza and Punzo, 2016). To avoid the risk of confounding the two sources of clustering, spatial inhomogeneity is estimated following a case-control approach. Cases will be, in turn, migrants of each nationality, while the controls will be a random sample of native household locations.

The model: the inhomogeneous K-function

The spatial distribution of household locations may be represented by a point pattern. The simplest theoretical model for a spatial point pattern is the homogeneous Poisson process (HPP), in which the expected number of events occurring within a unitary region $u \in R$ follows a Poisson distribution, whose intensity $\lambda(u)$ is uniformly distributed over R (Diggle, 2003). The inhomogeneous point process (IPP) is a generalization of the HPP obtained replacing the constant intensity λ by a spatially varying intensity function $\lambda(u)$; clustered patterns occur, with regions where $\lambda(u)$ is higher receiving a higher number of events. Ripley's *K*-function, usually denoted with $K(d)$, is used to detect clustering (or inhibition) in point processes with constant intensity; at every spatial distance d, $\lambda\ K(d)$ is the expected number of additional points of the process X located in a circle b of radius d surrounding an arbitrary event x. Baddeley et al. (2000) generalized Ripley's K-function to non-homogeneous point processes, by weigthing each point x_j by $w_i = \frac{1}{\lambda(x_i)}$. The inhomogeneous K-function is defined as

$$K_{\text{inhom}}(d) = \mathrm{E}\left[\sum_{x_j \in X} \frac{1}{\lambda(x_j)} 1\{0 < \|u - x_j\| \leq d| \quad u \in X\}\right] \qquad (1)$$

where $\|u - x_j\|$ is the Euclidean distance between points u and x_j and $1\{\dots\}$ is the indicator function. Clustered point patterns arise from the joint action of spatial inhomogeneity and spatial attraction. To distinguish between these two sources is challenging task, since many spatial processes are "equifinal", i.e. one realization of a point process may be consistent with underlying processes involving clustering due to either spatial inhomogeneity or spatial attraction (Harvey, 1966). Mazza and Punzo (2016) deal with this issue using a case-control approach. In case-control studies, data consist of a realization of two spatial point processes, one representing cases of a condition of interest and the other representing controls drawn at random from the population at risk. Assuming that the size of the population at risk tends to infinity and the sampling fraction to zero, the controls constitute a realization of an IPP with intensity $\lambda^*(u)$; the cases form a second, independent, point process – which may or may not be a Poisson process – with intensity $\lambda(u)$. The question of interest is whether the cases form an IPP with intensity proportional to that of the controls, i.e., $(u) = \rho\lambda^*(u)$, or whether they exhibit additional spatial structure.

Diggle et al. (2007) propose that the ratio between the intensity functions may be modeled to depend on a vector of m spatially referenced covariates

$$\mathbf{z}(u) = \left(z_1(u), \dots, z_m(u)\right)', \text{i.e. }; \lambda(u) = \lambda^*(u) f(\mathbf{z}(u); \boldsymbol{\theta}) \qquad (2)$$

where $f(\cdot)$ is any nonnegative function, either nonlinear parametric with parameters θ as in (2), or nonparametrically specified.

Our controls are a random sample of native household locations, and economic constraints are taken into account through explanatory spatial covariates related to the cost of residential property and job availability.

Data at hand come from the administrative register of the city of Messina, at December 31, 2016. Register data have been integrated within a geographical information system, and all the residential addresses geocoded. Other data used are the rent cost per square meter for private residential properties, as collected by the Italian Revenue Agency (OMI database). Computations are carried out by means of the spatstat package (Baddeley and Turner, 2005) for the R computing environment (R Core Team, 2017).

Main results

Our investigation will cover the most representative groups of foreign immigrants in the city of Messina. We expect a differentiated spatial trend for the different nationalities, with higher values in the more central parts of the city for those groups involved mainly in retailing and peddling activities. We will compute the estimates for the inhomogeneous K-function at various distances d. We will also report confidence envelopes based on Monte Carlo simulations for the null hypothesis of absence of interaction. First results show more dispersed settlements patterns for ethnicities more involved with household services than for those devoted to commerce and peddling. Furthermore, a higher spatial attraction emerges for those groups for whom the effects of chain migrations and family reunifications have been stronger.

References

1. Andersson, R.: Reproducing and reshaping ethnic residential segregation in Stockholm: the role of selective migration moves. Geografiska Annaler: Series B, Human Geography 95(2): 163–187. (2013).
2. Baddeley, A., Turner, R. Practical maximum pseudolikelihood for spatial point patterns. Australian and New Zealand Journal of Statistics 42: 283–322. (2000).
3. Baddeley, A., Turner, R.: spatstat: An R Package for Analyzing Spatial Point Patterns. Journal of Statistical Software 12(6): 1–42. (2005).
4. Clark, W.A., Fossett, M.: Understanding the social context of the Schelling segregation model. Proceedings of the National Academy of Sciences 105(11): 4109–4114. (2008).
5. Diggle, P.J., Gómez-Rubio V., Brown P.E., Chetwynd A.G., Gooding, S.: Second-Order Analysis of Inhomogeneous Spatial Point Processes Using Case–Control Data. Biometrics 63(2): 550–557. (2007).
6. Harvey, D.: Geographic Processes and the Analysis of Point Patterns: Testing Models of Diffusion by Quadrat Sampling. Transactions of the Institute of British Geographers XL, p.81–95. (1966).
7. Kalandides, A., Vaiou. D,: 'Ethnic' neighborhoods? Practices of belonging and claims to the city. European Urban and Regional Studies 19(3): 254–266. (2012).
8. Mazza, A., Punzo, A.: On the upward bias of the dissimilarity index and its corrections. Sociological Methods & Research 44(1): 80–107. (2015).
9. Mazza A, Punzo. A,: Spatial attraction in migrants' settlement patterns in the city of Catania. Demographic Research, vol. 35, p. 117-138. (2016).
10. Mucciardi, M. Mazza. A. Altavilla, A.M.: Analysis of the residential pattern of foreign immigrants in Catania using GWR modelling, SIEDS, vol LXXI n.2, (2017).
11. Openshaw, S.: The modifiable areal unit problem. Geobooks, University of East Anglia, Norwich, England. (1984).
12. R Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. (2017).
13. Schelling, T.C.: Dynamic models of segregation. Journal of Mathematical Sociology 1(2): 143–186. (1971).
14. Wong, D.: The modifiable areal unit problem (MAUP). In: Fotheringham, A. and Rogerson, P. (eds.). The SAGE handbook of spatial analysis. London: SAGE Publications: 105–123. (2009).

Supervised Learning for Link Prediction in Social Networks

Link prediction nelle reti sociali attraverso l'utilizzo di metodi e modelli di apprendimento supervisionato

Riccardo Giubilei, Pierpaolo Brutti

Abstract Link prediction is an estimation problem that has drawn a great deal of attention in recent years. In this work, a supervised learning approach is adopted to perform link prediction on data retrieved from Facebook. The specific goal, then, is to estimate the probability of two users to become friends in order to recommend them to one another whenever this probability turns out to be sufficiently high. On social platforms like Facebook, friendship recommendation is clearly a crucial ingredient since, when properly implemented, it plays a key role in determining the network growth. The contribution of this work consists in performing friendship recommendation on Facebook using a supervised learning approach that takes explicitly into account vertices' attributes; that is, all the personal information that users make available on their profiles.

Abstract La link prediction ha attirato molta attenzione negli ultimi anni. In questo lavoro, un approccio di apprendimento supervisionato viene utilizzato per fare link prediction su dati provenienti da Facebook. L'obiettivo è quindi quello di stimare la probabilità che due utenti diventino amici, in modo da suggerire gli uni agli altri quanto tale probabilità è alta. La raccomandazione delle amicizie è un problema molto importante poiché il suo corretto funzionamento è fondamentale per la crescita delle reti, che è l'obiettivo primario di siti come Facebook. Il contributo di questo lavoro è quello di fare ciò utilizzando un approccio di apprendimento supervisionato che prenda esplicitamente in considerazione anche gli attributi dei vertici, ovvero le informazioni personali che gli utenti inseriscono nel proprio profilo.

Key words: Link Prediction, Social Network Analysis, Network Science, Graph Theory, Supervised Learning, Machine Learning, Binary Classification.

Riccardo Giubilei
Sapienza University of Rome, e-mail: riccardo.giubilei@uniroma1.it

Pierpaolo Brutti
Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction and background

A social network is a popular way to model the interaction among people belonging to a group or a community. It can be represented using a graph, where each node is a person and each link indicates some form of association between two people.

In this framework, performing link prediction consists in predicting which nodes are likely to get connected. More precisely, the goal is to predict the likelihood of a future association between two unconnected nodes. This is carried out supposing the likelihood of link formation depends on the similarity between the two nodes.

In 2004, Liben-Nowell and Kleinberg [3] proposed one of the earliest link prediction models that works explicitly on social networks. The learning paradigm in this setup typically extracts the similarity between a pair of vertices exploiting various graph-based similarity metrics and uses the ranking on the similarity scores to predict the link between two vertices.

Subsequently, Hasan et al. [2] extended this work in two ways. First, they showed that using external data outside the scope of graph topology (namely, the vertices' attributes) can significantly improve the prediction results. Second, they used several similarity metrics as features in a supervised learning setup where the link prediction problem is posed as a binary classification task. Since then, the supervised classification approach has been popular in various other works on link prediction.

2 Motivation

The popularity of online social network services has thrived in recent years, attracting an increasingly large number of users. By connecting users with similar professional backgrounds or common interests, they open up new channels for information sharing and social networking. Creating connections not only helps to improve user experience, but also increases the chance of producing larger and more well-connected networks, which is the primary goal of these sites.

Consequently, link formation is fundamental. In Facebook, a link is formed whenever two people become friends. To increase the probability of link formation, users with the highest probability of becoming friends may be suggested to one another. This is achieved through the friendship recommendation system, that aims to find the most similar users in terms of their profile contents or their behavior so as to offer them to each other.

The scope of this analysis is to apply link prediction methods and techniques to perform friendship recommendation on Facebook data. The problem is tackled using a supervised learning approach that blends together both topological features and users' personal information. Incorporating the latter as covariates is everything but trivial so, in the following, we introduce a relatively simple method to handle effectively this crucial modelling step.

3 Data

The data was collected in 2014 by Julian McAuley (UC San Diego) and Jure Leskovec (Stanford University) using a Facebook application that asked to a pool of volunteers the permission to download their Facebook's profile information via Facebook API. In order to ensure the volunteers' privacy, all the data have been completely anonymized by assigning users and features sequential IDs. The data collection proceeded in the form of ego networks, i.e. starting from a central node (the person who gave the permission), and then expanding the network considering his friends and the mutual friends between them and the central node. 110 ego networks were collected, making a total of $27,520$ Facebook users. For each of these users, public information contained in their profile was also recorded.

Therefore, the dataset is composed by 110 distinct files, which correspond to the 110 ego networks, and by the additional file that contains the users' attributes. In this work we focus our attention on two specific ego networks. The first one is associated to user $6,934$ and has been selected as the train dataset being the largest among those that do not contain links to users from other ego networks. It is formed by 773 nodes, including the central one, and by $26,023$ links between them. Since the number of nodes is 773, the number of potential links in the network, given by all the possible combinations between the nodes, is $268,278$. Therefore, the number of actual links is approximately the 9.70% of all the possible links. The second one is the ego network of user $3,236$, and has been chosen as the test dataset for being structurally different from the first one. Indeed, it is composed of 345 nodes and $4,013$ links among them, which correspond to the 6.76% of the $59,340$ possible links.

4 Experimental setup

The link prediction problem is formalized as a supervised classification task, where each instance corresponds to a pair of vertices in the social network graph. Instances are characterized by features describing the similarity between the two nodes and a label denoting their link status. In particular, the instance is classified as positive if there exists a link between the nodes, or negative otherwise. The output of the models is a score for each non-observed link which quantifies how likely it is that it will actually become a link. The instances classified as positive are those that exceeds a certain threshold score.

Since each instance corresponds to a pair of vertices, the features should necessarily represent some form of proximity between them. In existing research works on link prediction, the vast majority of the features are related to the graph topology. Typically, they are built by computing similarities based on the node neighborhoods or on the set of paths that connect those two nodes. However, as anticipated in the Introduction, Hasan et al. [2] have proposed to extend the set of features in order to include also the vertices' attributes.

Now, coming back to our specific application, since we are dealing with ego networks, their topology immediately implies a diameter equal to 2. As a consequence, any feature based on paths is definitively not very informative and will not be included in the analysis. On the other hand, five neighborhood-based similarity indices that single out different aspects of the link formation phenomenon are selected. More specifically we consider: *Common Neighbors*, *Jaccard Index*, *Preferential Attachment Index*, *Adamic-Adar Index* and *Resource Allocation Index* [5]. For what concerns other *local* indices available in the literature, it is enough to say that they will not be considered here mainly because they have already shown to not lead to significant improvements in similar analyses.

Attributes-based features are built considering the file containing the users' attributes related to their personal Facebook profile. However, this file contains some redundant information, and, in addition, many of the attributes collected are not available for the majority of the users. Redundant attributes, such as the first, the middle, the full name and the ID, are excluded from the analysis. Likewise, all attributes that have been recorded for less than $1,000$ users are not considered. Among the remaining ones, some additional feature selection is carried out, eliminating variables with little to no informativeness. In order to build similarity indices from the remaining attributes, it is important to underline that a user may have more than a value for the same attribute. Consequently, the idea is to count, for each pair of nodes and for each attribute, the number of values they have in common for that attribute. This is motivated by the belief that the larger the number of characteristics two unconnected users have in common, the higher the probability that they will be linked in the future. This procedure leads then to a data-matrix, with the rows corresponding to the pair of vertices, and the columns being the attributes. The generic entry for this matrix is the number of times the values of a certain attribute coincide for the pair of nodes considered.

5 Models and results

Five binary classification models are considered: *Random Forest*, *Neural Network*, *Gradient Boosting*, *Naive Bayes* and *Logistic Regression*. For each model, a careful parameter tuning is carried out. In order to evaluate the predictive abilities of these models, a 10-fold cross validation is performed on the train data. The models are then evaluated using a number of metrics, including *accuracy*, *specificity*, *recall*, *precision*, *F1 score*, *Area Under the Receiver Operating Characteristic curve* (AUROC) and *Area Under the Precision-Recall Curve* (AUPRC).

Table 1 shows the performance comparison for the different classifiers considered. For the fixed-threshold metrics, the threshold has been set to 0.5. The results are very good for almost every metric. However, the choice of the best model is performed by considering only the metrics that are independent of the threshold chosen to convert the probability scores to class labels, i.e. AUROC and AUPRC.

Consequently, the best model is the Gradient Boosting, which is then used to make prediction on the test data.

Model	Accuracy	Specificity	Recall	Precision	F1 score	AUROC	AUPRC
Random Forest	89.97%	89.48%	94.57%	49.11%	64.65%	97.24%	79.43%
Neural Network	94.37%	97.53%	64.88%	73.87%	69.09%	97.11%	78.74%
Gradient Boosting	94.49%	97.64%	65.15%	74.81%	69.45%	97.24%	79.53%
Naive Bayes	92.31%	93.29%	83.26%	57.15%	67.77%	96.44%	74.86%
Logistic Regression	91.07%	91.01%	91.69%	52.27%	66.58%	96.99%	76.72%

Table 1: Evaluation metrics for the models considered using a 10-fold cross validation.

6 Prediction

The results obtained using the Gradient Boosting model on the test data are reported in Table 2. In addition to the threshold-independent metrics AUROC and AUPRC, also recall and precision are included, being of interesting and useful interpretation in the specific context. In fact, a recall of 82.63% indicates by definition that a little more than 8 people out of 10 a user may want to add are indeed suggested. On the other hand, a precision equal to 70.17% means that users would add approximately 7 people out of 10 suggested.

Model	Recall	Precision	AUROC	AUPRC
Gradient Boosting	82.63%	70.17%	98.06%	84.66%

Table 2: Evaluation metrics for the prediction on the test data.

Figure 1 allows to visualize the predictive results obtained using the model. In particular, all the effectively existing links are reported in the figure, coloring them

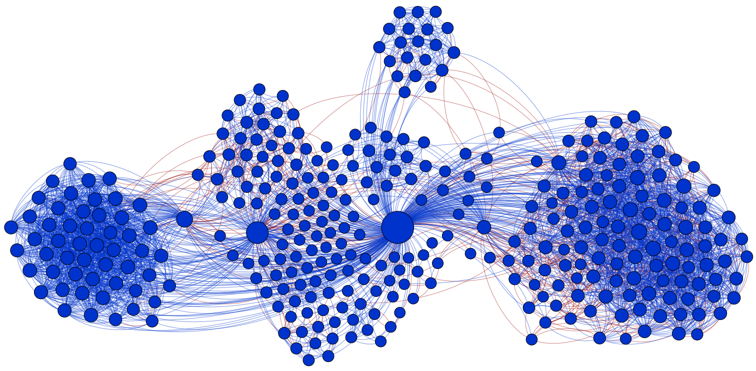


Fig. 1: Correctly predicted links (blue) and missed links (red) in the test ego network.

differently based on the predictions: correctly predicted link are colored in blue and missed links are colored in red. Therefore, it is possible to note that the model manages to reconstruct a massive part of the graph. This is achieved at the cost of suggesting only 3 people that a user would not be interested in adding every 10.

7 Conclusions and future work

The results obtained show that the link prediction task has been accomplished in a satisfying way. In particular, the inclusion of attribute-based features seems to be extremely useful, allowing to make very good predictions. This aspect is of great importance for both confirming that including them actually improves the predictive abilities of the models, and for validating the procedure used to build them starting from the users' personal information. In addition, the supervised learning approach has proved to be effective also for performing link prediction on Facebook data, with particular reference to ego networks.

In the future, it would be interesting to consider larger graphs, and study specific methods and techniques that scale well on "big data" networks.

In addition, the possibility to retrieve and consider the time-stamps of the links is certainly an aspect which would help in link prediction. For example, they may be included by treating more recent links as more important than older ones. Important contributions on this extension, also defined time-aware link prediction, are those by Ahmed et al. [1] and by Tylenda et al. [4].

The techniques and models presented here may be exploited to perform link prediction outside the specific task of recommending friends on Facebook. For example, the case of directed networks may be considered, being of great importance in other online social networks like Twitter. In the end, link prediction is a very relevant problem in almost every kind of networks, and it would be interesting to consider applications also in these other domains.

References

1. Ahmed, A., Xing, E. P.: Recovering time-varying network of dependencies in Social and biological studies. In: Proceedings of the National Academy of Sciences of the United States of America, 106(29): 11878–11883 (2009)
2. Hasan, M. A., Chaoji, V., Salem, S., Zaki, M.: Link Prediction using Supervised Learning. In: Proceedings of SDM Workshop of Link Analysis, Counterterrorism and Security (2006)
3. Liben-Nowell, D., Kleinberg, J.: The Link Prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology, 58(7): 1019–1031 (2004)
4. Tylenda, T., Angelova, R., and Bahadur, S.: Towards Time-aware Link Prediction in Evolving Social Network. In: SNA-KDD '09: Proceedings of the third Workshop on Social Network Mining and Analysis (2009)
5. Wang, P., Xu, B., Wu, Y., Zhou, X.: Link Prediction in Social Networks: the State-of-the-Art. Science China Information Sciences, 58(1): 1–38 (2005)

Women's empowerment and child mortality: the case of Bangladesh

Il ruolo della donna nella lotta alla mortalità infantile: il caso del Bangladesh

Chiara Puglisi[1] and Annalisa Busetta[2]

Abstract Bangladesh is the Southern Asian country that has been experiencing the highest absolute decline in the Under Five Mortality Rate in the past 15 years. This paper focuses on the importance of women's education and empowerment variables in explaining this extraordinary result. We use a two-level multilevel logistic regression to take into account the great differences among territorial communities in terms of child mortality reduction. It emerges that the importance of woman's empowerment - measured as individual and as mother - remains relevant even when the context is considered. A sensitivity analysis has been conducted to test the relevance of different indicators of female empowerment.

Introduction

The importance of female education for child survival in developing countries has been widely acknowledged [3;12]. Mosley and Chen [13] have been among the first scholars highlighting the importance of the mother's role, by arguing that women's education has positive effects on children's chances of survival. They argued that education allows mothers to increase their capabilities concerning nutrition, hygiene and treatment of illnesses, which in turn allows

[1] Chiara Puglisi, University of Bologna; email: chiara.puglisi93@hotmail.it
[2] Annalisa Busetta, Dept. Of Economics, Business and Statistics (SEAS) University of Palermo; email: annalisa.busetta@unipa.it

them to properly take care of their children. Women's education appears to affect child mortality both directly and indirectly, i.e. through its influences on gender equality and female empowerment [1].

In particular, research also showed that female empowerment is positively associated to child survival [7;8;10]. Authors who have been exploring such relationship, though, have been conferring to the term "empowerment" a wide meaning, for instance as a general holding of autonomy [7], by measuring it through several proxies: as decision-making power in the household, freedom of movement and employment status [8], as control of resources [10] and so on, hence implying for the relationship between female empowerment and child mortality to always retain its positivity, independently of how empowerment is measured. Kabeer [9], on the other hand, argued that "what mattered for achievements in relation to children's wellbeing was women's agency as mothers rather than as wives" [9:450]. She suggested to focus on indicators that are strictly related to women's ability to take action in safeguarding their children's healthcare, rather than on the ones concerning, for instance, equality within the marital relationship.

In this paper we aim to evaluate the relative importance of female empowerment and whether its importance changes if measured as possibility of agency towards the children's healthcare compared to female empowerment measured in terms of equality in the relationship between the parents of the children or in terms of women's agency in the social context.

The case study: Bangladesh.

Bangladesh makes an especially interesting case study because of its impressive improvements in terms of child survival: it is the country that, among all Southern Asian states, has been experiencing the highest absolute decline in the Under Five Mortality Rate in the past 15 years [6].

Comparing several rounds of Bangladesh Demographic and Health Surveys (BDHS) it emerges that as female schooling increase child mortality decreases, and this relation is steady over time. In Bangladesh DHS 2004 women who had no education have an Under Five Mortality rate (U5MR) equals to 114‰ live births whereas those with higher education have an U5MR of 58‰ live births. In BDHS 2014 they decrease respectively to 63‰ live births and 24‰ [6]. Moreover, a great deal of attention has been paid to Bangladesh in the literature concerning the role of female education and empowerment in child survival: empowering women has been pointed out as one of the most important factors behind the so-called "Bangladesh paradox", consisting in the sharp contrast between excellent health performances and reduction of mortality, and the malnutrition and low use of basic health services that still characterize the country [5].

Data and methods

The data utilized for this study derive from the Bangladesh DHS 2014. It is the 7th survey implemented by the DHS Program in the country, and it involved men and women in reproductive age, who provided also information about their children. The survey has been conducted on all the ever-married women, aged 12 to 49 years, selected through a two stage stratified sampling procedure: the first stage involved the selection of 600 territorial units (clusters) - 207 urban clusters and 393 rural ones; in the second stage, for every cluster has been selected an average of 30 households. Our study has concerned all the children born in the 10 years preceding the survey and their mothers. The response variable of the analysis was under-five death - that is, the death of a child before reaching his or her fifth birthday - recorded as a dichotomous variable[4]. The data for estimating child mortality derive from the birth history of the Women’s Questionnaire, where were collected information about all their live born, such as sex, month and year of birth, survival status, age at the moment of the interview or age at death in case of deceased child [10]. The subsample utilized in our study concerned a total of 16,439 children born in the 10 years preceding the survey and 10,856 women, that is, the mothers of such children.

As already mentioned, the main focus of this work is the effect of different indicators of female empowerment. In particular, we compare the results of:

- A measure concerning female empowerment as education, categorized in four modalities: no education, primary education, secondary and higher education. Note that this variable only concerns women who had completed their education before getting married;
- Three indicators measuring women’s empowerment in terms of equality within the couple: an indicator concerning the woman’s freedom of movement; a variable measuring the number of decisions to which women participate in the household; a variable quantifying the number of occasions in which women tend to justify violence from the husbands;
- An indicator concerning the woman’s possibility to act towards her children’s wellbeing: a variable investigating who has the final say final say on children’s health care.

All the models estimated in the next section control for individual characteristics of the child (sex, birth order, birth interval), of the mother (mother’s cohort and age at birth), of the household (wealth index (WI)[4] and availability of electricity)

[4] We excluded those born less than 59 months before the end of the survey, as they were not exposed to five full years of risk of dying.

[5] WI is a composite measure of a household's cumulative living standard calculated by DHS. It relies on household’s ownership of selected assets (such as televisions and bicycles), materials used for housing construction and types of water access and sanitation facilities.

and of the context (percentage of uneducated men, uneducated women, rich and richest households, households with electricity in the cluster, and type of residence).

We employ a two-level multilevel logistic regression approach, based on the calculation of adjusted odds ratios of under-five mortality according to the selected explanatory variables. The choice of a multilevel approach is consistent with the very clustered nature of the DHS data; the use of a multilevel approach allows for the correction of the bias that may occur in the estimation of the parameters resulting from such clustered structure. This option has been selected to give relevance to the influence that territorial context may have on children's survival. As Matteson et al. argued, even though there may be optimal individual-level conditions, "*the absence of adequate health care in the community, living in a congested urban environment or living in the context of high poverty may have independent negative consequences for her [a mother's] infant's well-being and survival*" [11:1884].

Preliminary results

To analyse how the five different measures of women empowerment influence under-five mortality in Bangladesh we fitted 10 multilevel random slope models (see Table 1) with individual and contextual variables: models 1, 3, 5, 7 and 9 control only for individual determinants, whereas models 2, 4, 6, 8 and 10 add the control of contextual determinants to highlight the significance of the effect of community.

Model 1 and 2 provide odds ratios of under-five death according to the educational level of the mother without and with the contextual-level variables. It emerges that mother's education has a strong protective effect in terms of child survival: the odds of under-five mortality among children born to women with a secondary or higher education are 29% lower than the probability of those born to uneducated women. Such effect is retained also including contextual factors (model 2): having a "higher" educational level retains its protective effect in terms of under-five survival, and this result is still statistically significant even when controlling for the percentage of uneducated males and females in the cluster, which means that the low educational level of the mother has a significant impact on child mortality independently from the average educational level of men and women within the cluster. As the literature on the matter showed, educated mothers are more likely to bring their children to healthcare facilities [4], are more conscious and more enlightened (which helps them to make better decisions for themselves as well as for the family) and can better communicate and interact with health service providers to get required care for themselves as well as for their newborn [2].

Models 3 to 8 display the odds ratios of under-five mortality in correspondence of a) mothers that have freedom of movement; b) mothers who participate to decisions within the household and c) mothers who justify wife-beating. We can see that mothers' possibility to move freely (that is, alone or with her children), her decisional power inside the couple and her attitudes towards wife-beating do not affect the odds of child mortality. Such results remain unchanged even when controlling for contextual-level variables.

Table 1 – Results of multilevel random slope models (odds ratio and significance level) with individual and contextual covariates

Variables	*Model 1*	*Model 2*	*Model 3*	*Model 4*	*Model 5*	*Model 6*	*Model 7*	*Model 8*	*Model 9*	*Model 10*
Mother's cohort ***[ref.1964-68]***										
1969-1973	.57*	.57*	.57*	.57*	.52**	.52**	.56*	.56*	.54**	.54**
1974-1978	.56*	.56*	.58*	.59*	.56**	.54**	.58*	.58*	.57*	.57*
1979-1983	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
1984-1988	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
1989-1993	.39**	.38**	.43*	.42*	.44*	.44*	.41*	.41*	.46	.46
1994-1999	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Mother's age at child birth ***[ref. 15-19 ys]***										
20-24 ys	.76*	n.s.	.73*	.74*	n.s.	n.s.	.73*	.73*	.79	n.s.
25-29 ys	.46***	.46***	.44***	.44***	.46***	.47***	.41***	.42***	.48**	.49**
>29 ys	n.s.	n.s.	n.s.	n.s.	.68**	n.s.	n.s.	n.s.	n.s.	n.s.
Sex										
Female	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Birth order ***[ref. Firstborn]***										
2°	.65***	.65***	.68**	.67***	.68**	.67**	.67***	.67***	.66***	.66***
3°	.69*	.68*	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
4°	n.s.	n.s.	n.s.	n.s.	1.42*	n.s.	1.42*	n.s.	n.s.	n.s.
Birth interval ***[ref. <24 months]***										
>24 months	.61***	.61***	.62***	.62***	.58***	.59***	.60***	.61***	.59***	.59***
Wealth index [ref. Poorest/poor]										
Medium/ Richest HH	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Electricity in the HH ***[ref. No electricity]***										
Yes	.68***	.77*	.68***	n.s.	.70**	n.s.	.68***	n.s.	.70**	n.s.
Education of the mother ***[ref. No education]***										
Primary	n.s.	n.s.								
Secondary/higher	.71**	.74*								
Mother has freedom of movement [ref. No]										
Yes		n.s.	n.s.	n.s.						
Mother participates to household's decisions [ref. No]										
Yes					n.s.	n.s.				
Mother justifies wife-beating [ref. No]										
Yes							n.s.	n.s.		
Mother participates to her children's healthcare decisions [ref. No]										
Yes									.72***	.73***

n.s.: not statistifically significant at 10%, *p<0.10; **p<0.05, *** p<0.01; all models include constants. Note that all the odds ratio should be interpreted in terms of association, and not of causality.

Eventually, models 9 and 10 allow us to shed more light on the relationship between the women's empowerment as mothers – measured as possibility of agency when it comes to their children's healthcare - and their children's survival. It appears that such measure of women's empowerment does have a protective effect in terms of child survival: children born to women who have the final say in decisions concerning their kids' healthcare have 29% lower odds to die before reaching their fifth birthday, compared to those born to women who do not have any say on kids' healthcare. Such protective effect is retained even when the contextual factors are included in the model as control, and the result stays statistically significant.

Hence, it appears that women's education has a statistically significant effect on child mortality both when controlling only for individual-level variables and when including the contextual-level ones, and so does women's empowerment – when measured in terms of possibility to act when it comes to their children's healthcare. This latter result is consistent to what Kabeer [9] argued, that is, that what matters is women's agency as mothers rather than as wives. Indeed, in our analysis the measure focusing on women's ability to take action in safeguarding their children's healthcare has a significant, protective effect on children's survival, whereas measures of empowerment in terms of equality within the marital relationship do not show any significant impact on the chances of child survival.

References

1. Alemayehu, Y. K., Theall, K., Lemma, W., Hajito, K. W., & Tushune, K. (2015). The role of empowerment in the association between a woman's educational status and infant mortality in Ethiopia: Secondary analysis of demographic and health surveys. *Ethiopian journal of health sciences*, *25*(4), 353-362.
2. Bloom, S. S., Wypij, D., & Das, G. M. (2001). Dimensions of women's autonomy and the influence on maternal health care utilization in a north Indian city. *Demography*, 38, 67-78.
3. Caldwell, J.C. (1979). Education as a factor in mortality decline: an examination of Nigerian data. *Population Studies* 33(3):395-413.
4. Caldwell, J., & McDonald, P. (1982). Influence of maternal education on infant and child mortality: levels and causes. *Health policy and education*, *2*(3), 251-267.
5. Chowdhury, A. M., Bhuiya, A., Chowdhury, M. E., Rasheed, S., Hussain, Z., & Chen, L. C. (2013). The Bangladesh paradox: Exceptional health achievement despite economic poverty. *The Lancet*, 382:1734– 1745.
6. DHS Stat Compiler. (2018). Retrieved from https://www.statcompiler.com/en/#cc=BD&ic=CM_ECMR_C_U5M&scl=0&scc=BD&dt=0&pt=0&ss=0&gr=1. Accessed on February 12, 2018.
7. Griffis, H. M. (2012). *Women's empowerment and infant and child mortality: Incorporating social institutions and context*. The Florida State University.
8. Hossain, B. (2015). Women empowerment and infant mortality in Bangladesh. *Applied Economics*, *47*(51), 5534-5547.
9. Kabeer, N. (1999). Resources, agency, achievements: Reflections on the measurement of women's empowerment. *Development and change, 30*(3), 435-464.
10. Maitra, P. (2004). Parental bargaining, health inputs and child mortality in India. *Journal of health economics*, *23*(2), 259-291.
11. Matteson, D. W., Burr, J. A., & Marshall, J. R. (1998). Infant mortality: a multi-level analysis of individual and community risk factors. *Social Science & Medicine*, *47*(11), 1841-1854.
12. Masuy-Stroobant, G. (2002). The determinants of infant mortality: how far are conceptual frameworks really modelled? In *The explanatory power of models* (pp. 15-39). Springer Netherlands.
13. Mosley, W. H., & Chen, L.C. (1984). An analytical framework for the study of child survival in developing countries. *Population and development review, 10*, 25-45.