ASA CONFERENCE 2019
Statistics for Health and Well-being

BOOK OF SHORT PAPERS

Maurizio Carpita and Luigi Fabbris
Editors

ASA Conference 2019 - Book od Short Papers
Statistics for Health and Well-being
University of Brescia, September 25-27, 2019
Maurizio Carpita and Luigi Fabbris (Editors)

ISBN: 978-88-5495-135-8
This Book is published only in pdf format.

INTRODUCTION

This Book includes a selection of 53 peer-reviewed short papers submitted to the Scientific Conference "*Statistics for Health and Well-Being*", held at the University of Brescia from 25 to 27 September, 2019.

The Conference, aimed at promoting applications that use statistical techniques and models suitable for health and well-being analyses, was organized by the ASA (Association for Applied Statistics) and the DMS StatLab (Data Methods and Systems Statistical Laboratory) of the Department of Economics and Management, University of Brescia.

The programme of the Conference included 25 parallel sessions with a total of 82 contributions with about 100 attendants, 4 plenary sessions (organised by ISTAT, the Italian National Statistical Institute, and USCI, the Statistical Union Italian Municipalities; SIS, the Italian Statistical Society, and ASA; AICQ-CN, the Italian Association for Quality Culture-North and Centre of Italy, and AISS, the Italian Academy for Six Sigma; and DBSPORTS, Big Data Analytics in Sports Project, respectively) and 4 special events (ISTAT and ASA Open Conference with the President of ISTAT, IASA Sensory Experiment, Visit to Capitolium, and Kick-off meeting ISI-SPG in Sports Statistics). Thank you very much to Eugenio Brentari, Chair of the Local Program Committee. For more information about the programme and other material visit the website www.sa-ijas.org/statistics-for-health-and-well-being/.

As co-chairs of the ASA Conference 2019, we are very grateful to the authors for submitting their interesting research with various real application of statistics in so many contexts of health and well-being, and to the members of the Scientific Committee for collaborating to the peer-reviewing process.

October, 2019

Co-chair Scientific Program Committee

Maurizio Carpita
University of Brescia

Luigi Fabbris
University of Padova

Conference session topics include, but are not limited to, the following areas of special interest:

Health and healthcare
Education and health
Health Psychology
Work and life balance
Economic well-being
Social relationships and social health
Welfare and well-being
Safety and security
Subjective well-being
Environment and pollution
Innovation, research and creativity
Quality of health services
Equitable and sustainable well-being
Resilience and vulnerability
Sport, Health and wellbeing
Sport analytics
Health and fitness
Sport psychology
Statistics and tourism
Food and beverage, health, well-being and life quality
Qualitative and quantitative methods for sensory analysis
Psychology and food
Food and beverage industries and markets
Methods and models for health and well-being analysis
Technology for health analysis

Scientific Program Committee:

Luigi Fabbris (University of Padua, co-chair)
Maurizio Carpita (University of Brescia, co-chair)

Giuseppe Arbia (SIS - Università Cattolica di Milano)
Rossella Berni (University of Florence)
Matilde Bini (SIS - European University of Rome)
Giovanna Boccuzzo (University of Padova)
Eugenio Brentari (University of Brescia)
Vittoria Buratta (ISTAT)
Giulia Cavrini (University of Bolzano-Bozen)
Alessandro Celegato (AICQ-AISS, PSV Project Service and Value)
Giuliana Coccia (ISTAT)
Adriano Decarli (University of Milan)
Tonio Di Battista ('G. D'Annunzio' University of Chieti and Pescara)
Simone Di Zio ('G. D'Annunzio' University of Chieti and Pescara)
Benito Vittorio Frosini (Sacred Heart Catholic University of Milan)
Antonio Giusti (University of Florence)
Silvia Golia (University of Brescia)
Maria Gabriella Grassia (Federico II University of Naples)
Maria Iannario (Federico II University of Naples)
Domenica Fioredistella Iezzi (Tor Vergata University of Rome)
Michele Lalla (University of Modena and Reggio Emilia)
Fabio Lucidi (SIPSA - La Sapienza University of Rome)
Marica Manisera (University of Brescia)
Paolo Mariani (University of Milan-Bicocca)
Francesco Mola (University of Cagliari)
Antonio Mussino (La Sapienza University of Rome)
Luigi Odello (International Academy of Sensory Analysis)
Francesco Palumbo (Federico II University of Naples)
Maurizio Pessato (Assirm)
Alessandra Petrucci (University of Florence)
Alfonso Piscitelli (Federico II University of Naples)
Marco Trentini (Unione Statistica Comuni Italiani)
Fabio Vernau (Federico II University of Naples)
Domenico Vistocco (Federico II University of Naples)
Paola Zuccolotto (University of Brescia)

Local Program Committee:

Eugenio Brentari (University of Brescia, chair)

Maurizio Carpita (University of Brescia)
Silvia Golia (University of Brescia)
Marica Manisera (University of Brescia)
Manlio Migliorati (University of Brescia)
Anna Simonetto (University of Brescia)
Marika Vezzoli (University of Brescia)
Mariangela Zenga (University of Milano-Bicocca)
Paola Zola (University of Brescia)
Paola Zuccolotto (University of Brescia)

Visit to the Capitolium. Brescia, 26th September 2019

INDEX OF SHORT PAPERS

Observed expenditures vs estimated burden of health care: a comparative evaluation based on spatial analysis

Giuseppe Alfonzetti[a], Laura Rizzi[a], Luca Grassetti[a], Michele Gobbato[b]
[a] Department of Economics and Statistics, University of Udine, Udine, Italy;
[b] S.O.C. Epidemiologia Oncologica – CRO, IRCCS, Aviano (PN), Italy.

1. Introduction and aims

The worldwide increase in the proportion of population older than 65 has become a subject of concern for policymakers (Gray, 2005), specifically for its drawbacks on health care expenditure (HCE). Some literature spread the concern that an ageing population would bring HCE to unsustainable levels. This scenario is based on the assumption that per capita health expenditure increases by age at constant rates over time. However this thesis is not unanimously supported for at least three primary considerations: health care expenditure growth is affected by other factors also, as technology and economic or institutional factors; the increasing life expectancy leads health care costs to shift in later years of individuals' life; and, finally, even if ageing really brings higher costs, resources could just be reallocated among the population.

This study aims to deepen the time and spatial patterns of health care expenditure and burden in the elderly population of Friuli Venezia Giulia (FVG), and to derive some evidence-based suggestions for the management of public resources. To this end, the time trend of general population HCE is fist analysed to identify the specific population ageing patterns. However, the assessment of spatial heterogeneity in the elderly healthcare burden is the primary goal of the paper. For this reason, a specific analysis is developed to identify the factors influencing the spatial heterogeneity of the health care burden in the elderly population. The use of observed HCE in this framework brings to unsatisfactory results. In particular, the formal tests for spatial correlation show no significant results. The challenge is to consider an indicator summarising the patients' chronicity and health severity while taking into account the intricate pattern of socio-economic determinants for health care demand. The Resource Utilization Band (RUB) indicator, provided in the Adjusted Clinical Groups (ACG) System developed by The John Hopkins University, seems to overcome the HCE issues.

After a brief introduction to the empirical settings, the paper focuses on the temporal and spatial descriptive analysis of HCEs and of RUBs measures and on the results of formal analysis of spatial heterogeneity. Conclusions are finally given.

2. Data and methods

The data used in the present analysis relates to the whole residents in FVG, even if this study focuses on people aged more than 65. The final dataset includes expenditures for pharmaceutical, hospital inpatient, and outpatient services, provided by the public regional health service in the period 2002-2017, and health severity, measured by the RUB indicator, in 2017 and 2018. Using the John Hopkins ACG System methodology, the population is divided into mutually exclusive groups sharing a similar morbidity profile, where each group requires a certain level of healthcare resources. For the RUB classification, ACGs are merged according to their use of resources and mapped into an ordinal variable with six categories (indexed as 0-5).

Data are collected clustering the population by age classes and gender within each municipality. At the municipality level, other measures such as population size, the ratio between

ISBN: 978-88-5495-135-8

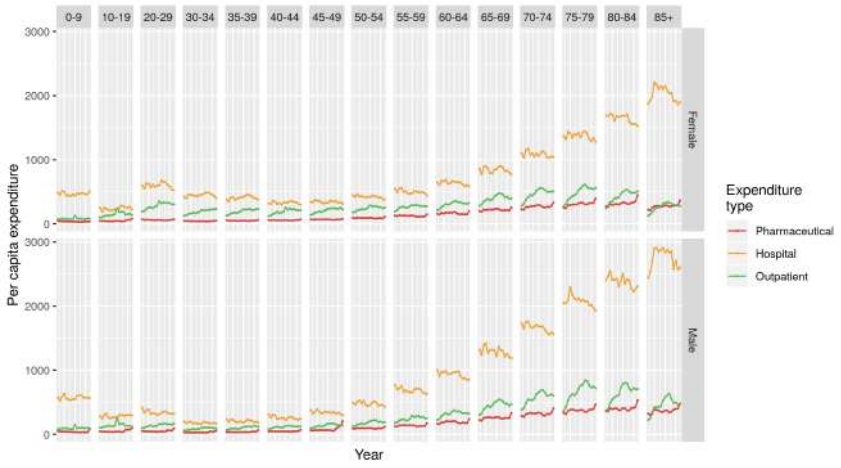


Figure 1: Per capita HCE time series dissaggregated by gender and age classes

males and females, death rate, count of 21 chronic conditions and variables relative to the population economic conditions are also considered.

The spatial heterogeneity of health care need is described focusing only on the elderly population. The presence of spatial autocorrelation is studied considering the classical Moran's I test. Further, some spatial econometrics models (such as those discussed in Elhorst, 2014; Moscone and Tosetti, 2014; LeSage and Pace, 2009) have been compared. The model selected for the analysis is the Spatial Durbin model. The model specification is as

$$Y = \rho WY + X\beta + WX\theta + \epsilon \tag{1}$$

where $\epsilon \sim N(0, \sigma_\epsilon I)$ and the neighbouring effects are introduced both in the spatial lagged response variable (WY) and in the regressors (WX). The direct effects of the esplicative variables are also included.

3. Descriptive analysis and preliminary results

The analysis of demographic trend and ageing phenomenon points out that the crowning in the population pyramid, which now lies within the 45-64 class, is moving upward, leading to a deflation of the 45-64 class and further swelling of the over 65 one. In particular, the cumulative share of the population in 45-64 and 65+ classes is worth at least the 50% in all the provinces.

Focusing on the elderly population, the total amount of HCE shows a relevant increase. Each component of the expenditure increased by almost 150 millions of euros in fifteen years. While the total HCE for hospital and outpatient services grew significantly from 2002 to 2010, the increase in total pharmaceutical expenditure is lower. The trend of the former types of services drives the overall trend of total expenditure because of their weight in the regional composition of healthcare services. The trend in pharmaceutical expenditure shows constant growth rates, with a steep acceleration from 2016 to 2017. These results assess partially the role of population ageing, while the effect of age class and gender on per capita HCE is described in Figure 1. The per-capita HCE scenario points out, however, that the increase can be attributed both to higher per capita expenditures in pharmaceutical and outpatient services and to the demographic pressure towards older age classes.

The analysis of RUBs distribution shows that more than the 50% of people fall within RUB 0 and RUB 1, the healthiest bands, while 2 and 3 stay at similar levels of band 0, and cover, in couple, almost the 40%. Levels 4 and 5, instead, the most expensive ones, reach the 7%

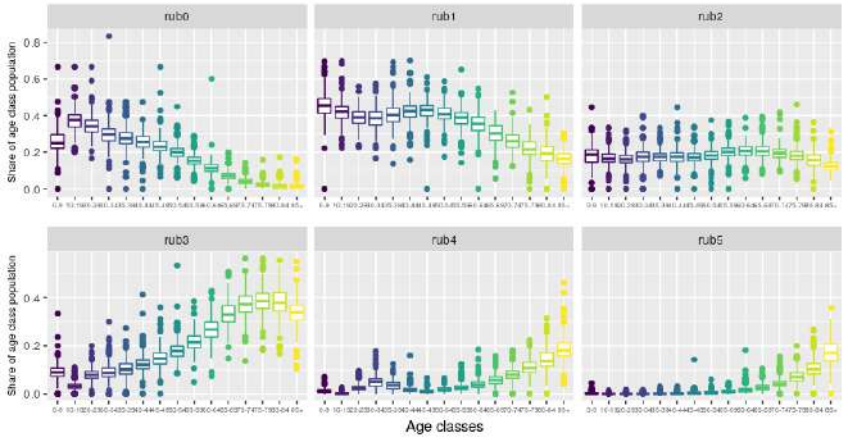


Figure 2: RUBs shares across age classes in 2017

of the population barely. Undoubtedly the need for healthcare resources increases with age. Moreover, the overall growth in the need for health care resources is mainly due to the share of RUB 3 people, namely people with a multimorbidity situation. In the seventies, this proportion is stable, and bands 4 and 5, which represent people with a severe health profile, show an expansion (see Figure 2).

In terms of health burden, the RUBs 3, 4 and 5 account together for more than the 50% of the population older than 65. For this reason, the proportion of older people with RUB 3-5 within each municipality is considered as the outcome variable. Its deviation from the regional mean level is the study variable whose spatial heterogeneity is evaluated through the SDM model. Descriptives of deviations of RUBs for elderly population are computed and mapped. In particular, the maps for the different levels of RUB point out a contrast between the province of Pordenone, the northern area, and the rest of the region. Then over 65 people seem to have a more severe healthcare burden in the northern area of the region. Relevant dependence patterns characterise the health severity distribution of the elderly population, with the areas of Pordenone and Gorizia identified as the healthier part of the over 65 regional population, as given in Figure 3.

The presence of spatial correlation in the RUBs proportion within elderly people is formally tested, and the estimation results concerning the Spatial Durbin Model (omitted for space reason) show that the neighbouring effect is significant. Most of the regressors generate higher indirect impacts rather than direct ones, indicating the power of the spillover effects accounted within the model. The spatial patterns of morbidities play an essential role in the explanation of the healthcare burden, together with the economic characteristic of the municipality, in particular, those related to the yearly amount of income. The role of the proportion of over 65 people is entirely unexpected since its estimated impact is negative. Higher the percentage of older people corresponds to lower values of the healthcare burden. In other words, when higher proportions of their peers surround older people, they seem to be healthier and need fewer healthcare services. Finally, the model provides further insights on the diseases whose spatial patterns mostly influence the healthcare burden, namely age macular degeneration, human immunodeficiency virus and low back pain.

4. Conclusions

The demographical ageing process in FVG is firmly higher than the national levels. In particular, people older than 65 account for 25% of the total population, against the 22% national

Figure 3: Map of RUB deviations in 2017

average. Moreover, the population average age in 2017 is 46.6 years old, in FVG, against the national average of 44.4. Following the Red Herring literature Zweifel et al. (1999), the descriptive analysis highlights how per capita expenditure, in particular for hospital services, decreased with years, being postponed to older ages thanks to the increase of life expectancy. In the present framework, the RUB indicator is used as a proxy for the healthcare burden. The spatial analysis points out any spatial pattern for HCE deviations from the regional average, while a robust geographical clustering characterises the RUB indicator deviations even after controlling for the demographical structure of municipalities. An econometric approach is used to model the spatial dependence and to identify the factors determining the healthcare burden, and the results of the analysis can be directly used to support the evidence-based decision-making processes.

Further developments would consider the use of disaggregated data and a possible relationship between HCE and RUB indicators. The idea is to model the phenomenon at the individual level through multinomial spatial regression.

References

Elhorst, J.P. (2014), Spatial Panel Data Models in *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*, Springer, Berlin Heidelberg, (DE), pp. 37–93.

Gray, A. (2005). Population Aging and Health Care Expenditure. *Ageing Horizons*, **2**, pp. 15–20.

LeSage, J. and R. K. Pace (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC, New York.

Moscone, F. and Tosetti, E. (2014). Spatial econometrics: Theory and applications in health economics, in *Encyclopedia of Health Economics*, eds. A. J. Culyer, Elsevier, San Diego, pp. 329–334.

Zweifel, P., Felder, S. and Meiers, M. (1999). Ageing of population and health care expenditure: a red herring?. *Health Economics*, **8(6)**, 485–496.

Computing ordinal consistency thresholds for pairwise comparison matrices

Pietro Amenta [a], Antonio Lucadamo [a], Gabriella Marcarelli [a]
[a] Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Benevento, Italy

1. Introduction

Pairwise comparison matrices (PCMs) are widely used for representing preferences in multi-criteria decision problems. To derive the ranking of preferences by means of pairwise comparisons, a positive number a_{ij} is assigned to each pair of elements (x_i, x_j) with $i, j = 1, \ldots, n$. This number expresses how much x_i is preferred to x_j as regards a given criterion. By comparing all the elements, a positive square matrix $A = (a_{ij})$ of order n is then obtained. The value $a_{ij} > 1$ implies that x_i is strictly preferred to x_j, whereas $a_{ij} < 1$ expresses the opposite preference, and $a_{ij} = 1$ means that x_i and x_j are indifferent (Saaty, 1980; Saaty, 1994). This matrix is at the heart of several methods that have been proposed in the literature to derive a priority vector, $w = (w_1 \ldots w_n)$, representing the ranking of preferences: the Eigenvector Method (EVM), the Arithmetic Mean Method (AMM), the Row Geometric Mean Method (RGMM), the logarithmic Least Squares method, the Singular Value Decomposition, to cite just a few (Aguaron *et al.*, 2003; Gass and Rapcsak, 2004; Pelaez and Lamata, 2003; Saaty, 1980). Regardless of the method chosen for the prioritisation procedure, before applying any methods, it is then necessary to check the consistency of these judgements. Consistency may be ordinal or cardinal. The cardinal consistency implies that the judgements are transitive and proportional: a decision maker is perfectly consistent in making estimates if his or her judgements satisfy the following consistency condition $a_{ij} * a_{jk} = a_{ik}$ for each $i, j, k = 1, 2, \ldots, n$ (Saaty, 1980). For example, if $a_{12} = 2$ and $a_{23} = 3$ then a_{13} must be equal to 6 to ensure that a 3×3 pairwise comparison matrix is perfectly consistent. In the case of perfect consistency, the following equality holds: $a_{ij} = w_i/w_j$.

The ordinal consistency implies instead only the transitive property; meaning that, if $a_{ij} > 1$ and $a_{jk} > 1$, then $a_{ik} > 1$. Transitivity is a condition weaker than consistency. Perfect consistency is unattainable in practice, but a degree of inconsistency can be considered acceptable. The consistency of judgements is strictly connected with the reliability of the preferences expressed by the priority vector. If the judgements are not consistent, then the prioritisation methods could provide different results. If the judgements are instead only ordinally consistent (that is, only transitive), then most methods provide vectors representing the same ranking, expressing in this way the same preferences: only the intensity of the preferences can vary (Siraj *et al.*, 2015). Due to its relationship with the reliability of the preferences, the consistency of judgements has been widely analysed by many authors. Several indices have been proposed to measure the degree of consistency of the judgements expressed by the decision maker. Each index is a function that associates pairwise comparisons with a real number that represents the degree of inconsistency in the judgements. Here we introduce some.

Saaty proposed the Consistency Index (CI), given by

$$CI = \frac{\lambda_{max} - n}{n - 1}, \tag{1}$$

for $i, j = 1, \ldots, n$, where λ_{max} represents the maximum eigenvalue of the pairwise comparison matrix. If the matrix is perfectly consistent, then $CI = 0$. Saaty suggested also the Consistency

ISBN: 978-88-5495-135-8

Ratio

$$CR = \frac{CI}{RI}, \tag{2}$$

where RI is the Random Index, which is obtained as the mean value of the CI derived from randomly generated matrices of order n.

Crawford and Williams, (Crawford and Williams, 1985) suggested a measure of inconsistency based on the estimator of the variance of the perturbation, when the Row Geometric Mean Method (RGMM) is used as prioritization procedure. Assuming their proposal, Aguaron and Jimenez calculated the thresholds for that measure, called Geometric Consistency Index (Aguaron *et al.*, 2003). The GCI index, based on the logarithmic residual mean square, is defined as:

$$GCI = \frac{2}{n(n-1)} \sum_{i<j} \log^2 e_{ij}, \tag{3}$$

where $e_{ij} = a_{ij} \times \frac{w_j}{w_i}$ represents the error obtained when the ratio w_i/w_j is approximated by a_{ij} and w is the vector derived by the RGMM.

Koczkodaj has defined the following consistency measure:

$$CM_K = \max_{i,j,k} \left[\min \left(\left| 1 - \frac{a_{ik}}{a_{ij}a_{jk}} \right|, \left| 1 - \frac{a_{ij}a_{jk}}{a_{ik}} \right| \right) \right], \tag{4}$$

based on the triplet of the elements of a pairwise comparison matrix, with $1 \leq i < j < k \leq n$.

The Salo-Hamalainen Consistency Index (Salo and Hamalainen, 1997) is defined as:

$$CM_{SH} = \frac{2}{n(n-1)} \sum_{i>j} \frac{\bar{r}(i,j) - \underline{r}(i,j)}{(1 + \bar{r}(i,j))(1 + \underline{r}(i,j))} \tag{5}$$

where

$\bar{r}(i,j) = \max_k (a_{ik} \cdot a_{kj})$ and $\underline{r}(i,j) = \frac{1}{\bar{r}(j,i)}$ CM_{SH} can be applied to all reciprocal matrices, regardless of the scale used and, like CM_K, is not linked to any prioritisation method (Salo and Hamalainen, 1997). Consistency indices and their thresholds may be useful to face cardinal consistency but they do not take into account the ordinal consistency (transitivity). Consequently, we focus on the transitivity and propose a transitivity threshold that could be useful because it may provide meaningful information about the reliability of the preferences and it may also allow us to avoid the revision of judgements. If the decision maker is interested in the ordinal ranking of elements and not in the intensity of preferences, then a transitivity threshold represents an important tool for this task: an index value less than the transitivity threshold ensures (with a high probability) that the ranking of preferences is unique on varying the prioritisation methods, only the intensity of preferences may be different.

2. Ordinal consistency thresholds

Although transitivity has represented a cornerstone of normative decision theory, many authors have criticised that principle because it forces us to assume that judgements satisfy this property. Saaty's consistency threshold has been criticised because it may allow us to accept many intransitive matrices or reject many transitive ones. For this reason we introduce a method to verify if a PCM is transitive or not.

To define the transitivity thresholds we generate 500000 random comparison matrices of size n and, using an algorithm based on the approach introduced by Gass (Gass, 1998), we check how many transitive and intransitive matrices are generated (the proportion of two categories varies as n varies). We then compute CR, GCI, CM_{SH} and CM_k for all matrices and, in order to define the thresholds associated with these indices, we introduce the following notation:

- let λ and $1-\lambda$ be the proportion of random generated intransitive and transitive matrices respectively;
- let α be the percentage of random intransitive matrices that are accepted according to the threshold value to be set;
- let β be the percentage of random transitive matrices that are rejected according to the same value.

Given λ, we suggest to define the threshold as the value that minimizes the quantity $\lambda\alpha + (1-\lambda)\beta$. Table 1 shows the transitivity-intransitivity thresholds for each index and the corresponding percentage of misclassified matrices for sizes 3 to 8.

Table 1: Transitivity-intransitivity thresholds for each index and different matrix size orders (n)

n		CM_{SH}	CR	GCI	CM_K	λ	$1-\lambda$
3	Threshold	0.586	1.405	3.968	0.956	0.2494	0.7506
	Misclassification rate	0.348	2.155	2.155	2.194		
4	Threshold	0.527	0.647	1.862	0.969	0.6248	0.3752
	Misclassification rate	6.597	6.118	7.147	3.199		
5	Threshold	0.511	0.440	1.353	0.969	0.8815	0.1185
	Misclassification rate	5.046	4.959	5.684	1.886		
6	Threshold	0.510	0.327	0.947	0.969	0.9787	0.0213
	Misclassification rate	1.509	1.645	1.758	0.588		
7	Threshold	0.507	0.256	0.798	0.969	0.9977	0.0023
	Misclassification rate	0.206	0.221	0.226	0.089		
8	Threshold	0.506	0.254	0.856	0.969	0.9998	0.0002
	Misclassification rate	0.014	0.015	0.015	0.007		

3. Application

In order to highlight the usefulness of our proposal, we consider a tourist accomodation that want to evaluate the satisfaction of the tourists about its services. In particular, each customer is asked to give a pairwise comparison among the following characteristics: food service, cleanliness, staff, price/quality ratio and comfort. The aim is to evaluate which service is preferred to the others, via an aggregation method. Before applying any procedure is anyway necessary to check the consistency of the judgments. If the decision makers are consistent, then they can be considered in the analysis, otherwise, according to classical procedure, the judgments must be revised. In many occasions anyway it is not possible to contact the decision makers and in these cases, the matrix can not be used in the analysis or some procedure to force it to be consistent must be introduced. Let consider for example the following matrix filled in by a customer:

	Food	Cleanliness	Staff	Price/quality	Comfort
Food	1	1/3	7	8	1/4
Cleanliness	3	1	3	3	1/7
Staff	1/7	1/3	1	1/8	1/6
Price/quality	1/8	1/3	8	1	1/2
Comfort	4	7	6	2	1

ISBN: 978-88-5495-135-8

Looking at table 2 it is easy to notice that according to the classical consistency thresholds for all indices, the judgments in this matrix should be revised. If we are interested instead only in ordinal consistency, we can consider the transitivity thresholds we propose and we can avoid to revise the matrix, using it for any aggregation and prioritization method (Pelaez & Lamata, 2003).

Table 2: Consistency and Transitivity thresholds associated with a PCM in the case of n=5 and indices' values corresponding to the previous matrix

	CR	CM_{SH}	GCI	CM_K
Consistency thresholds	0.10	0.33	0.36	0.88
Transitivity thresholds	0.440	0.511	1.353	0.969
Indices' values for the matrix	0.3844	0.5055	1.2734	0.9375

4. Concluding remarks

Due to some criticisms on consistency thresholds, particularly regarding its inability to capture the ordinal consistency, we propose approximated transitivity threshold for some consistency indices. This threshold is useful because it may allow to avoid the revision of the judgements if the decision maker is only interested in the ordinal consistency. If the value assumed by the consistency index is ranged between the consistency and the transitivity threshold values, then we are confident about the reliability of the preferences. In this case, the decision maker avoids the need to revise his or her judgements.

References

Aguaron, J., Moreno-Jimenez, J. (2003) The geometric consistency index: Approximated threshold, *European Journal of Operational Research* 147, 137–145.

Crawford, G., Williams, C. (1985) A note on the analysis of subjective judgment matrices, *Journal of Mathematical Psychology* 29 387–405.

Gass, S.I. (1998) Tournaments, transitivity and pairwise comparison matrices. *The Journal of the Operational Research Society* 49(6) 616-624

Gass, S., Rapcsák, T. (2004) Singular value decomposition in ahp, *European Journal of Operational Research* 154 573–584.

Koczkodaj, W. (1993) A new definition of consistency of pairwise comparisons, *Mathematical and Computer Modelling* 18, 79–84.

Peláez, J., Lamata, M. (2003) A new measure of consistency for positive reciprocal matrices, *Computer and Mathematics with Applications* 46 1839–1845.

Saaty, T. (1980), *Multicriteria Decision making: The Analytic Hierarchy Process*, McGraw-Hill, New York.

Saaty, T. (1994) *Fundamental of decision making and priority theory with the AHP*, RWS Publications, Pittsburgh.

Salo, A., Hamalainen, R. (1997) On the measurement of preference in the analytic hierarchy process, *Journal of Multi-Criteria Decision Analysis* 6, 309–319.

Siraj, S., Mikhailov, L., Keane, J. (2015) Contribution of individual judgments toward inconsistency in pairwise comparisons, *European Journal of Operational Research* 242 557–567.

ISBN: 978-88-5495-135-8

Household wealth and consumption in Italy: Analysis by density-weighted quantile regression

Ilaria Lucrezia Amerise [a], Agostino Tarsitano [a]
[a] Dipartimento di Economia Statistica e Finanza, University of Calabria, Rende (Cs), Italy.

1. Introduction

With the re-awakening of interest in the social-welfare aspect of economic change, the study of the relationships between wealth and consumption has begun to draw renewed attention. In this paper we examine the distributional impact of various features of the wealth distribution at household level, using data drawn from the Bank of Italy's survey of household income and wealth (SHIW). The purpose of our analysis is twofold. One is the notion that there is some relation between the amount of consumption and wealth either in terms of real assets, or in terms of financial assets, or as payroll income. The other is that the major effect of those covariates can be better portrayed and analyzed with the aid of weighted quantile regression rather than with the tools currently being used such as least squares.

Consider the consumption with observed values $\{y_1, y_2, \cdots, y_n\}$ and a corresponding set of covariates $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ where $\mathbf{x}_i$ is a $m \times 1$ vector for a sample of n observations from a distribution function F and $\boldsymbol{\beta}_\tau$ is a $(m \times 1)$ vector of unknown parameters which are to be estimated. We assume that $\mathbf{x}_i$ are rows of a $n \times m$ design matrix $\mathbf{X}$ with full rank m and $m < n$. We assume further that $x_{i,1} = 1 \; \forall \; i$ so that the design matrix $\mathbf{X}$ contains a column of ones.

Let $Q(\tau|\mathbf{x})$ be the τ-th conditional quantile regression. It is evident that, this is a function of $\mathbf{x}$. Depending on F, its expression can be quite complex. However, as a first approximation, it seems reasonable to search for a best linear fit

$$Q_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^t \boldsymbol{\beta}_\tau \qquad i = 1, 2, \cdots, n; \quad 0 < \tau < 1 \tag{1}$$

The rest of the paper is organized as follows. Section 2 gives an overview of the density-weighted quantile regression and presents the method as an iteratively re-weighted quantile regression method. In section 3 we explore and discuss the empirical results from an empirical analysis on consumption expenditure and wealth at household level.

2. Density-weighted quantile regression

As a robust data analysis technique, quantile regression has attracted extensive interest. Substantial efficiency might be gained by incorporating an appropriate weight function to account for inhomogeneity of the variances in the model specification and to attenuate the effects of heavy tails in model fitting. In this study, the weights are based on the unconditional sparsity function. This suggestion builds on the idea of Koenker (2005)[p. 160], who states that, rather than weighting by the reciprocal of the standard deviations of the observations, QR weights should be proportional to the unconditional density of the response.

The proposed estimator is the vector $\widehat{\boldsymbol{\beta}}_{\eta,\mathbf{w}^{(0)}}$ that solves the following problem

$$\min_{\boldsymbol{\beta}_\mathbf{w}} \left[\eta \sum_{y_i \geq \mathbf{x}_i^t \boldsymbol{\beta}_{\eta,\mathbf{w}}} w_i \left(y_i - \mathbf{x}_i^t \boldsymbol{\beta}_{\eta,\mathbf{w}} \right) + (1-\eta) \sum_{\mathbf{x}_i^t \boldsymbol{\beta}_{\eta,\mathbf{w}} > y_i} w_i \left(\mathbf{x}_i^t \boldsymbol{\beta}_{\eta,\mathbf{w}} - y_i \right) \right] \tag{2}$$

Minimization of (2) involves two sets of unknowns: m coefficients $\boldsymbol{\beta}_\eta = (\beta_{1,\eta}, \beta_{2,\eta}, \cdots, \beta_{m,\eta})$ and n weights $w_i, i = 1, 2, \cdots, n$. A single step is impossible because the parameters cannot be

ISBN: 978-88-5495-135-8

estimated without knowing the weights and the weights cannot be determined until the model is fitted. To estimate both components, we apply an iteratively re-weighting method.

Initially, we suppose that the weights are known, for example, $w_i^{(0)} = 1 \; \forall \; i$. Let $\boldsymbol{\beta}_{\eta,\mathbf{w}^{(0)}}$ be a solution of (2) obtained, for example by using the well-known package $quantreg$. To update weights we have to consider the unconditional density function of the residuals. This, however, is generally not known. We can get around this problem by noting that the density function is the reciprocal of the sparsity function that is, the derivative of the quantile function

$$f\left[Q_\eta\left(y|\mathbf{x},\mathbf{w}^{(0)}\right)\right] = \frac{1}{s\left[Q_\eta(y|\mathbf{x},\mathbf{w}^{(0)})\right]} \quad \text{with } s\left[Q_\eta\left(y|\mathbf{x},\mathbf{w}^{(0)}\right)\right] = \frac{dQ_\eta\left(y|\mathbf{x},\mathbf{w}^{(0)}\right)}{d\eta} \tag{3}$$

An approximation of $s()$ can be readily obtained by using the symmetric difference quotient. Let d_n be a bandwidth that tends to 0 as $n \to \infty$ and let $\widehat{Q}_{\eta+Kd_n}\left(y|\mathbf{x}_i,\mathbf{w}^{(0)}\right) = \mathbf{x}_i^t\boldsymbol{\beta}_{\eta+Kd_n,\mathbf{w}^{(0)}}$ be an estimate of the conditional quantile function at $\eta + Kd_n$, $K = -2,-1,1,2$. At any value $\mathbf{x}_i$, a five-point stencil in one dimension provides an approximation of the sparsity function

$$\widehat{s}_{i,\mathbf{w}^{(0)}} = \frac{\mathbf{x}_i^t\left[\boldsymbol{\beta}_{\eta-2d_n,\mathbf{w}^{(0)}} - \boldsymbol{\beta}_{\eta+2d_n,\mathbf{w}^{(0)}} + 8\boldsymbol{\beta}_{\eta+d_n,\mathbf{w}^{(0)}} - 8\boldsymbol{\beta}_{\eta-d_n,\mathbf{w}^{(0)}}\right]}{12d_n}, \quad i = 1,2,\cdots,n \tag{4}$$

with quartic truncation error. To make the difference quotient (4) operational, it is necessary to specify the bandwidth. There is extensive literature on the choice of d_n, but limited to the symmetric difference quotient (i.e. the slope of a nearby secant line through the points $\eta - d_n$, $f(\eta - d_n)$ and $\eta + d_n$, $f(\eta + d_n)$). To our knowledge, no indication has been given in the case of higher-order methods for approximating the first derivative. We have employed the bandwidth proposed by Bofinger (1975)

$$d_n = n^{-0.2}\left\{\frac{4.5\phi^4\left[\Phi^{-1}(\eta)\right]}{\left[2\left(\Phi^{-1}(\eta)\right)^2+1\right]^2}\right\}^{1/5} \tag{5}$$

where ϕ and Φ are the density and the standard Gaussian distribution function, respectively, with $\Phi[z_\alpha] = 1-\alpha$.

Despite the appeal, interpolation of the sparsity function has met with disappointment because quantile functions should be monotone increasing and consequently estimated quantile functions should not cross. To avoid this complication, we estimate the numerator of (4) under the following monotonicity restrictions

$$\operatorname{sgn}\left[\mathbf{x}_i\boldsymbol{\beta}_{\eta_2} - \mathbf{x}_i\boldsymbol{\beta}_{\eta_1}\right] \geq \epsilon \; \forall \; \mathbf{x}_i, \quad \eta_2 > \eta_1 \tag{6}$$

where $\operatorname{sgn}(.)$ equals to $-1, 0, 1$ according to whether the argument is negative, zero or positive and $\epsilon_i = 1.49{\times}10^{-8}$ is a small number to preclude confusion between overlapping and crossings. If conditions (6) are satisfied then the crossing will occur beyond the convex hull of observed data.

Now, we have to decide which level of η is the most appropriate for our analytical and general purposes based on the unconditional sparsity (and hence density) function. We propose using $\eta = 0.5$ because the asymptotic variance of quantile estimator is proportional to $\eta(1-\eta)/f[Q(\eta)]$, which is minimized at $\eta = 0.5$. Thus, the estimated median quantile is relatively more accurate than other quantiles. The interpolated sparsity (4) serves to re-define the weights as $w_i^{(1)} = 1/\widehat{s}_{i,\mathbf{w}^{(0)}}$ thus favoring the achievement of local efficiency in the sense of Koenker (2005)[p. 161].

At this point, the sparsity function can be re-interpolated using the updated weights $\mathbf{w}^{(1)}$, thus producing new weights $\mathbf{w}^{(2)}$, which, in turn, determines $\widehat{s}_{\mathbf{w}^{(2)}}$ and so on, until the relative absolute changes between $\mathbf{w}^{(i+1)}$ and $\mathbf{w}^{(i)}$ are below some specified tolerance. Despite the

complexity, we find that in practice, the process stabilizes after few iterations. Absolute convergence of the algorithm has not yet been established, but the method has converged for all the problems attempted by the authors.

The asymptotic properties of the density-weighted quantile regression estimators are developed in Theorem 5.1 in Koenker (2005)[Sect. 5.3.1]

3. Application to Bank of Italy survey data

This section describes a paradigmatic example of how density-weighted quantile regression may be useful in assessing whether the impact of wealth (prevalently, real assets) and disposable income (prevalently, net wages and salary) on consumption is stronger (or weaker) when the level of consumption is unusually high (or low). Data for this application are supplied by the survey Bank of Italy (2018), which covers $n = 7421$ households. In particular, we consider Consumption (C) , Real estate (R), Financial asset (F), Payroll income (Y), Household size (H). Following Jawadi and Sousa (2014), we estimate the relationship

$$(C_i) = \beta_{0,\tau} + \beta_{1,\tau} R_i + \beta_{2,\tau} F_i + \beta_{3,\tau} Y_i + \beta_{4,\tau} H_i + e_i, \quad i = 1, 2, \cdots, n \tag{7}$$

Table 1 reports the results of the proposed estimators. In passing, we have to remark that Bank of Italy (2018) gives sampling weights at household level. Nonetheless, the use of weights that come from a survey plan and not from an assumed covariance structure is a divisive subject in regression. For the current application we have ignored the survey weights.

Table 1: Density-weighted quantile regression estimates

η		Intercept	Real asset	Financial asset	Payroll income	Household size
0.05	β	4741.9	0.014	0.014	0.159	509.1
	t	18.61	13.31	4.12	12.07	3.49
0.10	β	5375.8	0.017	0.019	0.180	808.5
	t	29.28	21.37	6.65	19.23	8.48
0.25	β	7446.5	0.022	0.027	0.213	924.7
	t	42.91	30.12	7.79	23.19	11.37
0.50	β	9813.9	0.028	0.046	0.263	1150.7
	t	49.97	32.21	12.97	27.18	10.44
0.75	β	11661.4	0.035	0.070	0.296	2004.2
	t	44.30	34.51	10.73	23.65	14.45
0.90	β	14113.6	0.042	0.103	0.356	2967.6
	t	32.07	29.87	10.71	13.87	10.87
0.95	β	15739.3	0.048	0.127	0.359	3838.3
	t	24.96	15.61	6.28	11.10	12.68

The findings reveal that all the covariate have a direct effect on consumption and the effect grows with the level of the quantile. Moreover, the elasticity of consumption with respect to financial wealth is larger than the sensitivity of consumption with respect to real asset. This, in part, contradicts the tendency to high interest of household wealth in housing. For example, an empirical observation (Cannari and D'Alessio, 1990) about households in Italy is that they concentrate their wealth in housing and hold relatively limited financial assets. A possible

ISBN: 978-88-5495-135-8

explanation for restraints on consumption from wealth is that, for most families, the holding of financial wealth is in restricted accounts for pensions and insurance. They cannot easily withdraw these funds for current consumption, nor can they borrow against the collateral. The holding of unrestricted financial wealth is virtually nonexistent among lower-income families, and it is relatively limited even within high-income families. Presumably, tax policy has favored households concentrating their debt against their housing collateral, and by using larger mortgage balances and home equity lines to finance consumption.

Another important finding is that all the covariates employed in (7) appear to be highly significant in determining the amount of consumption. It is interesting to note that the asymptotic t-Student coefficients computed for the median quantile regression ($\eta = 0.5$) attains the most significant p-values. Conversely, the ordinary least squares (Ols) estimates fall in an intermediate category between the quantile levels 0.50 and 0.90.

The adjusted $\bar{R}^2$ obtained by least squares regression is 0.44. Even with all its limitations, a small $\bar{R}^2$, together with significant p- values, indicate that some important covariate is missing. In the quantile context we can use the average goodness-of-fit statistic proposed by Koenker and Machado (1999), which yield $\bar{R}_1 = 0.66$, thus confirming the omission of some data.

The Wald test of homoskedasticity in form of an F statistic (see Koenker, 2005[Sect. 3.3]) indicates the presence of a sever form of heteroskedasticity in the residuals of quantile regressions. Although, the rejection of the hypothesis of homoskedasticity does not prevent consistent point estimation of the parameters, it typically entails inefficiencies in confidence intervals very much like classical regression model. The differences in variances are also corroborated by the estimated density-weights that yield a Gini concentration index of 0.24 whereas a value near zero should be expected in the case of homoskedastic residuals.

In summary the study shows that there is a direct relationship between consumption and wealth in the context of Italian households: higher level of consumption are associated with higher wealth. This is not surprising as it is hardly surprising the positive link between consumption and household size. On the other hand, our empirical results are supportive of that the elasticity of consumption with respect of real estate, financial asset and payroll income increase with the quantile level, thus enlightening the gap between rich and poor households.

References

Bank of Italy (2018). Household income and Wealth in 2016. Supplement to the Statistical Bulletin - Sample surveys

Bofinger, E. (1975). Estimation of a density function using order statistics. *Australian Journal of Statistics*, **17**, pp. 192–195.

Cannari, L., D'Alessio, G. (1990) Housing assets in the Bank of Italy's survey of household income and wealth. In *Income and Wealth Distribution, Inequality and Poverty: Proceedings of the Second International Conference on Income Distribution by Size: Generation, Distribution, Measurement and Applications*, University of Pavia, Italy, September 28–30, 1989, pp. 326–334

Jawadi, F., Sousa, R. M. (2014). The relationship between consumption and wealth: A quantile regression approach. *Revue d'économie politique*, **124**, pp. 639–652

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

Koenker, R., Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**, pp. 1296–1310.

Ecotourism and food geographic areas

Fabrizio Antolini[a], Francesco Giovanni Truglia[b]
[a] Department of Communication Sciences, University of Teramo, Teramo, Italy
[b] Italian National Institute of Statistics, Istat, Agricultural Units, Rome, Italy

1. Introduction

One of the most important limitations in tourism concerns the intangibility of the products or services provided at the destination. In fact, the products and services provided contribute to the people's final choice that allows a geographical area to become a tourist destination. Food and wine traditions contribute to the attractiveness of locations, thereby transforming them into tourism destinations. While food is a tangible good, its contribution to the growth of tourism in locations is not, so food traditions, as well as wine traditions, are simultaneously tangible and intangible goods. When food is combined with tourism, it develops a particular type are able to develop a particular type of tourism called "slow tourism" or "ecotourism." Sustainability agendas suggest examining how the agricultural and tourism sectors can be combined in order to realize a sustainable development of territories. "Increasingly, regional tourism development initiatives are utilizing locally-produced foodstuffs and beverages to: strengthen tourism product areas; enhance visitors' experiences; and help to maintain and enhance the viability of local food production and processing sectors" (Boyne et al., 2003). In fact, rural tourism is more than a form of tourism: it represents a lifestyle that tourists want to live while experiencing sustainability. The objective of this paper is to analyze the flows of tourists seeking farm-holidays. This accommodation type indirectly shows an increasing sensitivity to the environment. Moreover, it expresses other aspects related to psychophysical lifestyle and well-being. In some regions of Italy, the increasing demand for this type of accommodation has been followed by a transformation of services, in particular, food services and wine traditions. In fact, many farm-holiday operators have linked their tourism activity to the local production of PDO-PGI goods. It will be interesting to analyze whether, in the regions considered (Tuscany and Apulia), there is a spatial convergence between the numbers of tourists, farm-holidays, and PDO-PGI products, which would thus identify touristic areas as "food vocations".

2. Spatial convergence among farm-holidays, food and tourism (FFT)

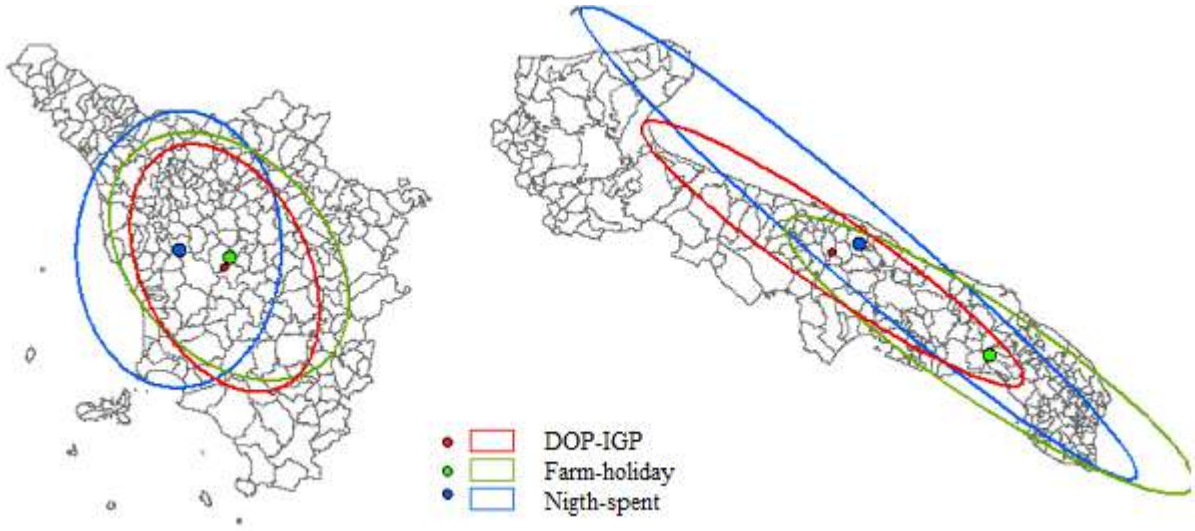
The regions considered are Tuscany and Apulia, in fact, in these geographical areas, the FFT factors seem to have been strongly developed. For sure, these regions present many differences, for example, Tuscany has a more developed entrepreneurial culture compared to Apulia. It means that a change in the services offered by farm-holidays should be present in Tuscany and to a lesser extent in Apulia, with a different identification of food vocation areas (FFT Areas). It could have a positive effect on the nights-spent density (nights- spent/square kilometres) in these areas.

First step is the identification of a centre of gravity or barycentre for the farm-holidays density (farm-holidays /square kilometres), the DOC-PGI products density (DOC-PGI/square kilometres), and the nights-spent density, thus using the official data at municipality level. As shown in Figure 1, it is possible to observe that the gravitational points of farm-holidays and the brand of quality for food products, are near in Tuscany but not in Apulia. It means that farm-holidays in Tuscany have changed their organization enriching their services with DOC and PGI products, too.

By looking at the night-spent indicators, we can observe that the representation suffers some limitations, in particular Tuscany where Florence inevitably shifts the centre of gravity. It does not concern Apulia, where the localization of this indicator seems to be more realistic. Nevertheless, the distance between the "central point" of the farms and the PGI-PDO products shows that the organizational structure of the farms has not changed. This aspect fits well with the different

ISBN: 978-88-5495-135-8

entrepreneurial cultures that characterize these regions.



Source: Our processing on Istat data (Istat, 2017a b c)

Figure 1: Barycentre and Standard Deviation Ellipse, PDO-PGIs density, farm-holidays density and nights spent density.

The second step is the application of the K function of Ripley (Reply, 1976; Diggle, 1983)

$$K(h) = \frac{E[N_0(h)]}{\lambda}$$

We can identify the number of the events N0 (h) in the H circle around a centre that is our "relevant event". In that case they are: the centres of gravity, obtained weighting the municipality with the number of the farm-holidays, the PDO-PGI products and the nights-spent (λ is the intensity of the process).

A homogeneous Poisson process has been used:

$$K(h) = \pi h^2$$

where

$$E(N_0(h)) = \lambda \pi h^2$$

are the expected values of the points inside the spatial circle. The Ripley transformation has been used for the distance (Boots and Getis,1988).

There is aggregation when

$$K(h) > \pi h^2$$

The third step is the application of non-parametric interpolation by Kernel Density Estimation (KDE) (Chainey et al.2002) to design new geographical areas not coincident with the administrative boundaries

$$\lambda_s^* = \sum_{i=1}^{n} \frac{1}{\tau^2} g\left(\frac{s - s_i}{\tau}\right)$$

where:

λ^*_s is the estimate of the intensity of the point-event distribution observed in the locality s;

S_i defines the points-event;

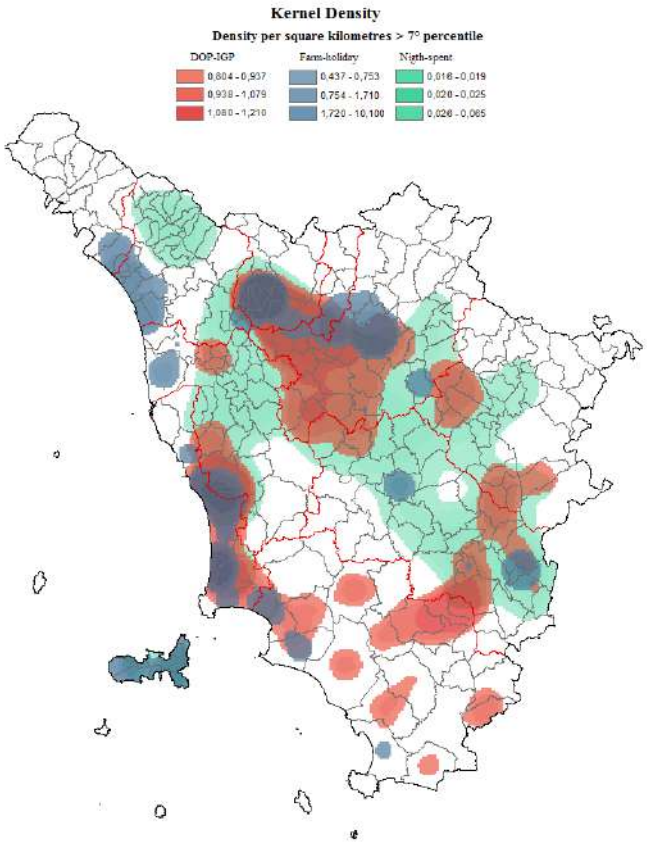
g(.) represents the three mobile functions (in our case we have a normal distribution as in Levine, 2004) of Kernel;

τ is the parameter controlling the smoothing effect and that takes into account the spatial variability of observations in the region.

3. Conclusion

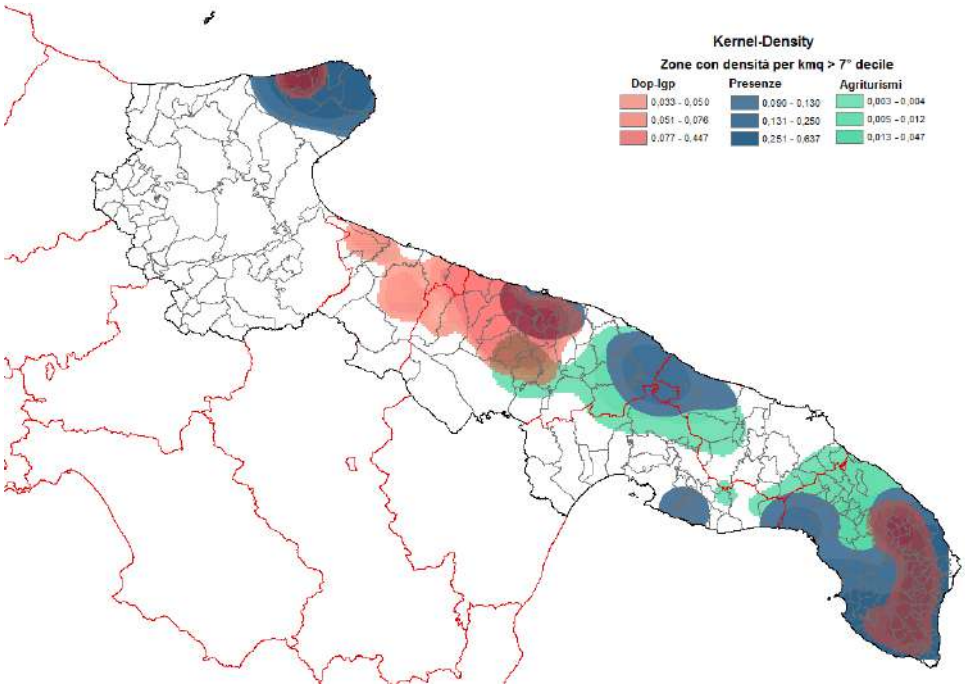
The analysis has carried out two different results. Figure 1 shows the centre of gravity for farms, PDO-PGI producers, and the nights-spent by tourists. By comparing the situation both in Tuscany and Apulia, we can observe that the barycentre of farms and PDO-PGI producers are closer to Tuscany, but farther from Apulia. It means that in Tuscany the farms have enlarged their services, offering quality on wine and food. It has also implied a change in the organizational structure of farms and the beginning of a modernisation process. We cannot find the same situation in Apulia, region in which the farms have maintained their original organizational structure.

Whereas, Figures 2 and 3 show the FFT areas individualized by a non-parametric interpolation (KDE), with a relevant difference between Tuscany and Apulia. Tuscany seems to have some FFT zones in the inner areas and on the coast. In Apulia, this situation does not seem to be so concentrated in the places; in fact, some areas are characterized by an important presence of tourists in the farmhouses, perhaps without producing PDO-PGI goods.



Source: Our processing on Istat data, (Istat, 2017a b c).

Figure 2: FFT areas in Tuscany.



Source: Our processing on Istat data, (Istat, 2017a b c).

Figure 3: FFT areas in Apulia.

References

Boyne, S., Hall, D., Fiona, W. (2003). Policy, Support and Promotion for Food-Related Tourism Initiatives. *Journal Travel & tourism marketing*, 14(3-4), pp. 131-154.

Boots, B., Getis, A. (1988). Point Pattern Analysis. *Sage University Paper Series on Quantitative Applications in the Social Sciences***,** series no. 07–001. Sage Publications.

Chainey, S, Reid, S., Stuart, N. (2002). When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime. *Innovations in GIS* 9, pp. 21-36.

Diggle, P.J. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London.

Istat (2017a) *Indagine sulle aziende agrituristiche*. Roma. http://dati.istat.it.

Istat (2017b) *Indagine sui prodotti agroalimentari di qualità DOP-IGP e STG*. Roma. http://dati.istat.it.

Istat (2017c) *Movimento negli esercizi ricettivi*. Roma. http://dati.istat.it.

Ripley, B.D. (1976). The Second-Order Analysis of Stationary Point Processes. *Journal of Applied Probability* **13**(2), pp. 255–266.

Levine, N. (2004). *CrimeStat III: a spatial statistic program for the analysis of crime incident locations*. National Institute of Justice, Washington DC.

Issues in prior achievement adjustment for value added analysis: an application to Invalsi tests in Italian schools.

Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto, Carla Rampichini
Department of Statistics, Computer Science, Applications "G. Parenti"
University of Florence

1. Introduction

Our work focuses on the estimation of the effect of student- and school-level characteristics on student achievement when a pre-test is available. More specifically, using Invalsi data (Martini, 2018), our focus is on the value added of the lower secondary schools on test scores at the 8th grade (post-test), accounting for the student test scores at the 5th grade (pre-test). We control for the available student and school characteristics using multilevel models (Goldstein, 2010).

Two main methodological approaches have been considered in the literature to deal with the estimation of a causal effect when a pre-test measure of the outcome is available (Penaloza and Berends, 2019; Kim and Steiner, 2019). The first approach consists in estimating the effect of the variable of interest on the test score, conditionally on the pre-test score (conditioning approach). In the second approach, the analysis is conducted on the gain score, namely the difference between the post-test and the pre-test scores (gain score approach). In the causal inference literature, the conditioning approach is implemented via regression models or matching on the pre-test score, which can be regarded as methods to remove confounding, when conditioning on the pre-test score is sufficient to make the unconfoundedness assumption plausible (Arpino and Aaassve, 2013). On the other hand, the gain score approach is related to difference-in-differences methods, which are devised to remove the effect of unobservable confounders under the assumption that such confounders have a time invariant effect, known as common trend assumption (Kim and Steiner, 2019). In such a case, taking the first difference of the outcome removes confounding (e.g. Lechner, 2011).

Recently, Kim and Steiner (2019) reconsidered the choice between the conditioning and gain score approaches. They consider a linear data generating model with constant effects across units. The treatment variable Z affects the post-test score Y, while an unobservable confounder, representing ability, A affects both Z and Y. In addition, the ability A affects the pre-test score P. If P is a reliable measure of A, conditioning on P removes most of the confounding effect of A. On the other hand, a low pre-test reliability suggests to consider the gain score approach, which is not affected by measurement error. However, the gain score approach is based on the common trend assumption. The authors derive formulas for the bias of the causal effects estimators under the two approaches, highlighting the assumptions required for unbiasedness. They also consider other scenarios, in particular a direct effect of the pre-test score on the treatment variable, which makes more problematic the assessment of the bias under the gain score approach (Allison, 1990).

In this paper, we aim at empirically comparing the two approaches, based on conditioning and gain scores, using multilevel models on Invalsi data. In this way we account for the hierarchical structure of Invalsi data with students nested into schools, while the aforementioned recent studies comparing conditioning and gain score approaches have focused on unstructured data.

ISBN: 978-88-5495-135-8

2. Invalsi data

Our data contains information on a cohort of students that participated in the Italian language Invalsi tests at grades 5th and 8th (i.e., the last year of the primary school and the last year of the lower secondary school, respectively). The data set has been obtained by merging data on students who attended the 5th grade in school year 2013-2014 with data on students who attended the 8th grade in school year 2016-2017. We retain data on students present in both occasions (about 90%).

The resulting data set consists of 427950 students whose tests are available in the data at both grades. The students are nested in 5777 Italian schools. The average number of tested students per school is 103.91 with a standard deviation of 54.97 (min $= 1$; max $= 334$).

Each of the two achievement tests (pre- and post-test) is composed of a set of items measuring the (unobservable) ability in language. Items are dichotomously scored, with value 1 for a correct answer and value 0 for a wrong answer. The selection of the set of items relies on internationally validated methods based on the Rasch model (Rasch, 1960). For this reason, the ability level of a student is measured by the raw score (i.e., the total number of correct answers to the test items). As the number of items is different across grades, we divide the raw scores by their maximum so that they are normalised in the range 0-100.

Several background variables are available both at student and school levels. Student covariates include gender, citizenship, and marks in language resulting from the school reports. Data also include information about parents' educational level and job condition, which are exploited by Invalsi to define an index of socio-economic status. In addition, a wide set of indicators measured at the end of the 5th grade provides information on student material deprivation, motivation and interest in learning, and relations with the class mates. School characteristics include information on the geographical location (municipality, urban area, altimetric area, and population density), the average number of students per class and the type of school (public vs private). Other school level variables are obtained averaging the student level characteristics (e.g., proportion of immigrants per school).

3. Preliminary results

We specify a linear multilevel model (Goldstein, 2010) with students at level 1 and schools at level 2. In order to compare the conditioning and the gain score approaches, we specify two versions of the model. In the first version, the response variable is the post-test score (8th grade test), while the pre-test score enters as a covariate. In the second version, the response variable is the gain score (difference between the 8th and 5th grade tests), while the pre-test score is omitted from the covariates. Both versions of the model include student and school independent variables as in a previous study by Invalsi (Martini, 2018). Among the student-level variables, we consider: gender (reference category: male), being immigrant of first generation and being immigrant of second generation (reference: Italian citizen), and socio-economic-cultural status (secs; continuous variable standardized in $[-3; +3]$, see Campodifiori et al. (2010)). In addition, we also account for some school-level characteristics, such as the type of school (reference: public school), the geographical macro-area where the school is located (reference: North West of Italy) and the cluster-means of first level variables: the school average pre-test score (CM_pre-test), the school proportion of females (CM_female), the school proportions of immigrants of first- and second- generation (CM_immI and CM_immII, respectively), the school average secs (CM_secs).

Table 1 displays the estimates for the two variants of conditional model and the gain score model. Estimated coefficients are all significant at 5%, with the exception of those in italic font.

Table 1: Model results

Variable	**Conditional model**	**Gain score model**
constant	48.55	48.59
Student-level covariates		
pre-test	0.53	–
female	3.74	2.58
immigrant I	-2.76	0.83
immigrant II	-1.97	0.30
secs	2.18	0.47
School-level covariates		
Cluster-means of student-level cov.		
CM_pre-test	-0.28	-0.74
CM_female	2.49	3.63
CM_immI	-10.77	-14.25
CM_immII	*1.38*	*-0.86*
CM_secs	2.12	3.85
Covariates defined at school-level		
private school	0.69	0.67
North East	-0.37	-0.37
Centre	-0.68	-0.68
South	-1.56	-1.53
Islands	-0.85	-0.81
Residual variances		
School level	4.15	4.00
Student level	12.30	14.24
LogL	-1675421.77	-1736763.45
n. parameters	22	21

Legend: not significant parameters in italic

Looking at results in Table 1, we first observe that the estimated coefficients of student-level covariates differ in the two models in the magnitude and often in the sign. In particular, we outline the differences in the sign of immigrant I and immigrant II. The difference between immigrants and natives is negative in the conditional model (estimated coefficient: -2.76) and positive in the gain score model (estimated coefficient: 0.83). This discrepancy between the estimated coefficients under the conditional and the gain score models can be due to the fact that the covariate X (e.g., immigrant I and immigrant II) acts on the pre-test score P, in addition to the post-test Y. In this case, the coefficient of X estimated in the gain score model is the difference between the direct effect on Y and the indirect effect on Y through P, thus the difference can be positive even if both the direct and indirect effects are negative.

As concerns the covariates measured at school-level a similar effect in the two models is observed. More precisely, estimated regression coefficients of covariates defined at school level (i.e., type of school and geographical macro-area) are the same, whereas those related with cluster-means of first level variables (i.e., CM_pre-test, CM_female, CM_immI, CM_immII, and CM_secs) differ in the magnitude but not in the sign. In detail, the contextual effect of secs (CM_secs) is positive, while those of the pre-test score (CM_pre-test) and the status of immigrant (CM_immI and CM_immII) are negative. Moreover, schools in the North West of Italy perform better than schools located somewhere else, and students attending a private lower secondary school have on average better results than students attending public schools. However, it is worth to note that the effect of private school is quite small compared to the effects of the other covariates. Indeed, it reduces from 4.08 (conditional model) and 3.50 (gain score model), when no control for other variables is introduced in the models (not shown here), to 0.69 and 0.67 (Table 1) after controlling for secs and the other covariates.

4. Future work

The preliminary results presented above suggest a certain caution in the interpretation of the covariate effects, in accordance with the hierarchical level of the covariate. This is particularly relevant when the aim of the study is the estimation of the causal effect of a treatment variable that may act at both student- and school-level.

For the future developments of the work, we intend to implement a Monte Carlo simulation study to evaluate the goodness of the conditioning and gain score approaches in estimating the treatment causal effect under different scenarios. We also aim at extending the study to take into account other student- and school-level characteristics, such as the student material deprivation and motivation in learning, as well as the school size and some information about the school location (e.g., level of urbanisation).

References

Allison, P.D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, **20**, pp. 93–114.

Arpino, B. and Aaassve, A. (2013). Estimating the causal effect of fertility on economic well-being: data requirements, identifying assumptions and estimation methods. *Empirical Economics*, **44**(1), pp. 355–385.

Campodifiori, E., Figura, E., Papini, M. and Ricci, R. (2010) *Un Indicatore di Status Socio-Economico-Culturale degli Allievi della Quinta Primaria in Italia* Working Paper n. 2, Invalsi.

Kim, Y. and Steiner, P. M. (2019). Gain scores revisited: a graphical models perspective. *Sociological Methods & Research*, DOI: 10.1177/0049124119826155.

Martini, A. (a cura di) (2018). *L'effetto scuola (valore aggiunto) nelle prove Invalsi 2018*. Invalsi.

Goldstein, H. (2010) *Multilevel Statistical Models, 4th ed.* Wiley.

Lechner, M.(2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, **4**(3), pp. 165–224.

Penaloza, R.V. and Berends, M. (2019). The Mechanics of Treatment-effect Estimate Bias for Nonexperimental Data. *Sociological Methods & Research*, DOI: 10.1177/0049124119852375.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Museum preferences analysis: an item response model applied to network data

Silvia Bacci, Bruno Bertaccini, Alessandra Petrucci
Department of Statistics, Computer Science, Applications "G. Parenti"
University of Florence, Florence, Italy

1. Introduction

Firenzecard is the pass of the municipality of Florence (IT) that offers the possibility to visit more than eighty museums and exhibitions all around Florence and its surrounding area in a period of three days (for details look at `http://firenzecard.it`). All the entrances in the museums by Firenzecard users are registered through digital devices, so that a huge amount of data about flows across museums and behavior of tourists is stored.

In this contribution we propose an analysis of the museum visits data collected in the year 2018. We first describe the network of Florentine museums using the indices that are adopted in the context of the (social) network analysis (Kolaczyk, 2009). Then, we propose a Latent Class Item Response (LC-IRT) model in order to identify homogenous classes of tourists that are distinguished for different museum preferences.

The novelty of the dataset together with a suitable statistical analysis allow us to provide the policy makers with useful information about the principal nodes of attraction of tourists, which can be used to improve the tourist services, also through the suggestion of alternative paths.

2. The network of Florentine museums

In the year 2018 the number of Firenzecards sold amounts to 127,092 corresponding to 884,389 visits to 40 different museums (out of more than 80 collections and exhibitions available in Florence and in the surrounding area). The number of entrances registered per card ranges between 1 and 31, with mean equal to 6.8 (median 6.0) and standard deviation 3.2; the 25% of users visits at most 4 museums, whereas the 75% of users visits at most 9 museums. In particular, the three most visited museums are the Galleria degli Uffizi (11.9% of visits in 2018), Galleria dell'Accademia (11.4%), and Opera del Duomo (10.7%).

The 40 museums visited with the Firenzecard define the nodes of a directed network whose edges represent the paths observed across museums (for details about definitions and instruments used in the network analysis see, among others, Kolaczyk, 2009). The amount of tourists that follow a certain path provides the weight of each edge. Looking at the increasing ordered distribution of the weighted edges, we observe that edges in the superior 10% account for the 79% of flows between pairs of nodes and those in the superior 25% account for the 94% of flows. Among these edges, there are the paths across Galleria degli Uffizi, Opera del Duomo, Galleria dell'Accademia, Palazzo Vecchio, and Pitti Boboli (along any direction), followed by Basilica di San Lorenzo and Cappelle Medicee.

Overall, the network is highly dense (density coefficient, defined as the ratio between observed edges and potential edges, is 0.849) and transitive (transitivity coefficient, which provides the proportion of two-stars that close in triangles, is 0.953). In addition, flows of tourists tend to move in a bidirectional way between museums as outlined by the reciprocity coefficient (i.e., ratio between mutual dyads and total number of dyads), which is equal to 0.905, and the

ISBN: 978-88-5495-135-8

share of complete triads (i.e., triples of nodes linked with bidirectional edges) that amounts to 62.7%.

Two common measures used in the network analysis to assess the importance of each node in terms of its capability of building incoming and outgoing links are given by the authority and hub scores. The authority score defines the importance of a node by how many nodes points to it, whereas the hub score defines the importance of a node by how many nodes it points to. Figure 1 (left) displays the authority and hub scores computed for the 40 Florentine museums. We observe that many museums have low scores on both the dimensions (range 0.00-0.20) and a

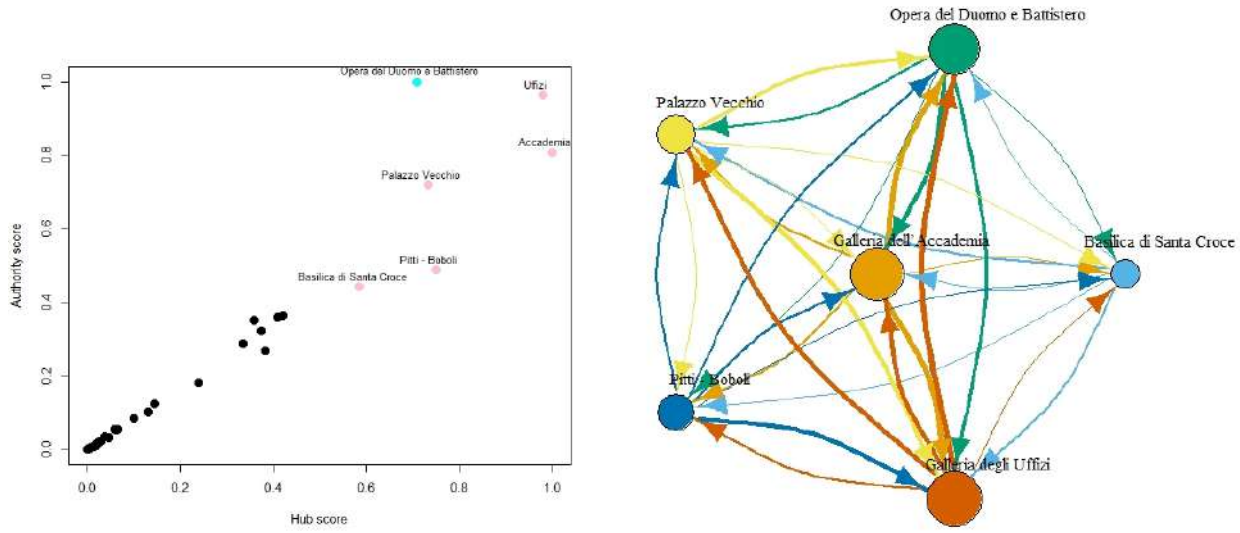


Figure 1: *Authority and hub scores (left) and network of museums with medium-high hub score (right).*

small sub-set of museums have intermediate scores (around 0.40). Only a few museums present high values (> 0.70) of authority and hub scores: Galleria degli Uffizi (authority = 0.979; hub = 0.964), Galleria dell'Accademia (authority = 0.808; hub = 1.000), Opera del Duomo (authority = 1.000; hub = 0.708), and Palazzo Vecchio (authority = 0.720; hub = 0.732); Pitti - Boboli and Basilica di Santa Croce follow with medium-high scores.

A synthesis of the above considerations is displayed in Figure 1 (right), limited to the network involving the six museums with the highest hub scores and the related links. The dimension of nodes reflects the number of visits per museum and the width of the edges refers to the amount of tourists that moved along the direction of the arrow.

3. The propensity to visit a museum: a latent class analysis

To characterize the typologies of Firenzecard users in terms of propensity to visit museums and identify museums whose level of popularity differs among tourists, we estimate a LC-IRT model (Bartolucci, 2007; von Davier, 2008; Bartolucci et al., 2016). We formulate a model for the vector of binary observed items $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ij}, \ldots, Y_{iJ})$, with Y_{ij} equal to 1 if tourist i visits museum j and 0 otherwise ($i = 1, \ldots, n$; $j = 1, \ldots, J$, $J = 40$). We also assume that the items are manifestations of the propensity to visit museums, which is represented through a discrete latent variable, Θ_i. Θ_i may assume a finite number of support points $\xi_1, \ldots, \xi_u, \ldots, \xi_k$ with probabilities $\pi_1, \ldots, \pi_u, \ldots, \pi_k$, respectively. Each support point represents a homogenous group (latent class) of tourists that share a similar behavior in terms of quantity and type of visited museums.

ISBN: 978-88-5495-135-8

The manifest distribution for the response vector is formulated as in the standard latent class model (Lazarsfeld and Henry, 1968; Goodman, 1974)

$$p(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \sum_{u=1}^{k} \prod_{j=1}^{J} p(Y_{ij} = y|\Theta_i = \xi_u)\pi_u,$$

with the conditional response distribution specified following the 2PL parameterization of Birnbaum (1968), that is,

$$logit\left[p(Y_{ij} = y|\Theta_i = \xi_u)\right] = \gamma_j(\xi_u - \beta_j) \quad i = 1, \ldots, n;\ j = 1, \ldots, J,$$

with γ_j and β_j item parameters that measure the level of attractiveness of each museum. Relying on the estimates of the model parameters and the individual observed item responses, the class membership posterior probability $p(\Theta_i = \xi_u|\boldsymbol{Y}_i = \boldsymbol{y}_i)$ is computed for each tourist, which is then allocated to the latent class having the maximum probability (maximum a posteriori criterion). Given the latent class, the conditional probability of visiting a museum, that is, $p(Y_{ij} = 1|\Theta_i = \xi_u)$, is predicted for each class.

On the basis of the Bayesian Information Criterion we select $k = 3$ latent classes of tourists. The three classes are ordered with respect to the propensity to visit a museum: the average probability of visit is 10% for Class 1, 18.4% for Class 2, and 28.9% for Class 3. The largest class is Class 2 ($\hat{\pi}_2 = 0.457$) followed by Class 1 ($\hat{\pi}_1 = 0.357$), whereas Class 3 collects the remaining 18.7% of tourists.

To provide a clearer idea about the differences in the tourist preferences, the two extreme classes 1 and 3 are compared in Figure 2 with respect to their conditional probabilities of visiting a museum, that is, $p(Y_{ij} = 1|\Theta_i = \xi_1)$ vs. $p(Y_{ij} = 1|\Theta_i = \xi_3)$. Museums whose probability of visit is above 80% for Class 3 are labelled in the figure.

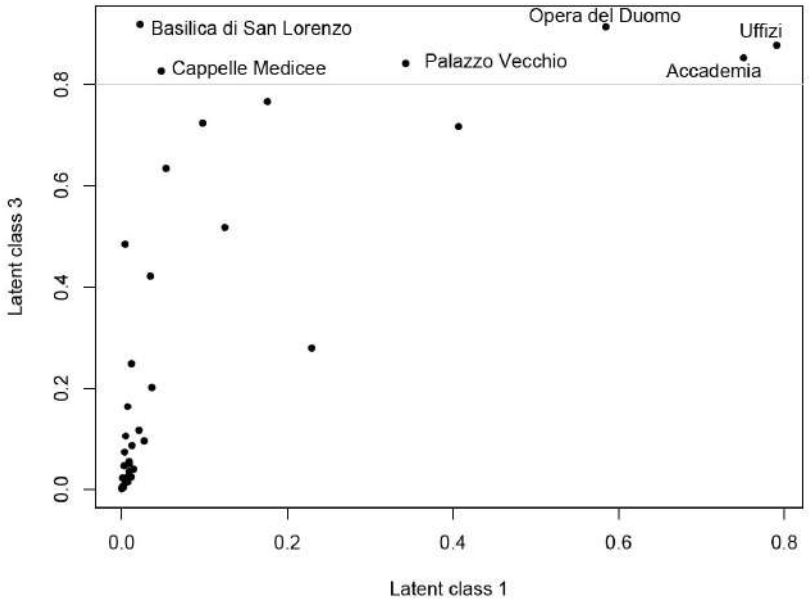


Figure 2: *Plot of* $p(Y_{ij} = 1|\Theta_i = \xi_u)$ *for* $u = 1, 3$.

On one side, we observe that Galleria degli Uffizi and Galleria dell'Accademia represent a sort of "must" for any type of tourist, as indicated by their high conditional probabilities. Opera del Duomo and Palazzo Vecchio follow with high conditional probabilities (above 80.0%) for

tourists in Class 3 and medium conditional probabilities (40.0-60.0%) for tourists in Class 1. Finally, it is interesting to observe the position of Basilica di San Lorenzo and Cappelle Medicee: these two museums are substantially ignored by tourists in Class 1 (conditional probabilities: 2.3% for Basilica di San Lorenzo and 4.8% for Cappelle Medicee), whereas they are at the top of preferences for tourists in Class 3 (conditional probabilities: 91.9% for Basilica di San Lorenzo and 82.7% for Cappelle Medicee).

4. Conclusions

The analysis of the network of Florentine museums originated by the Firenzecard records represents an important step in the comprehension of flows of tourists as well as their behavior and preferences. For the future development of this work we intend to characterize the classes of tourists in terms of individual characteristics (e.g., citizenship and age) in order to suggest paths of visit that are specifically tailored for the different typologies of tourists.

Authors thanks Linea Comune S.p.a. for making the Firenzecard data available.

References

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*. **72**: pp. 141–157.

Bartolucci, F., Bacci, S., Gnaldi, M. (2016). *Statistical analysis of questionnaires: a unified approach based on R and Stata*. Chapman & Hall, CRC Press, Boca Raton, (FL).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores*, eds. F.M. Lord and M.R. Novick, Addison-Wesley, Reading, (MA), pp. 395–479.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*. **61**: pp. 215–231.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data. Methods and Models*. Springer, New York, (NY).

Lazarsfeld, P. F., Henry, N. W.(1968). *Latent Structure Analysis*. Houghton Mifflin, Boston, (MA).

von Davier, M. (2008). A general diagnostic model applied to language testing data. *J. American Statist. Assoc.*. **81**: pp. 461–470.

Elderly with and without Children: Do They Report Different Health Conditions?

Chiara Bocci, Silvana Salvini
Department of Statistics, Computer Science, Applications "G. Parenti",
University of Florence, Florence, Italy

1. Introduction

The world is aging rather rapidly. As both the proportion of older people and the length of life increase throughout the world, key questions arise. Will population aging be accompanied by a longer period of good health, a sustained sense of well-being, and extended periods of social productivity, or are associated with more illness, disability, and dependency? And which is the relationship between childlessness and health, in a world where fertility does decline?

We intend to study the relationship between childlessness and perceived health for elderly in four European countries: France, The Netherlands, Poland and Italy. While Italy is a familistic society, the other three countries present few inter-generational exchanges and an early exit of children from the parental home. We have chosen to analyse these countries, which are characterised by different culture and welfare systems, to understand if the context is important to determine the relationship.

2. Health and Childlessness: The Relationship in the Literature

The influence of family behavior on health in old age has been increasingly recognised in literature. Apart from the physiological and psychological effects of pregnancy and childbirth, the health of both women and men may be influenced by stress, role changes, and changes in allocation of personal and family resources associated with child-rearing and by the emotional and social support benefits of parenthood.

Numerous studies in the past have considered the associations between fertility and mortality and different health indicators (Grundy and Read, 2015; Grundy and Tomassini, 2005). The literature on this topic suggests that there are several potential mechanisms that may cause different associations including selection into parenthood, direct biological factors, as well as indirect effects such as the relative costs and benefits of childrearing.

Many scholars have frequently speculated about why people have children and why some do not. A large number of costs and benefits in having children have been outlined by demographers and others (Ramu and Tavuchis, 1986). From the perspective of the elderly, some of these benefits include economic and social security, self-esteem gained from having acted in a normative way, health monitoring, companionship, achievement of a sense of continuity – however culturally defined – and pride, a family to be involved and feel needed in, the presence of grandchildren, and the potential and actuality of care. The burdens of having children may be fewer in number but are no less salient. For the elderly these may include such weights as continued unwanted involvement in childrens' lives, lack of independence, mental aggravations of various sorts, continuing financial demands, and an inability to transcend undesired aspects of the parental role.

Even though old childless individuals have social networks of less support potential than those who are parents, there are no differences in certain psychological wellbeing indicators

ISBN: 978-88-5495-135-8

Table 1: Percentage of older people reporting themselves childless, around year 2000.

Country	Men aged 65+ EHCP	Women aged 65+ EHCP	Women 65 – 70 EHCP	Women born in 1930 Official
Denmark	12%	13%	10%	–
Belgium	23%	24%	23%	17%
Luxembourg	25%	25%	27%	–
France	15%	17%	15%	13%
United Kingdom	20%	21%	16%	14%
Ireland	32%	27%	21%	–
Italy	35%	38%	40%	16%
Greece	13%	19%	18%	–
Spain	12%	16%	14%	15%
Portugal	33%	30%	27%	–

Source: data extracted from Iacovou (2000).

between the two groups. Apparently, childless old people find ways to cope with whatever negative effects of childlessness they may have experienced (Vikström et al., 2011).

3. Data and Methods

To examine France, The Netherlands and Poland we use data from the Generations and Gender Surveys (GGS) (for the years of 2005, 2002-2004 and 2010-2011, respectively), while for Italy we use the Italian Multipurpose Survey carried out in 2009, which focused on Family and Fertility (FFS). We are aware that the dataset are collected in different years, and that the comparison between countries could be affected by this fact. However, all data refer to a five years period and we can imagine that the theme under study is persistent enough in such a short time frame because the argument inherent to children concerns the context during time and is not influenced by conjuncture topics.

The core questionnaire of GGS contains over 1,000 questions or items, broadly classified as follows: parent-child relationships; parent's perspective; child's perspective; relationships between partners; partnership formation and dissolution; gender perspective; complex partnership and fertility histories, stepfamilies; contraception and infertility treatment; household; housing; economic activity, income and wealth; education; health; personal networks; welfare state; subjective well-being and values. In the FFS of 2009 the items are relative to household and focus on housing, economic activity, partnership formation and dissolution, complex partnership and fertility histories, contraception, education, occupation and values.

Even if the both GGS and FFS collect information on the whole adult population, we focus the analysis on people aged 50 and over, that, in the samples, are 18,033 in Italy, 3,331 in The Netherlands, 10,402 in Poland and 4,446 in France.

To investigate the possible relation between the perceived health status of old people and the number of children (controlling for age, sex, marital status, education, work status, region and ownership of household) we apply a multinomial logistic regression. The dependent variable is the perceived health status (classified in *Good*, *Fair* and *Bad*) and the covariate of interest is the number of children, treated as categorical, classified in 0, 1, 2, 3 and 4+.

In addition, we control for the possible confounding effects of the following explanatory variables:

- Age in classes ($50-64$, $65-74$, 75+ years), because we know that as age increases,

health deteriorates;

- Sex (*Male*, *Female*), since there are strong differences in health status and diseases between women and men and the relationship with presence of children may be differently influential;
- Marital status (*Single*, *Married*, *Divorced*, *Widowed*), because its relationship with health status is due to selection and causal effects;
- Education level (*High*, *Medium* and *Low*), because it influences health status directly and indirectly, through the different recourse to preventive medicine and use of health services;
- Work status (classified in *Ever worked* and *Never worked*), for the association and selection effect with health and children;
- NUTS1 regions, to control for the geographical context;
- Ownership of household (*Yes*, *No*), as a proxy of the economic status.

Finally, for Italy we include also the subjective evaluation of economic status (*Good*, *Not good*), not available for France, The Netherlands, and Poland.

4. Results

Generally, in every country we examined, we did not observe a significant association between the number of children and the perceived health of elderly (Figure 1), whereas covariates are often linked with health by a strong relationship (for example, in Figure 2 we show the

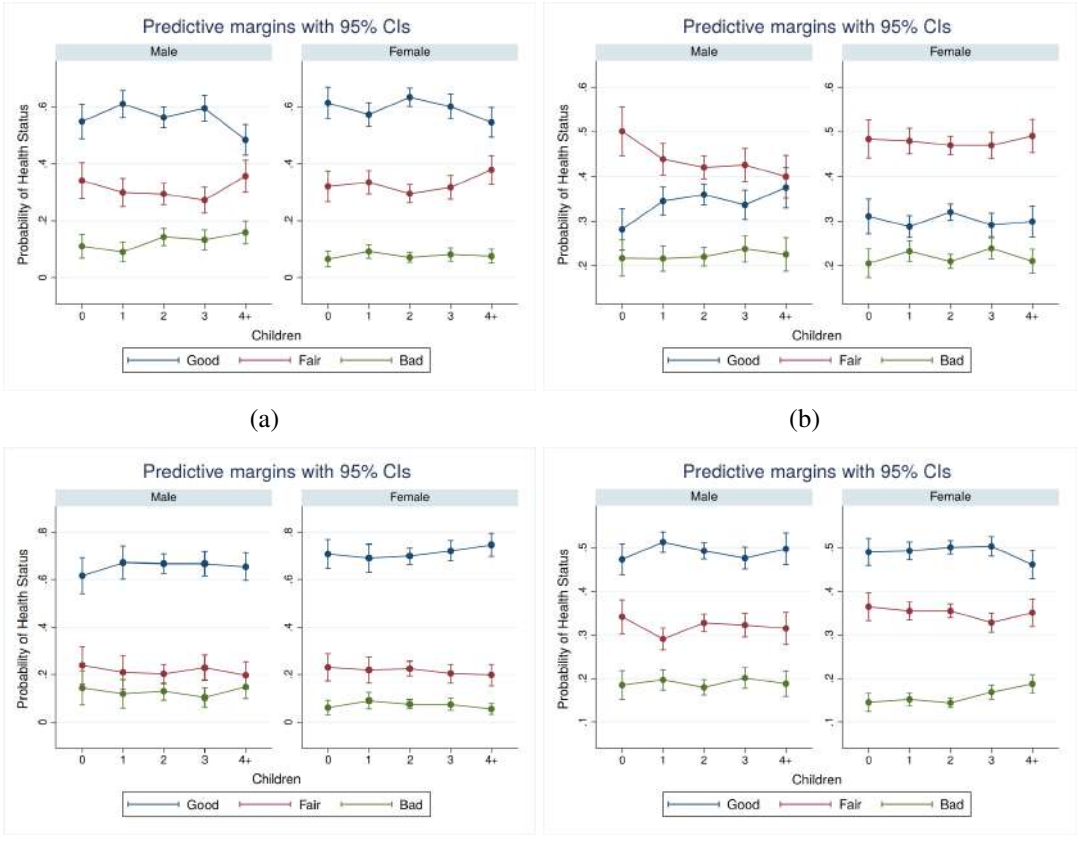


Figure 1: Predicted probability of the perceived health status according to the number of children: (a) France; (b) Poland; (c) The Netherlands; (d) Italy.

ISBN: 978-88-5495-135-8

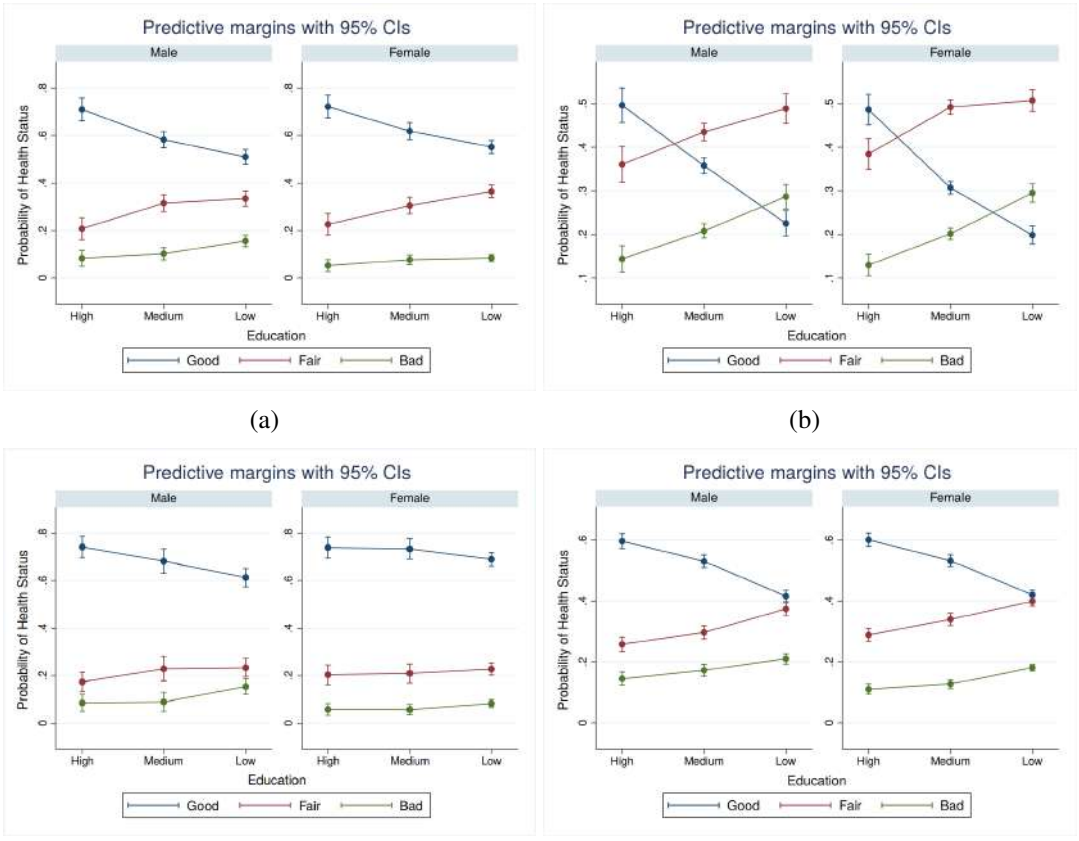


Figure 2: Predicted probability of the perceived health status according to the educational level: (a) France; (b) Poland; (c) The Netherlands; (d) Italy.

predicted probabilities of each health status with respect to the education level). In summary, and this is the result we outline, children do not have any impact on health, but for a weak positive effect of having 1 child in Italy and Poland (a lower probability of *Fair* vs *Good* health). Therefore, people with or without children do not seem to perceive a different health status.

References

Grundy, E., Read, S. (2015). Pathways from fertility history to later life health: Results from analyses of the English Longitudinal Study of Ageing. *Demographic Research*, **32**, pp. 107–146.

Grundy, E., Tomassini, C. (2005). Fertility history and health in later life: a record linkage study in England and Wales. *Social Science & Medicine*, **61**(1), pp. 217–228.

Iacovou, M. (2000). Health, wealth and progeny: explaining the living arrangements of older European women. *ISER Working Paper Series*, **2000-08**, Institute for Social and Economic Research.

Ramu, G. N., Tavuchis, N. (1986). The valuation of children and parenthood among the voluntarily childless and parental couples in Canada. *Journal of Comparative Family Studies*, **17**(1), pp. 99–116.

Vikström, J., Bladh, M., Hammar, M., Marcusson, J., Wressle, E., Sydsjö G. (2011). The influences of childlessness on the psychological well-being and social network of the oldest old. *BMC Geriatrics*, **11**.

A multi-inflated hurdle regression model for the total number of overnight stays of Italian tourists in the years of the economic recession

Chiara Bocci, Laura Grassini, Emilia Rocco
Department of Statistics, Computer Science, Applications "G. Parenti",
University of Florence, Florence, Italy

1. Introduction

During the years 2008-2013, consumption expenditures of Italian households was harshly hit by the world economic recession (ISTAT, 2014) with a remarkable reduction of purchasing power (-10.4% between 2007 and 2013). In that period, Italian households showed a reduction in tourism expenditure and a change in travel behaviour, as well. In particular, the expenditure share devoted to accommodation facilities passed from 2.8% in 2010 to 2.3% in 2013 and the annual decrease in the number of trips by resident was nearly -12% in 2010, -19% in 2013. Only in 2015, for the first time after seven years, there has been a remarkable increase (+13.5%).

Therefore, since tourism is an important driver of the economic development, the study and modelling of tourism demand of Italian residents becomes of extreme importance for knowing the determinants of household tourism behaviour and the impact of the economic crisis.

This work is concerned with individual tourism behaviour and, specifically, it investigates the participation in tourism of the Italian residents in a period covering the recent economic recession: whether or not they have travelled for vacation and they have reduced the length of stay (number of nights spent) over the period of analysis. Data on household and individual travel behaviour are derived from the survey on *Trips of Italian Residents in Italy and Abroad*, currently carried out by the National Statistics Office (ISTAT) for responding at the EU Reg.692/2011.

The theoretical framework for our analysis of overnight stays (OS) is the hurdle model, a modified count model which allows to consider the response as the results of a decision process in two steps: firstly a person decides whether to have a vacation trip and then, conditionally to a positive decision, he decides the number of OS. In a general hurdle model, at first a binary model is used to model the binary outcome of whether a count variable has a zero or a positive realization and then the positive realizations are modelled by a truncated-at-zero count data model. In our analysis, in order to account for the overdispersion of the OS and their concentration on some specific values, at the second step we employed a multi-inflated cumulative logit Negative Binomial model (Cai et al., 2018).

2. Literature review

Concerning the effects of economic crisis on tourist behaviour, we should consider that tourism has become a "normal thing" for some population (Bargeman and van der Poel, 2006); it is part of the lifestyle, quality of life and well being of an increasing number of people (Cracolici et al., 2013; Dolnicar et al., 2012). In addition, it is argued that many tourists follow a travelling career to fulfil their travel needs: they start to visit destination for basic-level needs passing later to higher-level needs (Ryan, 1998). Thus, we likely observe an inertia in tourism behaviour and a higher probability of a "slicing strategy" (e.g., cheaper holiday) rather than a "cutback or pruning strategy" (e.g, fewer trips, reduced length of stay). From these considerations and the investigations on the effects of economic crisis on tourist behaviour carried out by Bronner

and De Hoog (2012), Campos-Soria et al. (2015) and Wong et al. (2017) among others we can derive the following indications for our analysis: (i) to use an hurdle model, since tourist choice is a multi-stage decision process; (ii) behaviour determinants belong to three categories: socio-demographic, economic and trip related variables, and we must also include time and seasonal variables (with intra-annual data) and interaction terms; (iii) the assessment of specific strategies during an economic crisis can be addressed only with primary panel data, while from cross-section secondary data we can assess the effects of covariates on different levels of pruning strategy: the non-participation in tourism, reduction of the number of trips, of the length of stay, etc...

3. Methodology

Given the data characteristics described in Section 1 and the considerations derived in Section 2, this study examines whether and how the economic recession has affected the total number of OS in a quarter trough an hurdle model. Exploratory analysis of the data and common knowledge of the phenomenon show that it is restrictive to assume that the conditional mean and variance of the number of OS are equal; moreover, the number of OS is naturally concentrated on some values (like 2, 6, 7, 14, 20 nights). Thus, to handle the overdispersion (the mean of the positive number of OS is 9 whereas its variance is 86.6) and the multi-inflation of our data (as evident in Figure 1(b)), we model the multi-modal discrete distribution of the non-zero number of OS through a cumulative logit zero-truncated Negative Binomial regression model.

Formally, let $\boldsymbol{y}$ be the quarterly number of OS, and $\mathbf{X}$ and $\mathbf{Z}$ the covariates matrices included in the first and the second model respectively. Suppose that $\boldsymbol{y}$ contains a total of $M-1$ inflated values and, while these inflated values do not have to be consecutive in the model, for notational convenience we denote them as $\{1, ..., (M-1)\}$. Then, our hurdle model is given by:

- I stage: a *logit model* for the tourism participation

$$P(y_i = 0|\mathbf{X}_i) = exp(\mathbf{X}_i'\boldsymbol{\beta}_1)/(1 + exp(\mathbf{X}_i'\boldsymbol{\beta}_1))$$

- II stage: a *multi-inflated cumulative logit zero-truncated Negative Binomial model* for the number of positive OS

$$P(y_i = j|y_i > 0, \mathbf{Z}_i) = \begin{cases} p_{ij} + p_{iM}\dfrac{f_{NB}(j)}{[1 - f_{NB}(0)]} & for j = 1, ..., (M-1) \\ p_{iM}\dfrac{f_{NB}(j)}{[1 - f_{NB}(0)]} & for j \geq M \end{cases}$$

  in which $\sum_{j=1}^{M} p_{ij} = 1$ and the p_{ij}'s are formulated with the following cumulative logit (or proportional odds) model (McCullag, 1980)

$$logit\left[Pr(y_i \leq j)\right] = \frac{Pr(y_i \leq j)}{Pr(y_i > j)} = \gamma_{j0}$$

  for $j = 1, ..., (M-1)$, where the γ_{j0}'s are $M-1$ intercepts; and

$$f_{NB}(j) = P(y_i = j|\mathbf{Z}_i) = \frac{\Gamma(\lambda_i + y_i)}{\Gamma(\lambda_i)\Gamma(y_i + 1)}\left(\frac{\theta}{1+\theta}\right)^{y_i}\left(\frac{1}{1+\theta}\right)^{\lambda_i}$$

  denotes the probability mass distribution of the Negative Binomial model and in particular the well known NB2 model (Cameron and Trivedi, 2013) that has a quadratic variance function $\lambda_i(1 + \lambda_i/\theta)$. Γ denotes the gamma function, the parameter θ is assumed to be constant while λ_i depends on covariates by the function $ln(\lambda_i) = \mathbf{Z}_i'\boldsymbol{\beta}_2$.

The covariates of the logit model ($\mathbf{X}$) include variables 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, percentage of family income recipients included retired members). The covariates of the conditional model ($\mathbf{Z}$) include also trips-related variables (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. Finally, in the first model we have included as a specific interaction term the percentage of family income recipients per year, whereas in the second model we allow for different covariates effects for those who only take long vacancies (more than 3 nights at a time) than for the others. For this reason, all covariates in the second model are interacted with the dummy indicator *at least one short vacation*.

4. Participation in tourism of Italian residents in the period 2000-2013

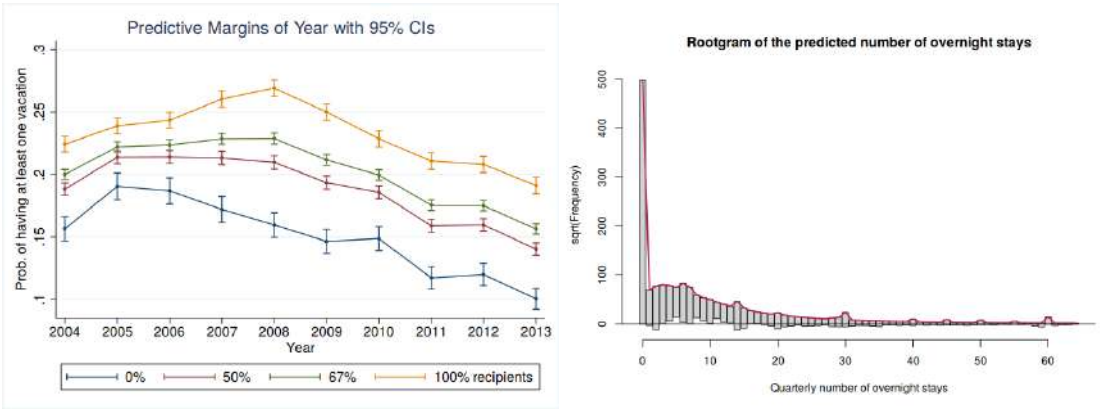
The household survey *Trips and Holidays of Italian Residents in Italy and Abroad* collects information about domestic and outbound travels of Italian residents. From 1997 to 2013, it has been carried out quarterly on a national annual sample of about 14,000 households. It offered an in-depth insight about the individual participation in tourism in terms of number of trips, nights spent and characteristics of the trip, but gave no information about tourist expenditure. From 2014 it has been associated with the *Consumer Expenditure Survey*, but, due to different sample designs, the two sources cannot be appropriately linked together for our aims. In addition, given the adoption of the Euro currency in 2002, we limit our analysis to the years 2004–2013.

First of all, our results, presented only partially in Table 1 and in Figure 1 due to lack of space, show that the economic crisis had a negative impact on both the tourism participation and the length of stay, particularly for those tourists that take only long vacations. Moreover, they confirm common knowledge that some types of trips strongly determine their length, that sea-

Table 1: Estimated parameters of the Truncated Negative Binomial regression model

Covariate	Coef.	Covariate	Coef.	Covariate	Coef.	Covariate	Coef.
Scaled age	0.122^{***}	(Scaled age)^2	0.040^{***}	Female	-0.006	Household size	-0.034^{***}
× short vac.	-0.059^{***}	× short vac.	-0.035^{**}	× short vac.	0.030^{*}	× short vac.	0.007
# of children	0.071^{***}	Univ. degree	0.046^{***}	Business trips	-0.054^{**}	OCC:housewife	0.048°
× short vac.	0.032^{**}	× short vac.	0.028°	× short vac.	0.028°	× short vac.	-0.057
OCC:student	0.031	OCC:retired	0.045	OCC:unable	0.029	OCC:manag.staff	-0.028
× short vac.	0.114^{**}	× short vac.	0.022	× short vac.	-0.01	× short vac.	0.083°
OCC:office work	-0.058^{*}	OCC:manu.work	-0.089^{**}	OCC:self-empl.	-0.066^{*}	OCC:professional	-0.048
× short vac.	0.086^{*}	× short vac.	-0.039	× short vac.	0.017	× short vac.	0.128^{**}
NUTS1:northeast	-0.079^{***}	NUTS1:centre	-0.063^{***}	NUTS1:south	-0.109^{***}	NUTS1:islands	-0.122^{***}
× short vac.	-0.015	× short vac.	0.009	× short vac.	0.007	× short vac.	-0.021
Quarter 2	-0.016	Quarter 3	0.414^{***}	Quarter 4	-0.094^{***}	2005	-0.055^{**}
× short vac.	-0.007	× short vac.	0.199^{***}	× short vac.	0.059^{**}	× short vac.	0.028
2006	-0.030^{*}	2007	-0.068^{***}	2008	-0.083^{***}	2009	-0.119^{***}
× short vac.	0.049^{*}	× short vac.	0.029	× short vac.	0.061^{*}	× short vac.	0.080^{**}
2010	-0.095^{***}	2011	-0.104^{***}	2012	-0.104^{***}	2013	-0.149^{***}
× short vac.	0.046°	× short vac.	-0.008	× short vac.	-0.052°	× short vac.	0.001
% income recip.	-0.049^{**}	# family visits	-0.080^{***}	# beach trips	0.105^{***}	#mountain trips	-0.049^{**}
× short vac.	0.066^{*}	× short vac.	0.151^{***}	× short vac.	0.030°	× short vac.	0.132^{***}
# art towns trips	-0.178^{***}	# tours	-0.015	other holidays[a]	-0.007	# free accom.	0.282^{***}
× short vac.	0.087^{***}	× short vac.	0.074^{**}	× short vac.	0.267^{***}	× short vac.	-0.328^{***}
total trips	0.330^{***}	Abroad	0.075^{***}	short vacation	-1.667^{***}	intercept	1.880^{***}
× short vac.	0.164^{***}	× short vac.	0.310^{***}	θ	4.084^{***}		

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)

Figure 1: Results: (a) Predictive margins of *Year* and of *% of income recipients* on tourism participation; (b) Comparison of predicted and observed values of the quarterly number of OS.

sonality is a universal factor in tourism and that the socio-demographic variables are relevant in determining the tourism behaviour. Focusing on the household economic condition (indirectly measured by the percentage of family income recipients), from Figure 1(a) it is possible to note that there is a positive association with the decision to travel and its effect is more pronounced in the period of economic crisis. On the other hand, it impacts negatively on the quarterly number of OS probably due to the possible time constraints deriving from the work activity. The hanging rootogram, Figure 1(b), compares predicted and observed values and indicates an overall good fitting of the estimated hurdle model and shows the multi-inflation of the OS variable.

References

Bargeman, B., van der Poel, H. (2006). The role of routines in the vacation decision-making process of Dutch vacationers. *Tourism Management*, **27**, pp. 707–720.

Bronner, F. De Hoog, R. (2012). Economizing strategies during an economic crisis. *Annals of Tourism Research*, **39**, pp. 1048–1069.

Cai, T., Xia, Y., Zhou, Y. (2018). Generalized inflated discrete models: A strategy to work with multimodal discrete distributions. *Sociological Methods & Research*, pp. 1–36.

Cameron, A.C., Trivedi, P.K. (2013). *Regression Analysis of Count Data*. Cambridge University Press, New York.

Campos-Soria, J.A., Inchausti-Sintes, F., Eugenio-Martin, J.L. (2015). Understanding tourists' economizing strategies during the global economic crisis. *Tourism Management*, **48**, pp. 164–173.

Cracolici, M., Giambona, F., Cuaro, M. (2013). Family structure and subjective economic well-being: some new evidence. *Social Indicators Research*, **118**, pp. 433–456.

Dolnicar, S., Yanamandram, V., Cli, K. (2012). The contribution of vacations to quality of life. *Annals of Tourism Research*, **39**, pp. 59–83.

ISTAT (2014). *Rapporto annuale 2014. La situazione del Paese*. ISTAT, Rome.

McCullag, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B*, **42**, pp. 109–142.

Ryan, C. (1998). The travel career ladder: an appraisal. *Annals of Tourism Research*, **25**, pp. 936–957.

Wong, I.K.A., Law, R., Zhao, X.R. (2017). When and where to travel? A longitudinal multilevel investigation on destination choice and demand. *Journal of Travel Research*, **56**, pp. 868–880.

Family lifestyle habits: what is passed down from adults to children?

Riccardo Borgia, Elena Castellari, Paolo Sckokai [a]
[a] Dipartimento di Economia Agro-alimentare, Università Cattolica del Sacro Cuore, Piacenza, Italy

1. Introduction

Noncommunicable diseases are responsible for 70% of deaths worldwide, primarily induced by the use of tobacco, unhealthy dietary habits, lack of physical activity (PA) and abuse of alcohol (WHO, 2017). Among these behaviors, the adoption of healthy eating patterns and a regular physical activity initiated in childhood may reduce the risk of occurrence of overweight and obesity, immediately, and the insurgence of chronic diseases, into adulthood (Nicklas et al., 2001). To promote the early adoption of such healthy practices family has always played a key role (Tinsley, 2003; Patrick and Nicklas, 2005). Nevertheless, the children's daily life environment is deeply and rapidly changing and the central role of the family in shaping children's lifestyle habits is being called into question (Crockett and Sims, 1995; Story, Neumark-Sztainer and French, 2002).

2. Data and research methodology

The study focuses on four behaviors indicated as highly – negatively and positively – related to the occurrence of overweight and obesity. The advisable behaviors are i) eating food for breakfast, ii) doing regular PA and iii) consuming five portions of fruit and vegetable (FV) per day. The unadvisable behavior is consuming one or more savory snacks a day (Table 1).

Table 1: Observed children behaviors and summary statistics.

Variables	Classes	Description	Relative frequencies (%)
Eating breakfast	1	Eating food for breakfast (with or without drinking)	81.44
	0	Not eating: breakfast skipping or just drinking	18.56
PA	1	Doing regular physical activity	49.36
	0	Otherwise	50.64
Snacking	1	Daily consumption of savory snack (potato chips, popcorn, etc.)	13.61
	0	Savory snack consumption lower than once a day or null	86.39
5 FV day	1	Consumption of the 5 daily-recommended portions of FV [a]	5.74
	0	Consumption lower than 5 portions of FV a day	94.20

[a] According to FAO/WHO (2003) one portion is defined as 80 g of fruit or vegetable.

Each habit is studied individually but by means of the same model (i.e. including the same regressors). The model employs a multinomial logistic (MNL) regression to relate the occurrence of the four investigated behaviors to the characteristics of the children (age and gender) and of the respective household. Specifically the latter outline health and socio-demographic features of the household: share of obese and overweight adults, household head's educational level, geographical area, and number of children. Moreover, to assess the household behavioral environment – main focus of the research – the share of adults manifesting the investigated children behaviors is included in the model. Lastly, the year of data collection is also observed – even if the study is cross-sectional – to monitor any significant variation from 2013 to 2016 (Table 2). The study is performed on the microdata of the *Italian Multipurpose Survey on Households Daily Life Aspects* provided by ISTAT (years 2013, 2014, 2015, 2016). The dataset consists of 25,265 children belonging to 16,893 households. Children are identified in the dataset as individuals younger than 18 years old.

ISBN: 978-88-5495-135-8

Table 2: Predictor variables and summary statistics.

Variables	Classes	Relative frequencies (%)		
Age Children age	3 - 5 y	19.24		
	6 - 10 y	33.42		
	11 - 13 y	19.97		
	14 - 17 y	27.37		
Gender Children gender	male	51.42		
	female	48.58		
BMI Overweight or obese adults within the household [a]	no one	32.84		
	at least one	48.82		
	all	18.34		
Education Household head's educational level	elementary school license or lower	7.03		
	secondary school license	35.93		
	high school diploma	41.50		
	university degree or higher	15.54		
Geographical Geographical distribution	Northern Italy [b]	42.27		
	Central Italy [b]	16.23		
	Southern and Insular Italy [b]	41.49		
Children Number of children within the household	1	32.86		
	2	50.09		
	3 or more	17.05		
Adults behavior Adults in the household manifesting the investigated behavior [c]		*Eating breakfast*	*PA*	*Snacking*
	no one	13.65	67.12	92.95
	at least one	31.42	24.4	5.92
	all	54.93	8.47	1.13
Adults FV Household mean number of FV portions daily-consumed [c]	2	42.77		
	3	34.32		
	4	17.86		
	5 or more	5.06		
Year Year of the survey	2013	25.96		
	2014	25.23		
	2015	25.08		
	2016	23.72		

[a] According to WHO a person with a *Body Mass Index* from 25 to 30 is considered overweight, obese from 30 or higher.
[b] Northern Italy: Piemonte, Valle d'Aosta, Liguria, Lombardia, Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna. Central Italy: Toscana, Umbria, Marche, Lazio. Southern and Insular Italy: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna.
[c] Only in the model for studying the children FV consumption the variable *adults behavior* has been substituted with the variable *adults FV*, to better observe the relation between children and adults' behavior.

3. Results

The study shows that the likelihood of observing the investigated behaviors significantly varies during the children's growth. Indeed, data suggest that the consumption of the daily-recommended portions of FV and daily-snacking increase with age, while the habit of eating for breakfast decreases. The practice of PA, finally, is significantly more observed in children from 6 years old, peaking between 11-13. The habit most gender-driven is the practice of regular PA, significantly more present in males. However, data show that females are less prone to snack and more inclined to consume the daily-recommended portions of FV (Table 3).

Focusing on the characteristics of the household, the analysis shows that the presence of overweight and obese adults is not significantly related to the insurgence of the investigated behaviors except for breakfast-eating. This habit, indeed, is less observed in children living in households of only overweight or obese adults. The educational level of the household head appears instead significantly associated with the children's practice of PA and daily-snacking: positively with the former, negatively with the latter. Breakfast-eating is instead more manifested in families with a graduated head of the family. The achievement of the daily-recommended number of portions of FV, finally, does not look significantly related to the education of the head of the family. The geographical distribution of the household demonstrates to be a key predictor as well. The practice of PA and the consumption of the recommended portions of FV are indeed less observed among children from Southern and Insular Italy, while daily-snacking is more present. The study does not show any statistically significant difference in the behavior of only children. If the children in the household are more than two, instead, they seem to be less inclined to the practice of PA and more prone to daily-snacking (Table 3).

Table 3: Discrete marginal effect of the predictors expressed as percentage change [a].

		Children behaviors			
Variables	**Classes**	*Eating breakfast*	*PA*	*Snacking*	*5 FV day*
Age *referent class: 3 - 5 y*	6 - 10 y	-1.41**	35.19***	3.82***	1.39**
	11 - 13 y	-8.77***	37.24***	7.09***	0.82
	14 - 17 y	-15.56***	27.87***	7.05***	3.56***
Gender *referent class: male*	female	-3.22***	-8.55***	-1.37***	1.17**
BMI *referent class: no one*	at least one	0.15	-0.03	0.42	-0.30
	all	-1.40*	-0.37	0.55	-0.35
Education *referent class: elementary school license or lower*	secondary school license	-0.88	9.14***	-0.65	0.34
	high school diploma	-0.31	16.89***	-3.47***	0.50
	university degree or higher	2.07*	26.45***	-5.98***	0.59
Geographical *referent class: northern*	central	1.10	1.62*	-0.80	-1.34**
	southern and insular	-0.18	-13.82***	5.57***	-1.46**
Children number *referent class: 1*	2	0.28	0.04	0.30	-0.16
	3 or more	0.21	-6.86***	1.65**	0.63
Adults behavior *referent class: no one*	at least one	11.88***	17.28***	28.22***	
	all	20.48***	27.54***	59.66***	
Adults FV *referent class: 2*	3				2.09***
	4				7.61***
	5 or more				43.02***
Year *referent class: 2013*	2014	0.43	2.61***	-0.01	-0.33
	2015	0.50	3.10***	-0.42	0.56
	2016	-0.97	3.39***	0.53	-0.21
Number of observations [b]		24,601	24,928	24,154	6,897

*** p < 0.01; ** p < 0.05 ; * p < 0.1
[a] Marginal effect computed as discrete change (%) of the partial derivative with respect to each class from the referent one.
[b] The difference of the number of observations is due to the presence missing values.

Finally, adult's behavior seems to be a key predictor of the insurgence of the investigated children's habits. Indeed, data show that the greater the number of adults manifesting the behavior, the higher the likelihood of observing the same behavior among children. Referring to the consumption of the daily-recommended portions of FV the tendency is the same, the greater the number of FV portions consumed by adults, the higher the likelihood of achieving the recommended number of FV portions among children (Table 3).

To conclude, the only behavior that manifests significant variations over the years is the practice of PA, increasing from 2013 to 2016 (Table 3).

4. Conclusions and policy implications

The findings of this study suggest that adults' behaviors still play a role – although inconstant – in the adoption of some children's habits. Indeed, the study shows that the regular practice of PA is much more likely passed down than breakfast-eating, and they are both less likely than daily-snacking. This might support the idea that unadvisable behaviors are more liable to be emulated than the advisable ones. To explore this issue, developments of the research will enlarge the number of observed – advisable and unadvisable – habits.

Furthermore, the research highlights the importance of children's characteristics in driving the adoption of some lifestyle habits. Indeed, jointly with the geographic information, these findings can thus suggest priority targets of intervention for children health promotion policies: Southern and Insular Italy and – limited to some habit – 14-17-year-olds. Another important issue arose in the study is the positive association between the presence of advisable behaviors and the education of the household head. This should hence prioritize health policy interventions to reach socially and economically less advantaged households.

Finally, further developments of the research will focus on the behavior of each individual within the household. This will attempt to understand, in years of profound changes in the environments in which children are growing, who is the key lifestyle promoter within the family.

References

Crockett, S. J., Sims, L. S. (1995). Environmental influences on children's eating. *Journal of Nutrition Education,* **27**(5), pp. 235–249.

FAO/WHO (2003). Diet, nutrition and the prevention of chronic diseases. *Report of a Joint FAO/WHO Expert Consultation*. WHO Technical Report Series, No. 916. Geneva, (CH).

Nicklas, T. A., Baranowski, T., Baranowski, J. C., Cullen, K., Rittenberry, L., Olvera, N. (2001). Family and child-care provider influences on preschool children's fruit, juice, and vegetable consumption. *Nutrition Reviews*, **59**, pp. 224–235.

Patrick, H., Nicklas, T. A. (2005). A review of family and social determinants of children's eating patterns and diet quality. *Journal of the American College of Nutrition*, **24**(2), pp. 83–92.

Story, M., Neumark-Sztainer, D., French, S. (2002). Individual and environmental influences on adolescent eating behaviors. *Journal of the American Dietetic Association,* **102**(3), pp. S40–S51. Supplement.

Tinsley, B. J. (2003). *How children learn to be healthy*. Cambridge (UK).

WHO (2017). *Noncommunicable Diseases Progress Monitor, 2017*. Geneva, (CH).

ISBN: 978-88-5495-135-8

Quantity and mood of final open-ended comments on an Erasmus+ VET mobility questionnaire

Elena Bortolato, Luigi Fabbris, Marco Vivian
Department of Statistical Sciences, University of Padua, Italy

1. Introduction

At the end of a questionnaire sent to students and apprentices who, in 2018, experienced an international internship in schools or companies (Fabbris and Boetti, 2019), the following open-ended question was posed to elicit suggestions from respondents on how to improve future youth mobility: "*Do you have any suggestion about possible ways to improve the aims or ease the mobility experience of future participants?".* Since the questionnaire was filled in by 1031 respondents and the number of responses to the question was 329, a provisional estimate of the probability of respondents writing a suggestion is 31.9%. Non-informative comments, such as 'No', 'Nothing to say' or 'Everything OK', were erased, and this reduced the number of effective suggestions to 269 (26.1% of respondents).

In this work, we analyse the relationships between the quantity and the mood of the suggestions received from respondents. The quantity of the suggestions is represented by the number of words and by typing characters used by the respondents. The mood is the result of an analysis of the words' content, which may reflect the respondent's judgement on his/her mobility experience or of the way the questions were posed.

2. Data and models

The mood was assessed by comparing the binary classifications obtained from an automated procedure with a manual classification executed by human experts. The classification had two possible outcomes: 1 if the mood was positively oriented and 0 otherwise. For the automatic classification of the comments, the *sentiment* command of the *sentimentr* package of the *R* project (Rinker, 2018) was applied. This software interprets the mood of a written response and the estimates are finally dichotomised as positive or negative. In addition, two experts were involved in the experiment: the comments were randomly partitioned into two halves, and each half was assigned to an expert who was asked to classify the comments as either positively or negatively oriented (without knowing the automatic classifications). If the automatic and expert classifications agreed, the agreed mood was considered the real one, while in cases of disagreement, the other expert was asked to reconcile the codes. No ties were admitted.

Various regression analyses were applied by adopting a stepwise selection of predictors. In the first, the mood was conceived as the criterion variable and descriptors of the experience and of the questionnaire quality were taken as predictors (Model 1). In a second analysis, the number of words, its logarithm, the number of typing characters and its logarithm were the criterion variables and the mood and a selection of descriptors of the experience and of the questionnaire quality were taken as predictors. In Table 2, we display an analysis of the logarithm of the number of words (Model 2) and of the number of characters (Model 3) because an initial analysis of the quantity of words and characters gave inexplicable results. The respondent gender and the evaluation score were included in all models.

3. Results and discussion

The mood of the script was classified as negative in 49 cases and as positive in the remaining 220 cases. The automated classifier diverged in 46 cases and the human experts in 25 out of the 269. This

ISBN: 978-88-5495-135-8

indicates that while the mood of a suggestion can be openly positive or negative, there are also cases in which the interpretation may vary. Most of the errors of the automated classifier were Mood = 0 instead of Mood = 1, while the opposite prevailed among the human errors.

The distribution of the number of words had a mean of 22.1 and a long right tail. The number of characters had a mean of 40.3 and again, a long right tail. The logarithms of both numbers showed close-to-normal distributions. By crossing the final mood with the number of words and the number of characters used in the script (Table 1), it was possible to state that the response length correlated highly, as did their logarithms, with content mood.

The data in Table 1 show that the shorter the duration of an internship, the lower its cost and the less relevant the participant's sacrifice of what s/he left at home, the more positive the mood reflected in the comments. In addition, an enhancement of psychological traits and social opportunities derived from mobility would imply a positive mood at the end of the questionnaire. That was why the experience evaluation score was significantly correlated with positive mood.

The regression analyses show that (a) all three models explain a large portion of the deviance of mood (about 52%) and of the logarithm of words and characters used in the sentences (40 and 39%, respectively); and (b) Models 2 and 3 are so similar in terms of their predictors as to induce us to comment on them jointly, just highlighting relevant differences.

The judgement score, which summarises in quantitative terms the internship experience, enables the prediction of both the mood and the length of the comments: the more qualified the mobility experience, the higher the probability of a positive mood. Although, it was the recall of problems encountered during the internship that probably induced respondents first to score lower-than-average their experience, and then to write longer comments.

Table 1: Differences between characteristics of participants and comments, by content mood

	Mood 0	*Mood 1*	*Difference*	*t-test*	*d.f.*	*p-value*
Number of words	34.8	19.3	15.5	3.97	267	0.000
Log(number words)	3.14	2.53	0.61	4.01	267	0.000
Number of characters	212.4	123.9	88.5	3.63	261	0.000
Log(number characters)	4.94	4.47	0.47	3.47	261	0.000
Duration (weeks)	9.49	8.09	1.4	1.74	262	0.083
Cost/week to family-Euro	169.9	120.9	49.0	1.95	241	0.052
Questionnaire interesting	5.71	6.48	-0.77	-2.00	264	0.047
Improved technical skills	0.61	0.76	-0.15	-1.93	267	0.036
Improved self-confidence	0.63	0.89	-0.26	-4.53	267	0.000
Improved life plans	0.49	0.70	-0.21	-2.84	267	0.005
Improved EU news	0.53	0.77	-0.24	-3.42	267	0.000
Integrated with school	0.45	0.61	-0.16	-2.12	267	0.035
Working abroad	0.76	0.89	-0.13	-2.43	267	0.016
Sacrificed comfort zone	0.17	0.31	-0.14	-1.94	254	0.054
Evaluation score	7.39	8.69	1.30	-4.65	260	0.000

Table 2: Regression models for the mood (Model 1), the logarithm of the number of words (Model 2) and the logarithm of the number of characters in final comments (Model 3).

	Model 1	*Model 2*	*Model 3*
Intercept	0.026	3.984***	5.343***
Being a male	-0.339*	-0.410.	-0.279
Male * More extrovert	-0.510***	=	=
Male * Use English at home	0.229**	=	=
Male * Use English at work	=	=	0.302.
Male * Origin country Portugal	-0.227**	=	=
Male * Improve professional autonomy	0.173*	-0.441*	=
Male * More able to control future	0.439*	=	=
Male * More sociable	0.248.	=	=
Male * Sacrificed job opportunities	0.342*	=	=
Male * Stressing questionnaire	=	-0.099**	-0.062*
Male * Days to prepare mobility	=	0.015*	=
Male * Improve finding a job chances	=	0.818***	=
Male * Sacrificed friends	=	-0.803**	=
Male * Improved consciousness	=	=	0.707*
Age * International environment	0.004*	=	=
Evaluation score	0.045**	-0.081*	0.003
Interesting q.re*Improve self-confidence	0.038***	=	=
Clear questions	0.032*	0.047*	=
Easy to fill questionnaire	-0.028.	=	=
Easy to fill q.re * Improve final degree	-0.015*	=	=
Easy to fill q.re * Services sector	0.015.	=	=
Stressing questionnaire	0.110***	=	=
Stressing q.re * Sector industry services	-0.061**	=	=
Stressing q.re* More able control future	-0.071**	=	=
Stressing q.re * Sacrificed friends	-0.152***	=	=
Sacrificed friends	0.530***	=	-0.417*
Sacrificed family	=	-0.352**	-0.268*
Mood	=	-1.023***	-0.693***
Mood * Improve responsibility taking	=	0.926**	1.128***
Mood * Language host country at work	=	0.443**	=
Mood * Language host country at home	=	0.408*	1.039***
Mood * Integrate with origin country	=	-0.428**	-0.331**
Mood * Destination Spain	=	-0.331*	=
Mood * Service sector	=	1.028*	1.495***
Mood * Follow news of EU countries	=	0.297.	=
Mood * Improved integration with school	=	=	-0.301*
Service sector	=	-1.293**	-1.942***
Use language host country at home	=	=	-0.692*
Use English at home	=	0.230.	=
Improve technical skills	=	-0.389**	-0.296*
Improve career chances	=	=	0.297*
Improve responsibility taking	=	-0.533.	-0.719**
Improve commitment to school/company	=	=	0.402.
Improve intercultural skills	=	=	0.185.
Sample size (n)	*(216)*	*(213)*	*(212)*
R^2	0.519	0.404	0.385

Significance: ***= 1‰; **= 1%; *= 5%; .=10%.

ISBN: 978-88-5495-135-8

The mood of the comments correlated negatively with their length, whether singly or together with other predictors: in general, problematic episodes induced longer writings. Instead, there are variables—such as those representing highly successful internships (positive interaction with the host country, both at work and outside; the development of professional skills and of a European feeling; working in the service sector, the most likely to engender fulfilment)—which, provided the mood was positive, increased people's willingness to deliver suggestions. A longer comment could therefore reflect both enthusiasm for the experience and a need to explain what did not work properly and could have been done better. A positive experience that did not prompt the expression of a positive mood at the end of the questionnaire prompted respondents to give more concise comments. The interaction between feelings of integration with the country of origin and mood generated shorter comments: this result was an exception to the relationships suggested above.

In addition, the quality of the questionnaire - which included its interest to the respondent, clarity and ease in answering its questions, or conversely, its tendency to annoy - partly explained the mood of the comments. We ascertained that interest in the content, together with strength of personality traits and clarity of the questions, generated positive feelings. An unexpected result was that a demanding questionnaire can generate a positive mood provided conditions, such as working in the industry service sector, improving the participant's ability to control their future life or feeling that leaving friends for the sake of mobility was a sacrifice are kept separate. These variables identified groups of respondents who had suffered highly from their mobility and so might have found the questionnaire annoying or stressful. Instead, the only descriptor of the questionnaire quality that significantly related to the number of words used was the clarity of the questions. This may mean that the questionnaire quality may have influenced the mood expressed in the open-ended responses but not their length, unless the questionnaire was perceived as stressful, and in this case, respondents commented briefly on it.

The comments of male participants correlated positively with negative mood and shorter comments (in terms of words), while females' comments were longer and more positively oriented. Also, there were a few interactions between being male and the mobility descriptors that correlated negatively with mood: these included becoming more extrovert and moving away from Portugal. In contrast, many variables interacted with being male in determining a positive mood: the use of English at home, improvements in professional autonomy, the perception of an increase in sociability and the capacity to master one's own life, and the sacrifice of job opportunities due to mobility. We can conclude that female participants showed more positive mood than their male counterparts, but also that male participants who had realised a full international experience and had enhanced their personality traits and their professional competence felt so good that they concluded the questionnaire with an informative suggestion.

References

Fabbris, L., Boetti, L. (2019). *ROI-MOB: Measuring the Return on Investment in VET Mobility in the European Union*. Padova: Cleup.

Rinker, T. (2018). sentimentr: Calculate Text Polarity Sentiment. R Package Version 2.6.1, https://CRAN.R-project.org/package=sentimentr.

Balancing multi-class imbalanced data into a training dataset using SCUT method

Rafaela Soares Bueno[a], Luiz Sá Lucas[a], Ana Carolina Sá Lucas[a]
[a] theopinionedge, Rio de Janeiro, Brazil

1. Introduction

The recent advances of technology in all fields have provided a fast data growth, either in quantity or in availability, through many types of source**s**. Machine Learning, the subarea of Artificial Intelligence, is a tool composed of various statistical and computational techniques applied to a huge mass of data. One of the main objectives of Machine Learning is to predict and classify data – things or people – that can be either labeled (Supervised Learning) or unlabeled (Unsupervised Learning) (Kuhn and Johnson, 2013; Lantz, 2015; Torgo, 2017). One of the key issues in Supervised Learning Models is the problem of class imbalance of the response variable, that is, when one or more classes of this variable do not have the same relative frequency with each other, whether the variable is binary or multi-categorical. The class imbalance in the response decreases accuracy, making it harder to predict and classify data, and may misclassify it toward the majority class (Mosly, 2013; Agrawal *et al.*, 2015; Barella, 2015; Lantz, 2015). When the outcome variable is binary, there are some methods to balance the imbalanced class, such as Down, Up, ROSE, and SMOTE (Chawla et al., 2002; Han, Wang, and Mao, 2005). However, when a response variable is multi-categorical, one method should be considered: SCUT (Sahare and Gupta, 2012). The imbalance class problem may be found in many situations. The SCUT method may be valuable to solve that difficulty in many fields. In this case, it was used for an Italian Football Championship dataset, which contains results of some seasons between 2008 and 2012. This dataset is composed by 380 observations and 482 features and, for this study, the response variable characterizes the three possible results of a football match: win, loss, or draw (Carpita *et. al.*, 2014 and 2015). In addition, some indexes measure the performance of the methods, such as Accuracy, Kappa, Sensitivity, Specificity, Precision, Recall, among others.

The objective of this paper is to *i)* present the SCUT function to balance multi-class imbalanced data; *ii)* indicate the best index in this case; *iii)* suggest the best sample size on this training dataset, considering the SCUT method.

The next three sections contain, respectively, the explanation about the function and the indexes, the results, and the conclusions.

2. Methodology

Indexes

There are several performance measures of a classification and prediction model, such as Accuracy, Kappa, Sensitivity, Specificity, among others. In general, the first two ones are mostly used and they may be applied both to binary and multi-categorical response variables.

Accuracy is based on the confusion matrix, which is a $n \times n$ matrix that relates the actual values and the predicted values. When the predicted value is equal to the actual value, the classification is correct. Alternatively, when both differ from each other, the prediction is wrong (Lantz, 2015). Considering the 2×2 confusion matrix and a variable (such as a soccer match win) class of interest, there will be four prediction values, so that two will be correct and two will be incorrect, as follows:

ISBN: 978-88-5495-135-8

- True Positive (TP): Correctly classified as the class of interest
- True Negative (TN): Correctly classified as not the class of interest
- False Positive (FP): Incorrectly classified as the class of interest
- False Negative (FN): Incorrectly classified as not the class of interest

Mathematically, Accuracy is defined as a proportion between the total number of true positives (whether or not the response variable is binary) over the sum of all predictions (Kuhn and Johnson, 2013):

$$Acc = \frac{TP}{TP+TN+FP+FN}.$$

On the other hand, Kappa is more precise than the Accuracy index, because it considers the possibility of a correct prediction only by chance. For this reason, it is very important to solve the class imbalance problem, since *"a classifier can obtain high accuracy simply by always guessing the most frequent class"* (Lantz, 2015, p. 323).

Accuracy and Kappa coefficients range from 0 to 1. The superior limit indicates the perfect agreement between the true and the prediction values.

Methods to balance imbalanced classes' problem

The simplest methods to balance imbalanced classes are Up and Down. The first one consists of duplicating observations from the minority classes through simply resampling with replacement. On the other hand, the method called Down involves randomly removing observations from the majority class to prevent them from dominating the algorithm. The most common technique for doing so is resampling without replacement (Kuhn and Johnson, 2013).

SCUT, which means *SMOTE and Clustered Under-sampling Technique*, is a hybrid sampling method, since it merges oversampling of minority classes, in this case by creating artificial observations, in a way similar to the SMOTE technique (Torgo, 2017) and under-sampling majority class, by applying cluster analysis to stratify the dataset. It is essentially a technique to reduce the imbalance between classes in a multi-class setting. According to Agrawal *et al.* (2015), SCUT is more effective when it is used on pre-processing data phase (higher accuracy values). In this case, SCUT was applied to the training dataset.

One way to implement the SCUT is by splitting the dataset, based on the response variable, into *n* parts / classes, thus calculating the average of observations of all the classes *(m)*. There will be some classes where the number of observations is less than the mean *m*, and for this reason the oversampling is performed in order to make this class size equal *m*. Otherwise, for classes in which the number of observations is greater than the mean *m*, the under-sampling technique is performed to obtain a number of instances equal to *m*. Moreover, the classes that have a number of instances equal to the mean *m* are kept untouched (Agrawal *et al.*, 2015).

3. Results

The dataset used in this study refers to the results of the Italian Football Championship and its target variable characterizes the three possible results of a match: win, loss and draw. However, there is a class imbalance problem, since the sizes of each category are different: 47.4% of win, 27.4% of draw, and 25.3% of loss.

The training dataset is composed by 70% of the dataset. Moreover, a bagged classification algorithm ("treebag") was applied, and in order to estimate the distribution of the Accuracy and Kappa indexes, 2 repeats of **a** 10-folders cross-validation were used.

ISBN: 978-88-5495-135-8

Table 1 shows the results containing the median value of Accuracy and Kappa indexes using the original data as well as three methods: Down, Up, and SCUT (sample size equal to 380 instances).

The Up method registered the highest values of Accuracy and Kappa indexes (0.929 and 0.858, respectively). However, the same indexes by the SCUT method are high as well (0.904 and 0.808), and very close to the ones in the Up method.

Table 1: Median of the Accuracy and Kappa indexes using original data and Down, Up, and SCUT methods

Index	Original data	Down	Up	SCUT
Accuracy	0.824	0.786	0.929	0.904
Kappa	0.501	0.572	0.858	0.808

In order to evaluate the Accuracy and Kappa indexes behavior in the case of the SCUT method, other sample sizes were also tested (400, 600, 900, 1,200, 1,800, and 2,400 instances) and, for that, a bootstrap technique was used. Figures 1 and 2 show that, as the sample size increases, there was an asymptotic trend towards 1.000 (the sample size increased about 8 times – from 380 to 2,400 observations – so that median values of Accuracy and Kappa get closer to 1.000).

Even though it is possible for the Kappa index to reach a median value of 1.000, a sample of 600 observations may be considered a good sample size in order to obtain a satisfactory Kappa index (greater than 0.900), since it increases more than 10 percentage points from 380 to 600 instances (it varied from 0.808 to 0.925).

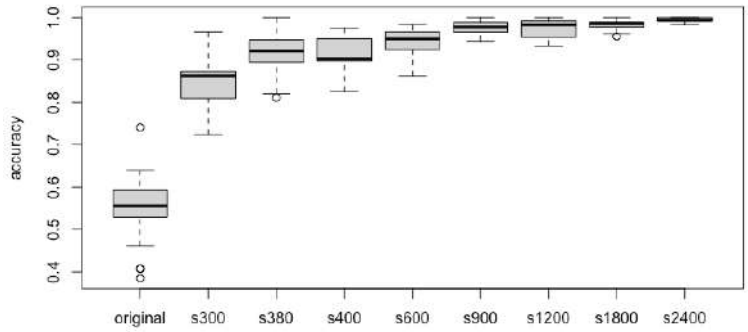


Figure 1: Boxplot of the Accuracy index in the case of the SCUT method with different sample sizes

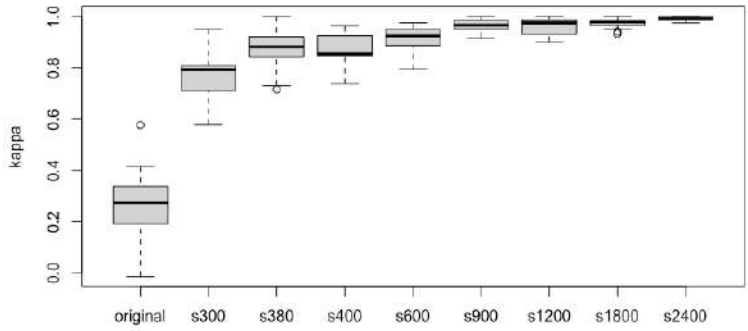


Figure 2: Boxplot of the Kappa index in the case of the SCUT method with different sample sizes

ISBN: 978-88-5495-135-8

4. Conclusions

This study showed the importance of the SCUT method in balancing the class imbalance in a training dataset, as well as proving to be the best index of the predictive model. In comparison to other methods, SCUT turned out to be the best one, since it merges two techniques (oversampling of the minority class and under sampling of the majority class). In addition, the performance coefficients associated to this method were high, being the Kappa index the preferable one, since it compensates for the possibility of a correct prediction only by chance. At last, the study also showed that a large sample size was not necessary in order to obtain good performance measures, because a dataset with 600 instances already produced a Kappa's coefficient greater than 0.900. However, when the sample size increased to 2,400 instances, through bootstrap technique, the Kappa index reached 1.000.

References

Agrawal, A., Viktor, H.L., and Paquet, E., (2015). *SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling*. 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, 2015, pp. 226-234.

Barella, V.H., (2015). *Técnicas para o problema de dados desbalanceados em classificação hierárquica*. São Paulo, Brazil.

Carpita, M., Sandri, M., Simonetto, A., and Zuccolotto, P. (2014). *Football mining with R*. In Zhao, Y. and Cen, Y., editors, Data Mining Applications with R, pp. 398–433.

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

Chawla, N., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., (2002). Smote: Synthetic Minority Over-sampling technique. *Journal of Artificial Intelligence Research*, **16**(1), pp. 321-357.

Han, H., Wang, W.-Y., and Mao, B.-H., (2005). *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning*. Advances in Intelligent Computing, Vol. 3644 of Lecture Notes in Computer Science. Springer Berlin. Heidelberg, pp. 878-887.

Kuhn, M., and Johnson, K., (2013). *Applied Predictive Modelling*. New York, USA.

Lantz, B., (2015). *Machine Learning with R.* Second Edition. Birmingham, United Kingdom.

Mosley, L., (2013). *A balanced approach to the multi-class imbalance problem*. Iowa, United States.

Sahare, M., Gupta, H., (2012). A Review of Multi-Class Classification for Imbalanced Data. *International Journal of Advanced Computer Research*, **2**(3-5), pp.160-164.

Torgo, L., (2017). *Data Mining with R – Learning with case studies.* Second Edition. Portugal.

ISBN: 978-88-5495-135-8

Investigating well-being at work via composite indicators

Stefania Capecchi[a], Carmela Cappelli[a], Maurizio Curtarelli[b], Francesca Di Iorio[a]
[a] Department of Political Sciences, University of Naples Federico II, Naples, Italy;
[b] Prevention and Research Unit, EU-OSHA (European Agency for Safety and Health at Work), Bilbao, Spain

1. Introduction

Psychological or subjective well-being is a multifaceted concept covering several related phenomena, involving emotional responses, feelings and global judgements of satisfaction about life (Howell et al., 2007, among many others), and its multiple domains (e.g. housing, family life, work and so on) (World Health Organization, 2012; OECD, 2013).

Work has long been recognised as having important influences (both positive and negative, indeed) on health and well-being (Litchfield et al., 2016). In modern workplaces - alongside to physical, chemical and biological hazards, depending on the type of industry -, hazards are frequently related more to the way work is organised, to the working environment and the nature of work itself rather than to specific agents, and harm is therefore more psychological than physical (Litchfield et al., 2016). The literature provides a comprehensive account of the topic and of job characteristics (and of their lack) which are considered as psychosocial risk factors for workers (EU-OSHA, 2013, among many others). Nonetheless quantitative evidence about the effect of psychosocial risks on health and well-being is still relatively scarce. Furthermore, analyses of the measure of interactions between physical and psychosocial risk factors seem not frequently reported in the relevant literature.

A common practice in analysing survey data regarding workers' SAH is to consider a few drivers covering a wide range of psychosocial and physical risk factors customarily measured by means of scales administered to respondents: interviewees are usually asked to select a response category out of a list, answering questionnaires often made of a number of question batteries.

Using some of the evidence of the European Working Conditions Survey (EWCS), carried out by the European Foundation for the Improving of Living and Working Condition (Eurofound), we present an empirical analysis where the variable of interest is the self-assessed health (SAH) as a proxy of workers' well-being.

More specifically, this paper focuses on well-being at work in order to build synthetic indicators[1], instead of providing a collection of individual results, aiming at understanding which individual risk factors exert a stronger impact on workers' health at the EU28 level, and whether psychosocial risk factors do affect well-being as much as the physical ones.

After a brief sketch of the data employed and of the implemented procedure, results of the synthetic indicators, as obtained from two subsets of risk factors, are discussed, and few concluding remarks end the paper.

2. Data and methods

Data employed in this exercise come from the Sixth EWCS[2] which provides a wide-ranging

[1] Current literature, as summarized in OECD (2008) handbook, emphasizes several steps to achieve an effective and consistent composite indicator.

[2] Eurofound carried out the 6th wave of the survey in 2015 interviewing 43,850 employees and self-employed workers in 35 European countries: the 28 European Union Member States plus, namely, the candidate countries for EU membership (Albania, F.Y.R.of Macedonia, Montenegro, Serbia and Turkey), and Norway and Switzerland. At country level, the sample size ranges from 1,000 to 3,300 people according to the sample design. Data can be downloaded from http://discover.ukdataservice.ac.uk, while detailed description of survey design and report can be found in Eurofound (2017).

ISBN: 978-88-5495-135-8

picture of Europe at work across countries, occupations, sectors and age groups.

Response variable of interest stems from question Q75 referring to self-reported health status[3], as it is common in literature on the subject: "How is your health in general? Would you say it is: (1 Very good; 2 Good; 3 Fair; 4 Bad; 5 Very bad)".

Common individual characteristics here considered are *gender*, *age* and *education level*. *Gende*r (from question Q2a) is expressed by the usual dummy variable where female = 1; age is expressed in years (Q2b). Education level (from the original Q106) is described by a dummy where holding a university degree =1 (*tertiary*). Given the high number of missing values with reference to net monthly earnings, to investigate the relationship with individual's economic status, information is derived from the answers to Q100: "Thinking of your household's total monthly income, is your household able to make ends meet " (*make-ends-meet*) rated on a six point wording scale from "Very easily" (1) to "With great difficulty" (6). With respect to job features, we introduce two dummies to distinguish full-time vs. part-time job (*fulltime*, where full-time=1) and permanent vs. non-permanent job (*permjob*, where permanent job=1). A dummy is used to consider the belonging of respondent to a country of EU12. Moreover, we reckon the number of working days per week (Q26, *d4w*) and the hours weekly (Q24) spent at work (*whours*).

With respect to risk factors at work as surveyed by the EWCS, we got two different sets[4]: the first one gathers all the *physical risk factors* (15 variables); the second one refers to the *psychosocial risk factors*, which include a list of variables related to the ways work is organised and managed as long as the social environment of workers (27 variables).

Missing values and "don't Know" responses have not been considered in the analyses; therefore, our target sample consists of 21,991 individuals at EU28 level.

Given the nature of the SAH variable, Ordered Probit models have been implemented, for the two subsets of risk factors separately and altogether (estimates are obtained using STATA14 where the dummy variables are treated as usual as factors). Although, with so many variables to be considered, the relationship between the type of risk and self-reported health was difficult to read and globally interpret. To cope with this problem, we have derived two composite indicators, obtained by Principal Component Analysis (Jolliffe, 2011), and then employed them as explanatory variables in a further model implementation.

3. Results

3.1 Derivation of the PCs and composite indicators

To better synthesize data, a Principal Components Analysis (PCA) has been performed distinctly on the two sets of variables, the one comprising the physical risk factors and the other referred to the psychosocial ones, and two composite indicators have been built. For homogeneity purposes, the analysis has been conducted on the correlation matrix.

For space constraints, we do not report in this paper tables and figures which may be available from Authors. To summarise, some variables show a strong correlation with health and well-being of workers. This is the case of variables related to positions or movements during work, and the same can be said for those related to work-life balance or to a positive and motivating work environment, in which workers have a sense of fulfilment with work, feel motivated, have a say, are consulted and participate in decisions, are supported by management and trust managers and, finally, experience good relationships with colleagues

PC1.1: *Physical risk factors.*

[3] The proportion of those claiming a bad or very bad SAH is about 2%, while more than 77% report a positive o very positive evaluation with reference to SAH, and there are not prominent differences in the frequency distributions with regards to gender and work sector.

[4] For brevity, we do not report here the lists of selected variables for the risks in the workplaces, as well as tables for detailed frequency distribution, break downs by gender, sector and type of contract, which are available from Authors.

For these set of covariates, the total inertia of the data is p=15. The eigenvalue associated with the first PC is 4.71 while the second largest is much smaller (1.53). Indeed, the variance of the first PC (PC1) accounts, alone, for 31.4% of the total inertia (4.71/15*100), therefore suggesting that one dimension, provided by the first PC, is enough to synthetize information.

The PC1 thus, ranges from better to worse physical working conditions, and it is indeed negatively correlated to SAH: the higher (with a couple of exceptions) the value of the covariates (which, given the direction of the scale employed for coding, indicates worse physical working conditions), the higher the value of the PC1. Along PC1 it is possible to identify at one end workers with a high exposure to physical risk agents (with positive high scores on PC1) and at the opposite end workers in desk-based jobs (low negative score on the PC1).

PC1.2: *Psychosocial risk factors.*

The synthetic indicator has been derived also for the second set of variables describing psychosocial risks. In this case, the total inertia of the data is *p=26*. The eigenvalue associated to the first PC is 6.17 while the second largest is 2.63. The first PC explains the 23.7% (6.17/26*100) of the total inertia. As in the previous case, also for the psychosocial covariates, one dimension (the first PC) captures most of information in the data, although in this case a second PC might be considered.

Most of the variables are positively correlated to PC1: the higher the value of these variables, which denote a positive and motivating work environment, the higher the value of PC1. In this case, higher values of the synthetic indicator denote better working conditions.

3.2 Ordered Probit models including composite indicators, discussion and limitations of the study

Based on the findings of the PCA, three Ordinal Probit models have been estimated (see Table 1), considering as explanatory the respondents' covariates and the first PC derived from the first set of variables, denoted as *PC1.1*, and from the second set of variables, denoted as *PC1.2*.

Table 1: Ordered Probit Models with Physical and Psychosocial risk factors indicators

SAH	**Mod 1 (PC1.1 and PC1.2)**			**Mod 2 (PC1.1: Physical synthetic indicator)**			**Mod3 (PC1.2: Psychosocial synthetic indicator)**		
	Coef.	*Std. Err.*		*Coef.*	*Std. Err.*		*Coef.*	*Std. Err.*	
2.gender	-0.083	0.016	***	-0.108	0.016	***	-0.023	0.016	
Age	-0.031	0.001	***	-0.029	0.001	***	-0.030	0.001	***
1.tertiary	0.071	0.017	***	0.076	0.017	***	0.073	0.017	***
1.permjob	-0.096	0.021	***	-0.089	0.021	***	-0.085	0.021	***
1.fulltime	0.066	0.025	***	0.085	0.025	***	0.075	0.025	***
1.private	0.027	0.016		0.021	0.016		0.013	0.016	***
Whours	0.001	0.001		-0.001	0.001		0.001	0.001	
d4w	-0.002	0.011		-0.007	0.011		-0.004	0.011	
endsmeet	0.112	0.007	***	0.146	0.007	***	0.130	0.007	***
1.deu12	0.088	0.016	***	0.061	0.015		0.094	0.016	***
Pc1.1	-0.063	0.004	***	-0.080	0.004	**			
Pc1.2	0.096	0.003	***				0.104	0.003	***
/cut1	-3.980	0.087		-3.830	0.085		-3.855	0.086	
/cut2	-3.051	0.0737		-2.926	0.072		-2.934	0.073	
/cut3	-1.732	0.070		-1.642	0.069		-1.629	0.070	
/cut4	-0.111	0.069		-0.061	0.068		-0.021	0.069	

In all the cases, the synthetic indicators built for the two sets of risk, either together (Mod1) or alone (Mod2 and Mod3), turn out to be significant, confirming that they provide an effective synthesis of the underlying variables which exert an impact on SAH. Also, the respondent related

ISBN: 978-88-5495-135-8

characteristics remain significant in the same way, for all the models.

It is worth to stress that the added value of building these synthetic indicators relies on that they allow either for simplifying an analysis or for disentangling specific drivers of work-related well-being, with the additional advantage of removing redundant information.

Nevertheless, it is important to also underline some limitations of this exercise, which stem directly from the data used and the survey itself. First, the physical risk factors are not extensively surveyed in the case of the EWCS and therefore they refer only to a small subset of the sample. Moreover, European and national legislations have targeted this type of risk factors for several decades now resulting in their steadily decrease. Another point regards the limitations stemming from the questionnaire that seem to include too many questions (and variables) and some questions appear to be repetitive as they seek to grasp sometimes the same concept. This seems confirmed by the circumstance that only the first principal component in the two groups of selected variables is significant; in addition, no great contrasts are captured. All in all, a questionnaire including fewer and more targeted questions would allow for gathering a better quality information and would be a more cost-effective solution.

References

EU-OSHA (2013). *Psychosocial risks and workers' health*. OSHWiki (contributor: Hupke M.) available at: https://oshwiki.eu/wiki/Psychosocial_risks_and_workers_health [accessed 24/7/2019].

Eurofound (2017). Sixth European Working Conditions Survey, 2015. [data collection]. 4th Edition. UK Data Service. SN: 8098, http://doi.org/10.5255/UKDA-SN-8098-4.

Howell, R.T., Kern, M.L., Lyubomirsky, S. (2007). Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. *Health Psychology Review*, **1**(1), pp. 83-136.

Jolliffe, I. (2011). *Principal component analysis* (pp. 1094-1096). Berlin Heidelberg: Springer

Litchfield, P., Cooper, C., Hancock, C., Watt, P. (2016). Work and Wellbeing in the 21st Century, *International Journal of Environmental Research and Public Health*, **13**(11), pp. 1065.

OECD (2008). *Handbook on constructing composite indicators. Methodology and user guide*. OECD, Paris.

OECD (2013). *OECD Guidelines of Measuring Subjective Well-Being*. OECD, Paris.

World Health Organization (2012). *Measurement of and target-setting for well-being: an initiative by the WHO Regional Office for Europe*, World Health Organisation: Geneva, available at: http://www.euro.who.int/__data/assets/pdf_file/0003/180048/E96732.pdf [retrieved on line, 20/06/2019].

Exploring the statistical structure of soccer team performance variables using the Principal Covariates Regression

Maurizio Carpita[a], Enrico Ciavolino[b], Paola Pasca[b]
[a] Department of Economics and Management. University of Brescia, Italy;
[b] Department of History, Society and Human Studies. University of Salento, Lecce, Italy.

1. Introduction

In the Data Science panorama, great room for indicators building, as well as predictive modeling is represented by sports data. Match outcome is a non-ambiguous, well-defined response variable that lends itself to the application of statistical learning models. In addiction, the availability of data related to sports players reveals what components of players' performance matter the most, thus representing a topic of particular interest for decision making and best choices in the competitive framework. The European Soccer database, available on Kaggle (KES database) incorporates data about both players and teams of about 20,000 soccer matches for seasons 2009-2015 in 10 different European countries (Carpita et al., 2019b-c). Experts of the EA Sports FIFA videogame (see the website *sofifa.com*) state that the performance of a soccer player is made up of 7 broad dimensions (*power, mentality, skill, movement, attacking, defending* and *goalkeeping*), each of which incorporates, in turn, more specific skills to be developed and mastered by players on the pitch (e.g. *finishing, volleys, crossing, short passing, heading* as components of the *attacking* ability)[1].

Relying on experts' suggestion, Carpita et al. (2019b) modify the original indicators related to the 7 *sofifa* dimensions by incorporating the four player roles (*forward, midfielder, defender, goalkeeper*): results showed that performance skills might play a more or less consistent role according to where players are located in the pitch. However, no statistical inquiry has been carried out on *sofifa* experts' performance indicators. Correlations among them revealed an unclear dimensional structure, making multicollinearity concerns, as well as the reconstruction of broad performance areas worth to be examined in detail. As a first development, Carpita et al. (2019a) used a non-supervised clustering technique for multivariate data which, however, did not significantly improve prediction of match results.

For this reason, it is worth to examine the KES database with clustering techniques that also encompass prediction objectives. *Principal Covariates Regression* (PCovR) fits this purpose: it simultaneously reduces the predictors to a few components and regresses the criterion on these components (De Jong and Kiers, 1992). The predictive performances of the PCovR components are compared with the experts' *sofifa* indicators using the *Skellam Model*, a regression variation that best fits the distribution of home and team goal differences (Karlis and Ntzoufras, 2008).

2. Methods

Principal Covariates Regression This procedure was developed by De Jong and Kiers (1992) to deal with the interpretational and technical problems that emerge when a regression analysis is performed on a relatively high number of predictor variables. The method simultaneously reduces the matrix of the predictor variables $\boldsymbol{X}$ (N, units $\times J$, variables) to a limited number

[1]The 33 original performance variables and their *sofifa* classification in 7 dimensions are in the first three columns of Table 1 at page 3 of this short paper.

ISBN: 978-88-5495-135-8

of components and regresses the vector of the criterion variable $\boldsymbol{y}$ ($N \times 1$) directly on these components. A parameter $\alpha \in [0; 1]$ allows to emphasize the *Principal Components Regression* (PCR, $\alpha = 1$) over the *Reduced-Rank Regression* (RRR, $\alpha = 0$), both being an integral part of PCovR. This translates into a flexible tuning on predictors reconstruction rather than on the predictive power of the regression model and vice versa. PCovR aims at minimizing the loss function:

$$L = \alpha \cdot \frac{||\boldsymbol{X} - \boldsymbol{T}\boldsymbol{P}_{\boldsymbol{X}}||^2}{||\boldsymbol{X}||^2} + (1 - \alpha) \cdot \frac{||\boldsymbol{y} - \boldsymbol{T}\boldsymbol{P}_{\boldsymbol{y}}||^2}{||\boldsymbol{y}||^2}.$$

The left part of L concerns dimension reduction: $\boldsymbol{T}$ is an $N \times R$ score matrix that contains the scores of the N observations on the R components, $\boldsymbol{P}_{\boldsymbol{X}}$ is the $R \times J$ loading matrix that contains the loadings of the predictor variables on the J components. In the right part of L, the criterion variable $\boldsymbol{y}$ is simultaneously regressed on the J components, thus the vector $\boldsymbol{P}_{\boldsymbol{y}}$ ($R \times 1$) contains the resulting regression weights for the criterion variable. The R package `PCovR` allows Vervloet et al. (2015) to carry out PCovR by flexibly setting:

- the number of components to extract;
- the value of the parameter α;
- the rotation option.

In this study, for the loss function L the difference between the home and away team of the first 28 performance variables[2] in Table 1 are used as $\boldsymbol{X}$, and the goals' difference is used as $\boldsymbol{y}$. Moreover, the choice of 4 components with the rotation option $varimax$ provide stable results independently to the α value (the automatic procedure would emphasize the PCR part of L).

Skellam Regression Consider the number of goals scored in a match as a pair of counts (H,A), where H is the number of goals scored by the home team and A the number of goals scored by the away team, so that $Y = (H - A)$ is the goals' difference (if $Y > 0$ the home team won; if $Y = 0$ the home team drew; if $Y < 0$ the home team lost). Assuming that (H,A) is generated by a bivariate Poisson distribution with positive parameters λ_H, λ_A and positive covariance parameter λ_{HA}, the random variable Y has the Skellam (or Poisson Difference) distribution, which does not depend on correlation between H and A. Under these assumptions, the Skellam regression model specification for the random variable of the goals' difference Y is the following (Karlis and Ntzoufras, 2008):

$$\begin{aligned} Y &\sim Skellam(\lambda_H, \lambda_A) \\ \log(\lambda_H) &= \mu_H + \boldsymbol{z}^T\boldsymbol{\beta}_H \\ \log(\lambda_A) &= \mu_A + \boldsymbol{z}^T\boldsymbol{\beta}_A \end{aligned}$$

where $\boldsymbol{z}$ is the ($K \times 1$) vector of the standardized differences between the *home* and *away* team performance indicators (simple averages of the variables grouped using the classification in Table 1) by each of the four players roles, and we expect that for the parameter's vectors $\boldsymbol{\beta}_H > 0$ and $\boldsymbol{\beta}_A < 0$ (Carpita et al., 2019b; Pelechrinis and Winston, 2018).

[2]The five *goalkeeping* variables have been excluded from the analysis for two main reasons: first, those variables only belong to the *goalkeepers* role, thus produced a large amount of NAs for other players' roles; second, from an interpretational point of view, the *goalkeeping* is a very specific role (e.g. variables such as *handling* or *diving* are allowed for *goalkeepers* role only) thus it has not been considered worth to be included in the PCovR.

3. Results

The last two columns in Table 1 gives the two classifications, according to experts (*sofifa*) and PCovR (*pcovr*) with $R = 4$ components and $\alpha = 0.5$. For the *pcovr* classification, the correlation between each variable x and its component with the max column value of the loading matrix $\boldsymbol{P_X}$ is shown in brackets; these correlations are positive and much higher than those with the other three components, with the exception of x_2, x_3 and x_6. The 1^{st} component contains variables belonging to heterogeneous dimensions in experts' classifications; the 2^{nd} component is still mainly characterized by the *defending* abilities; the 3^{rd} components incorporates most of the abilities in the *movement* dimension, along with the *stamina* variable, while the latter components is made up by all the variables related to an *aggressive* response in the match.

Variables		**Classifications**		**Variables**		**Classifications**	
Label	**Long Name**	*sofifa*	*pcovr*	**Label**	**Long Name**	*sofifa*	*pcovr*
x01	shot power	power	comp 1 (0.626)	x19	acceleration	movement	comp 3 (0.881)
x02	jumping	power	comp 4 (0.543)	x20	sprint speed	movement	comp 3 (0.848)
x03	stamina	power	comp 3 (0.422)	x21	agility	movement	comp 3 (0.769)
x04	strength	power	comp 4 (0.727)	x22	reactions	movement	comp 1 (0.651)
x05	long shots	power	comp 1 (0.770)	x23	balance	movement	comp 3 (0.659)
x06	aggression	mentality	comp 4 (0.486)	x24	crossing	attacking	comp 1 (0.695)
x07	interceptions	mentality	comp 2 (0.691)	x25	finishing	attacking	comp 1 (0.687)
x08	positioning	mentality	comp 1 (0.650)	x26	heading	attacking	comp 4 (0.787)
x09	vision	mentality	comp 1 (0.768)	x27	short passing	attacking	comp 1 (0.788)
x10	penalties	mentality	comp 1 (0.654)	x28	volleys	attacking	comp 1 (0.726)
x11	dribbling	skill	comp 1 (0.725)				
x12	curve	skill	comp 1 (0.766)	x29	diving	goalkeeping	goalkeeping
x13	free kick	skill	comp 1 (0.726)	x30	handling	goalkeeping	goalkeeping
x14	long passing	skill	comp 1 (0.702)	x31	kicking	goalkeeping	goalkeeping
x15	ball control	skill	comp 1 (0.805)	x32	gok_positioning	goalkeeping	goalkeeping
x16	marking	defending	comp 2 (0.881)	x33	reflexes	goalkeeping	goalkeeping
x17	standing tackle	defending	comp 2 (0.892)				
x18	sliding tackle	defending	comp 2 (0.886)				

Table 1: Summary of *sofifa* and *pcovr* classifications of the 33 variables of the KES database

The Regression weights vector with the correlations between y and the four components is $\boldsymbol{P_y} = (0.286, 0.097, 0.139, 0.126)^T$, so that the criterion variable (the goals' difference) is more positively correlated with the first and the third components. Note that correlations for the criterion variable y are lower than correlations for the performance variables $\boldsymbol{x}$: as a consequence, the proportion of explained variance for y is only 13% and for $\boldsymbol{x}$ is 66%, so that the weighted sum of the variance accounted for y and $\boldsymbol{x}$ by the four components is 39%. These results could be expected because, considering the correlations between all the 28 predictor variables $\boldsymbol{x}$: the average is 0.32, the median is +0.29, the third quartile is 0.48 and the maximum is 0.87.

Considering the results for the Skellam regression model for the goals' difference, both *sofifa* and *pcovr* predictors $\boldsymbol{z}$ have significant parameters with the expected positive signs for the home team and negative for the away team equation. Table 2 illustrates the main results, with some diagnostics obtained with a 75%-25% split for training and testing: results are very similar, and suggest that the use of the different predictors $\boldsymbol{z}$ does not modify the predictive abilities of the Skellam regression model, for what concerns the final match results. However, note that the number of *pcovr* predictors (13) is lower than the number of *sofifa* predictors (22).

Predictors	bic	n.ind	sigin.H	sigin.A	cor.OE	rmse	mae	acc.3	acc.2	sen.2	spe.2
sofifa	57,579	22	11	10	0.406	1.621	1.262	0.523	0.597	0.820	0.405
pcovr	57,454	13	7	7	0.405	1.621	1.261	0.518	0.591	0.816	0.398

Legend: **bic**: bayesian information criterion for the model; **n.ind**: number of indicators in each equation of the model; **sigin.H-A**: number of indicators with significance < 0.15 in equations H and A; **cor.OE**: correlation between observed and estimated goal differences; **rmse**: root mean square error of the model; **mae**: mean absolute error of the model; **acc.3**: accuracy for the prediction of 3 results (W-D-L); **acc.2**: accuracy for the prediction of 2 results (W-NW); **sen.2**: sensitivity for the prediction of 2 results (W-NW); **spe.2**: specificity for the prediction of 2 results (W-NW).

Table 2: Skellam regression model diagnostics for *sofifa* and *pcovr* predictors

Finally, Fig. 1 illustrates the calibration curves for the match results (win, draw and loss) in *sofifa* (left) and *pcovr* (right): as it can be seen, the prediction for *draws* represent a problematic category for prediction (Pelechrinis and Winston, 2018), while the prediction for *win* approximates the ideal the most, at least up to around 85%. For what concerns *loss*, for an observed probability $> 75\%$ the *pcovr* prediction tends to be under confident.

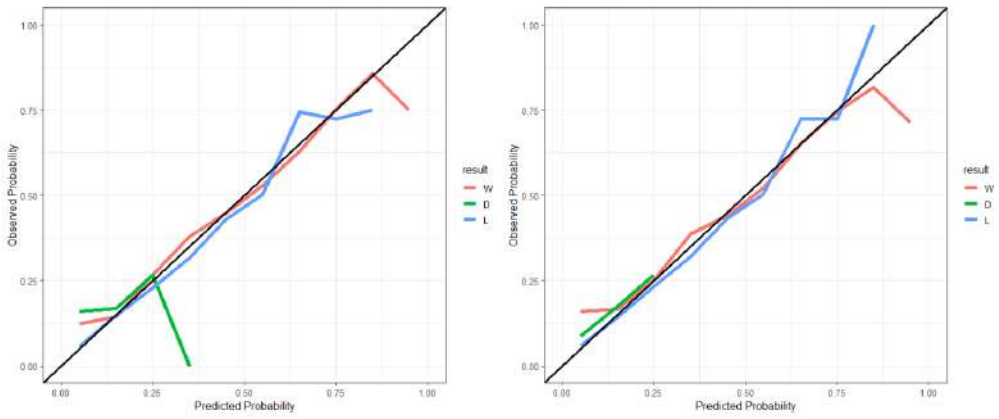


Figure 1: probability calibration curve for the Skellam regression model with 22 *sofifa* predictors (left) and 13 *pcovr* predictors (right)

References

Carpita, M., Ciavolino, E., and Pasca, P. (2019a). Composite indicators of the Soccer Players' Performance Indices. In Mariani P. (Editor): *Data Science & Social Research 2019 Book of Abstracts,* PKE Publisher, Milano (Italy), page 40.

Carpita, M., Ciavolino, E., and Pasca, P. (2019b). Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling*, **19**(1): pp. 74–101.

De Jong, S. and Kiers, H. A. (1992). Principal covariates regression: Part I. Theory. *Chemometrics and Intelligent Laboratory Systems*, **14**(1-3): pp. 155–164.

Karlis, D. and Ntzoufras, I. (2008). Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, **20**(2): pp. 133–145.

Pelechrinis, K. and Winston, W. (2018). Positional value in soccer: Expected league points added above replacement. *arXiv.org – arXiv: 1807.07536 [stat.AP]*.

Vervloet, M., Kiers, H. A., Van den Noortgate, W., and Ceulemans, E. (2015). Pcovr: An *R* package for principal covariates regression. *Journal of Statistical Software*, **65**(8): pp. 1–14.

The mobile phone big data tell the story of the impact of Christo's *The Floating Piers* on the Lake Iseo

Maurizio Carpita
Department of Economics and Management, University of Brescia, Italy;
DMS StatLab - Data Methods and Systems Statistical Laboratory, University of Brescia, Italy.

1. Short story of Christo's *The Floating Piers*

> "*Those who experienced The Floating Piers felt like they were walking on water – or perhaps the back of a whale. The light and water transformed the bright yellow fabric to shades of red and gold throughout the sixteen days*."
>
> Christo

Figure 1: Christo's *The Floating Piers* (*christojeanneclaude.net/projects/the-floating-piers*).

From June 18 through July 3 2016, the Italian Lake Iseo was reimagined, with an international event, *The Floating Piers* (Figure 1), which was free and open to the people. It was a temporary art installation on the water created - using canvases, cables and metal structures - by the contemporary artist Christo Vladimiroff Javacheff (1935-), originally conceived in 1970 together with his wife Jeanne-Claude Denat de Guillebon (1935-2009) as a 3-kilometer-long walkway, that crossed the shores of Lake Iseo (about 100 kilometers east of Milan), from Sulzano to Monte Isola and to the San Paolo's island (Figure 1). Local authorities estimated that 1,2 million people visited the site in the sixteen days of *The Floating Piers* event, an average of 72,000 visitors per day in an area where usually there are about 12,000 residents. Other sources (*marketingdelterritorio.info*) estimated 1,5 million visitors, with a daily average of 100,000 attendances and the peak of 115,000 attendances reached on Friday July 1, 2016. In 2019 the documentary of Andrey Paounov *Christo Walking on Water* told the story of *The Floating Piers* adventure, also remembering some controversies with the local administrations due to the danger that the pontoons would not been able to hold up to the continuous crowds.

In the *smart city* era, mobile phone big data are increasingly used to detecting presence and quantifying the number of people at a given moment in time with reference to a more or less wide area of interest. Analytics derived from mobile phone big data are very useful and used to understand city usage and mobility pattern, and to monitor big social events (Zanini et al., 2016; Manfredini et al., 2015; Carpita and Simonetto, 2014).

In this study, the mobile phone big data and the statistical approach are used to quantifying the impact of *The Floating Piers* event on the area of the Lake Iseo in summer of 2016.

ISBN: 978-88-5495-135-8

2. TIM big data structure and the statistical methodology

Thanks to a two-years agreement between the Statistical Office of the Municipality of Brescia and *Telecom Italia Mobile* (TIM), the DMS StatLab research team (*sites.google.com /a/unibs.it/dms-statlab/*) had access to the TIM mobile phone activity recorded in the period from April 1 2014 to August 11 2016 for the Province of Brescia[1].

Mobile phone big data are geo-referred data collected over a spatial grid and characterized by its latitude and longitude, and over the time. The simpler object containing this information is named *pixel* or *cell*, the elementary component of each *Geographic Information System* (GIS) that allows to gather, organize, manage, analyze, combine, develop and present geographically located information. The TIM big data for the Province of Brescia are into 923x607 *pixel* of 150 m^2 size each, available at intervals of 15 minutes, for a total of more than 40,000 millions of records collected (Metulini and Carpita, 2019). For each pixel and time interval, the corresponding record refers to the *density* (estimated average number) of mobile phones simultaneously connected to the TIM network in that geographic area and time interval. The mobility feature of these data is hidden and is not possible to trace the single TIM user over time. In the standard setting of geo-statistical analysis, a rectangle union of many pixels that cover an area of interest for the study is named *raster*.

In this study, the statistical procedure used to analyze raster big data is based on the *Histogram of Oriented Gradients* (HOG) method, that's a feature descriptor widely used successfully for object detection, pattern recognition and image compression (Tomasi, 2012). The HOG represents an image (in this case a raster) as a unidimensional feature vector, that can be analyzed using standard statistical procedures. In brief, the HOG procedure is the following. First, each raster is partitioned into smaller rasters, and for each pixel of these sub-rasters the vector of gradients $\boldsymbol{g} = (g_x, g_x)$ (differences right – left = g_x and up – down = g_y of density around the pixel) are computed. Second, for each $\boldsymbol{g}$ two measures are computed:

$$Magnitude = ||\boldsymbol{g}|| = \sqrt{g_x^2 + g_y^2} \qquad \text{and} \qquad Direction = \arctan(g_x/g_y)\ .$$

The final HOG object for each smaller raster is obtained binning the *Directions* and sum the *Magnitudes* for each bin; the final HOG vector of the full raster is obtained stacking the vectors of its smaller rasters. The matrix **X,** which has in column the days of the period of interest and in row the stacked HOGs for the 96 quarters of each day, was created and used with the *k-means cluster analysis* to classify the daily profiles (Metulini and Carpita, 2019).

3. Analysis and results of Christo's *The Floating Piers* impact

After a preliminary exploration of the area of interest, one decided to considered three rasters on the Lake Iseo side in the Province of Brescia that in the summer of 2016 were the residential areas mainly affected by *The Floating Piers* event (Figure 2). The first raster includes Monte Isola, the largest inhabited and car-free island of European lakes with about 1,800 residents, that has joined the *Club of the most beautiful villages in Italy*. The second raster includes the municipality of Sulzano, which dates back to Roman times and is a small village of about 2,000 residents just in front of the Monte Isola Island, and where the most important activities are related to tourism and sailing. The third and last raster includes Iseo, a town of about 9,000 residents major tourist center on the south-eastern shore of Lake Iseo, about 20 kilometers north of Brescia; in Roman times, Iseo was crossed by an important Roman consular road that connected Brescia to Val Camonica along Lake Iseo, and today this town is part of the famous *Franciacorta wine region*.

[1] Many thanks to Rodolfo Metulini (DMS StatLab, University of Brescia) and Marie Cointin (Institut Universitaire de Technologie, Université de Bretagne Sud), that worked with me to this research, and to Marco Trentini of the Municipality of Brescia to support this research.

ISBN: 978-88-5495-135-8

Figure 2: The three TIM rasters of the Lake Iseo area considered for the analysis. Dark colors signal absence of TIM users, more bright colors signal presence of TIM users.

The *R AnalyticFlow* (RAF)[2] interface (Figure 3) was used to develop the ETL (*Extraction, transformation and Loading*) with the HOG procedure of the TIM mobile phone big data for the three rasters on the Lake Iseo

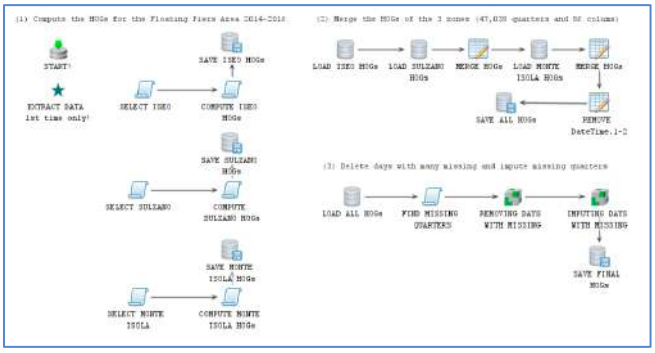


Figure 3: RAF 3-step ETL with HOG procedure for TIM big data and *The Floating Piers* analysis.

Using the *k-mean cluster analysis* of the HOGs respectively of June 2015 and 2016, the *scree-plot* supports the choice of 4 clusters of days: in 2015 the four clusters have about the same number of days, whereas in the month of *The Floating Piers* the cluster 4 contains about 50% of the days. Inspection of data shows that in this cluster there are days of the second half of June, the period of *The Floating Piers* event.

Figure 4 shows the *median TIM density profile* (hereafter *profile*) for weekdays (left) and weekends (right) in the 96 quarters of the day considered as the *Benchmark* (days of June and July 2014 and 2015) and for *The Floating Piers* (from 16 June to 3 July 2016). The range of the *Benckmark profile* is from 2,000 (around 6 am) to 2,500 (around 10 pm) in weekdays and reaches 3,000 in weekends. Applying the moltiplicative factor of 5 (i.e. assuming the TIM market share of 20%)[3], the estimated daily average people in the three areas of the Lake Iseo in June and July 2014 and 2015 was 10,000-12,500 in weekdays and reaches 15,000 in weekends; these estimates are consistent with the official statistics. *The Floating Piers profile* is very much higher: its range is from 3,000 to 6,000 in weekdays and reaches 6,500 in weekends, so that the estimated daily average people in the three areas of the Lake Iseo in the 16 days of the event was 15,000-30,000 in weekdays and reaches 32,500 in weekends.

[2] The *R AnalyticFlow* (*r.analyticflow.com*) is an open source data analysis tool with an intuitive user interface based on the *R* language and environment for statistical computing.

[3] For the newspaper *ilSole24Ore* (Finanza & Mercati, 2016-12-29), in 2016 the national market share of TIM was 30,3%: we use 20%, considering the lower estimate (22,5%) obtained in another study (Metulini and Carpita, 2019) and that during June and July there are many tourists not TIM users.

ISBN: 978-88-5495-135-8

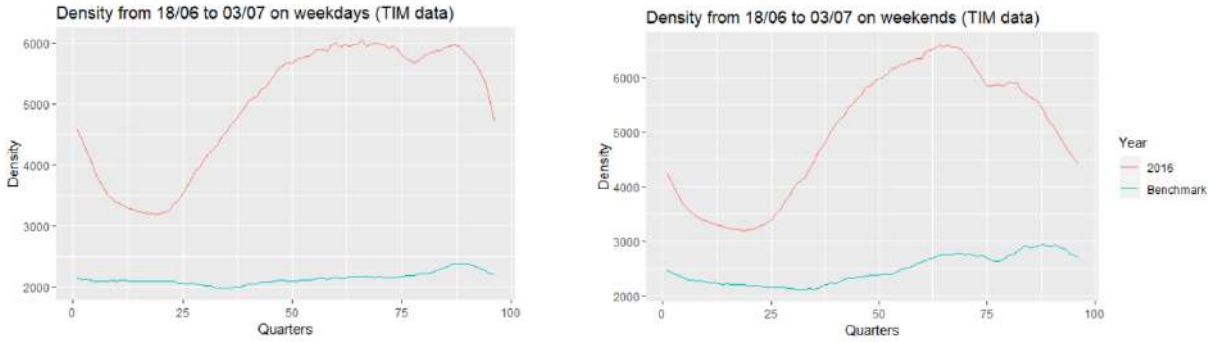


Figure 4: *Benchmark* and *The Floating Piers* profiles for weekdays and weekends.

Figure 5 tells the story of Christo's art installation impact on the Lake Iseo during the 16 days of the event. *The Floating Piers profile* is 2-3 times higher than the *Benchmark profile* and the people that visited the piers on the water increased in the period (the *media effect* has played an important role on this evidence). Applying the multiplicative factor of 5 to the two *profiles*, in the area of the Lake Iseo for the second half of June the estimated benchmark range is 10,000-15,000 people per day, whereas from June 18 through July 3 2016 the estimate of the daily attendances increases to 22,500-32,500 (+45%). As the installation was open from 8 am to 10 pm (14 hours), assuming an average time for the visit of about 4 hours, the median of the daily number of visitors of *The Floating Piers* is estimated from about 78,000 to about 115,000 on Friday July 1, 2016, results consistent with the official statistics.

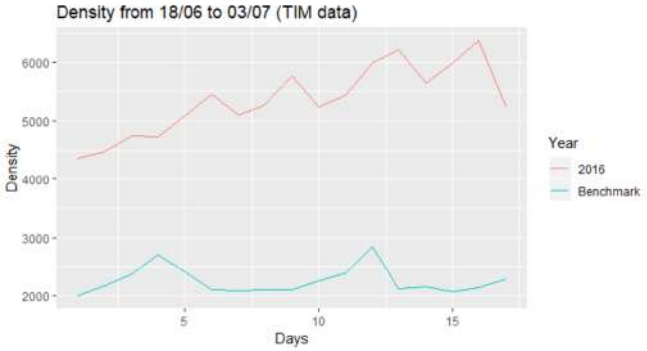


Figure 5. *Benchmank* and *The Floating Piers* profiles for the 16 days of the event.

References

Carpita, M., Simonetto, A. (2014). Big data to monitor big social events: Analysing the mobile phone signals in the Brescia smart city. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems*, **5**(1), pp. 31-41.

Manfredini, F., Pucci, P., Secchi, P., Tagliolato, P., Vantini, S., Vitelli, V. (2015). Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region. In *Advances in complex data modeling and computational methods in statistics*, eds. A.M. Paganoni and P. Secchi, Springer International Publ., pp. 133-147.

Metulini, R., Carpita, M. (2019). A strategy for the matching of mobile phone signals with census data. In SIS 2019 *Conference: Smart Statistics for Smart Applications - Book of short papers*, eds G. Arbia, S. Peluso, A. Pini and G. Rivellini, Pearson, pp. 427-434. Available at: *arxiv.org/pdf/1906. 11739.pdf.*

Tomasi, C. (2012). Histograms of oriented gradients. *Computer Vision Sampler*, pp. 1-6. Update version available at: *www2.cs.duke.edu/courses/fall15/compsci527/notes/hog.pdf*

Zanini, P., Shen, H., Truong, Y. (2016). Understanding resident mobility in Milan through independent component analysis of Telecom Italia mobile usage data. *The Annals of Applied Statistics*, **10**(2), pp. 812-833.

Athletes' mental skills, personality and other drivers to assess the performance in a study on volleyball

Daniela Caso[a], Maria Iannario[b], Francesco Palumbo[b]
[a] Department of Humanities , University of Naples Federico II, Naples, Italy;
[b] Department of Political Sciences, University of Naples Federico II, Naples, Italy.

1. Introduction

To investigate the relationship between sports performance, personality and athletes' mental skills a survey was carried out to a sample of young female athletes enrolled in women volleyball teams of several Series, in the Campania region. Personality was assessed by using the NEOFive Factor Inventory: a 15-item self-report measure that assesses five personality dimensions of extraversion, neuroticism, openness, agreeableness, and conscientiousness (Costa and McCrae, 1992). Participants were required to fill out the questionnaire indicating, on a 5-point scale (strongly disagree, disagree, neutral, agree, strongly agree), their dis/agreement to the statements. Mental skills were evaluated using the sport performance psychological inventory (IPPS-48) made up of 48 items in which respondents state how often (from 1 = never to 6 = always) they describe their sporting experience (Robazza et al., 2009).

Aim of this study is to identify those factors, personality traits, which can successfully assess the performance of team athletes measured by means of an overall composite observable variable which allows overcoming the drawbacks of the measurements obtained with game statistics (Piedmont et al., 1999). Criticism, in fact, arises by providing these direct indexes of the actual level of ability. Among the uncertainty reasons the difficult to generalize findings over different sports/or even from one position to another within a sport; some aspects related to the performance that are difficult to measure; the contribution of other factors which may facilitate or impair performance for the athlete and the team as, for instance, a great ability of some athletes that may have a disruptive behavior, especially in team sport. Recently, it is also considered coaches' ratings of qualities outside of actual performance for measuring a sort of 'collateral' abilities. In this contribution, the observable variable related to performance takes into account training for competitions and official competitions expressed by means of an athletes' rating on a scale with four behaviorally anchored options: 0 refers to *no success* in any type of competitions - 3 *success* in all participated competitions. Thus, an ordinal data model which considers the psychological process of selection was the natural candidate for assessing the stated level of performance as section 3 shows. The next section is about the psychometric scale reliabilities, the last one discusses the mixture model results and provides some concluding remarks.

2. Psychometric instruments

Scale reliability refers to the capability of a psychometric instrument to properly measure the latent dimension (variable) that it is supposed to grasp. In this context, the three considered inventories refer to as many personality traits. The literature considers two different reliabilities: internal and external. This work deals with the internal reliability that is defined as the consistency of the items within the scale. Several more or less sophisticated approaches exist

ISBN: 978-88-5495-135-8

to measure internal reliability, the most widely used is the Cronbach alpha that ranges in $[0, 1]$. Values over $\alpha = 0.8$ are considered satisfactory. The NEOFive Factor (Costa and McCrae, 1992) is a fifteen item extra short version of the well known big-five personality inventory that has fifty items. Both versions have five sub scales that refer to as many personality profiles. The second instruments - as mentioned in the introduction - is the Inventory of Psychological Performance Sports-48 (IPPS-48) (Robazza et al., 2009); it consists of 48 items recorded on six levels ordinal scales that range from "never" to "always". It focuses on both positive and negative mental abilities that influence sport performance. In details, it is referred to Race preparation, Self-talk, Concern, Confidence, Goal-setting, Mental practice, Concentration disorder, Emotional arousal control. First six subscales are referred to positive emotions whereas the former two to negative emotions, and they are supposed to be negatively correlated with the ones in the first group. The Rosenberg Self-Esteem Scale (Rosenberg, 1965) is a 10-item scale designed to assess the global self-worth by measuring both positive and negative feelings about the self. The scale is uni-dimensional. Item intensities are reported using a 4-point Likert scale format ranging from "strongly agree" to "strongly disagree".

For sake of space, exhaustive results about the study on scale reliability are omitted. Table 1 shows the Cronbach alpha and the average correlation between pairs of items. IPPS-48 and Self-esteem inventories ensure alpha index being over the 0.8 threshold.

Notice that some caution needs about the validity of the reliability analysis because the scales are administered in Italian, rather than in the original language. Results in Table 1 indicate

Instrument	# items	# subscales	Cronbach alpha	std alpha	average ρ
NEOFive Factor	15	5	0.49	0.52	0.07
IPPS-48	48	8	0.93	0.93	0.22
Self-esteem	10	2	0.84	0.84	0.35

Table 1: Psychometric scale reliabilities, measured by the Cronbach alpha index.

that the NEOFive Factor scale cannot be considered as validated with respect to our sample of female athletes. Diverse causes can affect the whole scale reliability, albeit the same occurs in if the five subscales are independently considered. Alpha indexes over the threshold correspond to the other two scales.

3. Methods and results

The assessment of perceived sport performance has been analyzed through a mixture model that assumes that the observable ordinal variable Y depends on two components: the first one conveys the personal feeling of the athletes and the second one the inherent uncertainty generated by external circumstances. For a given number of ordinal categories $m > 3$, the assessments $(y_1, y_2, \ldots, y_n)$ expressed by $n = 164$ subjects are the realization of a random sample $(Y_1, Y_2, \ldots, Y_n)$ collected with p covariates summarizing all the available information about respondents possibly useful to explain their perceived performance. Formally the mixture model denoted CUB (Piccolo, 2003) is specified by:

$$\begin{cases} Pr(Y_i = j \mid \boldsymbol{\theta}; \boldsymbol{x}_i) = \pi_i \, b_j(\xi_i) + (1 - \pi_i) \, p_j^U \,, \qquad j = 1, 2, \ldots, m; \\ \pi_i = \pi_i(\boldsymbol{\beta}) = \frac{1}{1+e^{-\boldsymbol{x}_i \boldsymbol{\beta}}} \,; \qquad \xi_i = \xi_i(\boldsymbol{\gamma}) = \frac{1}{1+e^{-\boldsymbol{x}_i \boldsymbol{\gamma}}} \,; \quad i = 1, 2, \ldots, n \,. \end{cases} \tag{1}$$

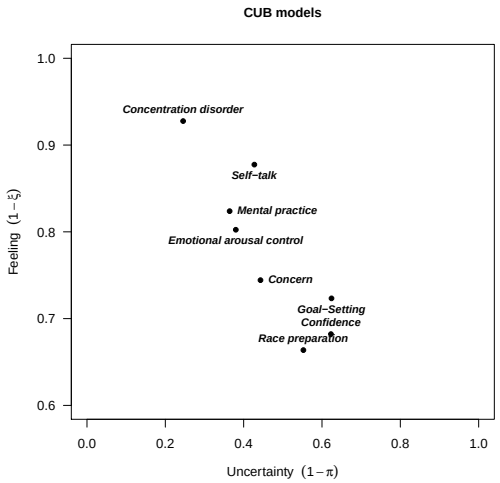


Figure 1: CUB models for IPPS-48 $(m = 6)$.

We set $b_j(\xi_i) = \binom{m-1}{j-1}\xi_i^{m-j}(1-\xi_i)^{j-1}$ and $p_j^U = 1/m$, $j = 1, 2, \ldots, m$, for the probability mass functions of the shifted Binomial and discrete Uniform random variable, respectively, and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$, with $\boldsymbol{\beta}, \boldsymbol{\gamma}$ denoting the parameter vector for the uncertainty and feeling components. Here, $\boldsymbol{x}_i$ is the information set extracted from the matrix $\boldsymbol{X} = ||x_{il}, i = 1, 2, \ldots, n; l = 1, 2, \ldots, p||$. It is used to specify the relationship of π_i and ξ_i with the corresponding (personal or derived by psychological scales) covariates. Given the finiteness of covariates, the parameter space $\Omega(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is an open set and the CUB model is well defined since $\pi_i \in (0, 1)$ and $\xi_i \in (0, 1)$, $i = 1, 2, \ldots, n$. Inferential issues are in Iannario and Piccolo (2016).

For the analysis of perceived performance $1 - \xi_i$ is a measure of high feeling since it increases the probability to give high ratings to the assessment whereas $1 - \pi_i$ expresses the level of individual perceived uncertainty.

Furthermore, to summarize the n responses to a given item in a parametric way (for instance, the different item of psychological scales), no covariates needs to be specified. Then a CUB model collapses to a discrete random variable with probability mass function $Pr(Y = j|\boldsymbol{\theta}) = \pi b_j(\xi) + (1 - \pi)1/m$. Here the baseline model implies the average subject-related parameters. An example is in Figure 1 where the items concerning the IPPS-48 are represented. In the present paper the estimated model for perceived performance is reported in Table 2. Here it is possible to observe that the uncertainty to express the perceived assessment decreases with a high level of *Self-esteem* and increases with mental practice (*Mind*) which is the cognitive (thinking) rehearsal of a physical skill without movement. The level of feeling instead increases with high race preparation (*Train*) and *Concern*, emotional arousal control (*Aurousol*) and practicing of other sports (*Other sports*). A reduction of the feeling related to perceived performance, instead, is related to increasing time spent to study (*H-Study*) in addition to a surplus of *Confidence*.

Thus, the model allows to consider the main aspects that affects the perceived performance on which it is possible to act. Higher level of self-esteem to identify higher performance (see Allen et al. (2013), among others) has been already discussed in the literature. A recent study reveals that psychological factors such as confidence and emotional arousal control, among others, are related to success in sports performance (Vaughan et al., 2018). The other latent traits concerning IPPS-48 (as train and concern) and the other drivers represent new disclosures on which discuss also with reference to a possible comparison with individual sport or male teams.

A more intensive thought has to be done on the use of five-factor inventory in the team sport

Table 2: Fitted model (1) for the assessment of perceived performance

Uncertainty	Estimates	Std. Error	*Feeling*	Estimates	Std. Error
constant	-4.839	3.995	*constant*	5.766	1.818
Self-esteem	0.414	0.200	*Train*	-0.046	0.029
Mind	-0.164	0.096	*Concern*	-0.078	0.032
			Confidence	0.072	0.035
			Arousal	-0.153	0.052
			Other sports	-1.303	0.353
			H-Study	0.319	0.169
Log-lik	-186.9799				

context for well understand athlete' personality. We consider to use the fifty item long version or an alternative shorter version. However, we believe that to administrate the big five inventory questionnaire after a training session might have strongly affected its validity.

The *biopsychosocial* model (Hase et al., 2019) is having a strong diffusion among sports performance studies and represents an alternative interpretative model. It integrates social levels to explain the motivational processes of human performance. A possible extension of the study in this direction may be one of the aims of further analysis.

Finally, identifying the role of psychological factors in sports performance has been recently considered as a remarkable research path that allows a coach the best strategy to help an athlete and its team achieve the best performance and to recognize and support the psychological needs that hinder or benefit performance (Limone and Toto, 2018). All possible results imply that sports coaches might benefit from trying to promote a challenge state in their athletes.

References

Allen, M.S., Greenlees, I., Jones, M. (2013). Personality in sport: a comprehensive review. *International Review of Sport and Exercise Psychology*, **6**, pp. 184–208.

Costa, P. T., McCrae, R. R. (1992). *Revised NEO personality inventory and NEO five-factor inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Hase, A., O'Brien, J., Moore, L. J., Freeman, P. (2019). The relationship between challenge and threat states and performance: A systematic review. *Sport, Exercise, and Performance Psychology.* **8**, pp. 123–144.

Iannario, M. and Piccolo, D. (2016). A comprehensive framework of regression models for ordinal data. *METRON*, **74**, pp. 233–252.

Limone, P., Toto, G. A. (2018) The psychological constructs and dimensions applied to sports performance: a change of theoretical paradigms. *Journal of Physical Education and Sport*, **18**, pp. 2034–2038.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, pp. 85–104.

Piedmont, R.L., Hill, D.C., Blanco B. (1999). Predicting athletic performance using the five-factor model of personality. *Personality and Individual Differences*, **27**, pp. 769-777.

Robazza C., Bortoli L., Gramaccioni G. (2009). L'inventario psicologico della prestazione sportiva (IPPS-48). *Giornale italiano di psicologia dello sport*, **4**, pp. 14–20.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Vaughan, R., Laborde, S., McConville, C. (2018). The effect of athletic expertise and trait emotional intelligence on decision-making. *European journal of sport science*, pp. 1–9.

Partial Least Squares Path Modelling Approach for Sustainability using qualitative information

Rosanna Cataldo[a], Maria Gabriella Grassia[a], Marina Marino[a]
[a] Department of Social Science, University of Naples Federico II, Naples, Italy

Today, the sustainability is defined as the greatest challenge of our generation. It's a complex multidimensional phenomenon, which was studied for couple decades already. It can be seen from different prospective and angles, but the most popular definition is provided by World Commission on Environment and Development (WCED) in the Brundtland Report: "Sustainable development is development that meets the needs of the present, without compromising the ability of future generations to meet their own needs" [WCED (1987)]. Moreover, the United Nations state that "for sustainable development to be achieved, it is crucial to harmonize three core elements: economic growth, social inclusion and environmental protection. These elements are interconnected and all are crucial for the well-being of individuals and societies"

In this view, in 2015, the 193 countries of the United Nations General Assembly adopted the 2030 Development Agenda titled "Transforming our world: the 2030 Agenda for Sustainable Development" and its 17 Sustainable Development goals (SDGs). The 17 Goals in turn hold 169 targets; each target has between 1 and 3 indicators used to measure progress toward reaching the targets. In total, there are 232 approved indicators that will measure compliance. The SDGs cover a range of ambitious objectives to end poverty, protect the planet, and ensure equality and prosperity for all. They are interdisciplinary and cross cutting, with many indicators repeated across Goals, highlighting that progress in any one area depends on simultaneous development in another. In Table 1, the Goals for each area are reported [1].

Table 1: SDGs and the three areas

Area		**Goals**
Social Area	Goal 1	No Poverty
	Goal 2	Zero Hunger
	Goal 3	Good Health and Well-Being
	Goal 4	Quality Education
	Goal 5	Gender Equality
	Goal 6	Clean Water and Sanitation
Economic Area	Goal 7	Affordable and Clean Energy
	Goal 8	Decent Work and Economic Growth
	Goal 9	Industry, Innovation and Infrastructure
	Goal 10	Reduced Inequalities
	Goal 11	Sustainable Cities and Communities
	Goal 12	Responsible Production and Consumption
Environment Area	Goal 13	Climate Action
	Goal 14	Life Below Water
	Goal 15	Life On Land
	Goal 16	Peace, Justice and Strong Institution
	Goal 17	Partnerships for the Goals

Starting from existing Elementary Indicators (EIs) for the estimation of the single areas, we propose a System of Model Based Composite Indicators (CIs), estimate through to Higher-Order PLS-PM, to compute the Global Sustainability CI, in which the hierarchical relationships between this different LVs are considered.

[1] For a detailed description of the individual Goals, refer to the site "Sustainable Development Goals" (www.un.org/sustainabledevelopment/development-agenda/).

Higher-order constructs in PLS-PM are considered as explicit representations of multidimensional constructs that exist at a higher level of abstraction and are related to other constructs at a similar level of abstraction, completely mediating their influence from or to their underlying dimensions [Cataldo et al. (2017), Chin (1998)].

According to PLS-PM approach (see Tenenhaus et al. (2005) for details), it is possible to define the SDGs as a multidimensional Latent Variables (LVs) not measurable directly and related to its single indicators or Manifest Variables (MVs) by a formative relationship.

The Global Sustainability model is reported in Fig. 1, that gives an overview of the approach used in the current research.

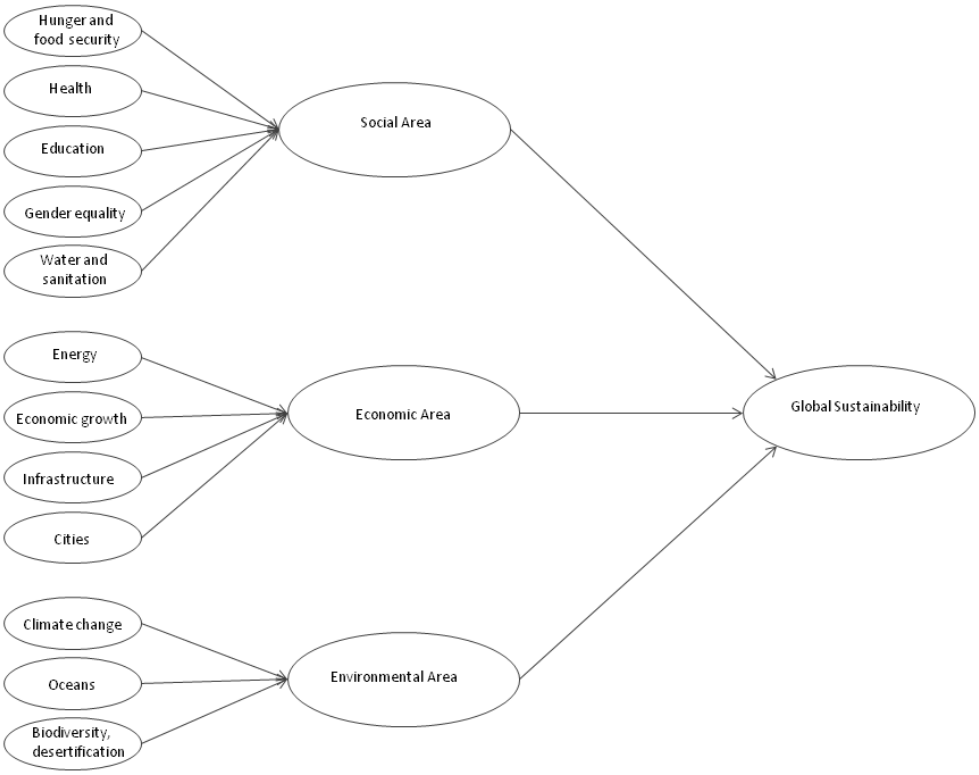


Figure 1: The Sustainability Composite Indicators System

Global Sustainability is, here, conceived as a Third-Order Construct affecting Second-Order dimensions, which in turn shape First-Order LVs underlying specific aspects of the Second-Order dimensions. The study focuses on a formative-formative measurement model, a model resulting from the combination of formative lower order and formative higher order constructs.

A key characteristic of the PLS-PM method is the extraction of CI scores. In the System of indicators built with PLS-PM, you can obtain the scores for each indicator, exogenous or endogenous, and for each indicator you can make a ranking among units. Moreover, PLS-PM provides information on the relative importance of constructs in explaining other constructs in the structural model. Information on the importance of constructs is relevant for drawing conclusions. For this reason, an Importance-Performance Matrix Analysis is a valuable decision making tool. The matrix contrasts the structural model's total effects (the importance) and the average values of the indicator (the performance). As a result, conclusions can be drawn on two dimensions (i.e., both importance and performance), which is particularly important in order to prioritize actions. The analysis is based on a scatter plot where each CI is positioned according to its mean and its path coefficient with respect to the target CI. In this way the scatter plot is divided into four areas:

- the first area is the most critical area, because the CIs have a high impact but a low mean value;
- the second is the area of the monitoring, in which the CIs have a low value for the mean and the path coefficient;
- the third is the area to improve because the CIs have a high mean value and a low path coefficient;
- the fourth is the area to be maintained, in which the CIs have a high value for the mean and the path coefficient.

A similar scatter plot can be considered also for the MVs. In this kind of matrix, we have the possibility to analyze the strengths, weaknesses, opportunities, and threats of constructs, that are considered in the model in order to estimate a latent concept.

It's worth to note that this approach cannot take into account the qualitative information considered in the Sustainable Development Agenda 2030. These indicators are mainly related to geographic area, currency and form of government of each country and constitute potential sources of heterogeneity. The eventual heterogeneity, if not considered, can lead to inappropriate model and inaccurate results. If heterogeneous data structures can be traced back to observable characteristics (observed heterogeneity), potential sources of heterogeneity can be considered by forming groups of data based on observable characteristics. Unfortunately, the sources of heterogeneity in data rarely is fully known a priori. Consequently, situations arise in which differences related to unobserved heterogeneity prevent the PLS path model from being accurately estimated. Indeed, even if the model suits all the standard validation criteria, it is difficult to establish whether it is valid for the whole population or it is merely an average artifact from several sub-populations, as stated in Lamberti et al. (2016). Since researchers never know if unobserved heterogeneity is causing estimation problems, they need to apply complementary techniques for response-based segmentation that allow for identifying and treating unobserved heterogeneity. Several techniques have recently been proposed that generalize statistical concepts such as finite mixture modeling, typological regression, or genetic algorithms to PLS-SEM (see Sarstedt (2008) for a review).

For our purpose, the path modeling segmentation tree (PATHMOX) algorithm proposed by Sánchez and Aluja (2006) is applied. This algorithm aims at automatic detecting heterogeneous segments within the PLS-PM methodology using explanatory variables that account for any heterogeneity in the path model but are not used as indicators in the model. Once an overall PLS path model is estimated, the algorithm, using external variables, identifies the optimal subdivision into two groups whose path models are as different as possible. The two groups are called child nodes, while the global PLS is named the parent node. As an optimal first split is detected, the algorithm proceeds iteratively computing the optimal split of those subgroups. The final outcome consists of a binary tree of models, where every branch of the tree identifies a segment of the population with a specific PLS-PM model in the terminal node of that branch. We apply this approach to analyze sustainability of the 28 European community countries. The data derive from the database of World Bank (www.databank.worldbank.org/data).

Considering that each EI has different units and values, for comparison purposes all units are first normalized to a value between 0 and 1, where 0 was assigned to the least sustainable while 1 was the value assigned to the most sustainable country for each EI. In the case of some EIs, where the larger value meant lower sustainability, the normalization had reverse order, so that the country with the highest EI received a value of 0, while the country with the lowest EI had a value of 1.

ISBN: 978-88-5495-135-8

References

Cataldo, R., Grassia, M. G., Lauro, N. C., and Marino, M. (2017). Developments in higher-order pls-pm for the building of a system of composite indicators. *Quality & Quantity*, 51(2):657–674.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. *Modern methods for business research*, 295(2):295–336.

Lamberti, G., Aluja, T. B., and Sanchez, G. (2016). The pathmox approach for pls path modeling segmentation. *Appl. Stoch. Model. Bus. Ind.*, 32(4):453–468.

Sarstedt, M. (2008). A review of recent approaches for capturing heterogeneity in partial least squares path modelling. *Journal of Modelling in Management*, 3:140–161.

Sánchez, G. and Aluja, T. (2006). Pathmox: a pls-pm segmentation algorithm. In *G.Schmiek (Eds.), Proceedings of KNEMO 2006, number ISBN 88-89744-00-6, Tilapia, Anacapri*, page 69.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., and Lauro, C. (2005). Pls path modeling. *Computational statistics & data analysis*, 48(1):159–205.

WCED (1987). *Our Common Future: Report of the World Commission on Environment and Development*. Oxford University Press.

A composite indicator via hierarchical disjoint factor analysis for measuring the Italian football teams' performances

Carlo Cavicchia [a,b], Pasquale Sarnacchiaro [a], Maurizio Vichi [b]
[a] Department of of Law and Economics, University of Rome Unitelma Sapienza, Rome, Italy;
[b] Department of Statistical Sciences, University of Rome La Sapienza, Rome, Italy.

1. Introduction

In the last years, with the data revolution and the use of new technologies, phenomena are frequently described by a huge quantity of information useful for making strategical decisions. In the current "big data" era, the interest of statistics into sports is increasing over the years. Football is assuredly the most popular sport in Italy, with millions of supporters and amateurs in every cities. 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.

Sportive and economic data are collected for all teams which use statistical analysis in order to measure and improve their performances. The main goal of any Football Championship club is to achieve sport results, by trying to increase their turnover as well.

For dealing with all this amount of information, an appropriate statistical analysis is needed. A priority is having statistical tools useful to synthesise the information arised from the data. Such tools are represented by composite indicators (CIs), that is, non-observable latent variables and linear combinations of observed variables.The strategy of construction of a CI used in this paper is based on a non-negative disjoint and hierarchical model for a set of quantitative variables. This is a factor model with a hierarchical structure formed by factors associated to subsets of manifest variables with non-negative loadings.

In according to the Handbook on Constructing Composite Indicators of the OECD OECD (2004), where the Factor Analysis (FA) methodology is presented as a weighting method used to combine observed indicators, we propose a hierarchical model with the non-negative loadings which best reconstructs the observed indicators according to the common factor model estimated by Maximum Likelihood Estimation (MLE) method. Therefore, loadings are not subjective, but statistically estimated summarizing the observed common relation among data. By hypothesising a two levels hierarchy, the complete system of loadings that best reconstruct the data according to the model is simultaneously estimated.

In this paper, we propose a CI for measuring the Italian football teams' performances, in terms of both sportive and economic variables.

The paper is organised as follows. In Section 2 a description of the methodology is provided. The real application on Italian football teams' performances is presented in Section 3.

2. Hierarchical Disjoint Non-Negative Factorial Analysis

Hierarchical Disjoint Non-Negative Factorial Analysis (*HDNFA*) Cavicchia et al. (2019) is a factorial model that considers two typologies of latent unknown constructs: H specific factors and a single (nested) general factor. *HDNFA* is identified by the two simultaneous equations:

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{A}\mathbf{y} + \mathbf{e}_{\mathbf{x}} \tag{1}$$

$$\mathbf{y} = \mathbf{c}\mathbf{g} + \mathbf{e}_{\mathbf{y}} \tag{2}$$

where $\mathbf{A}$ is the $(J \times H)$ matrix of unknown specific factors loadings, $\mathbf{c}$ is the $(H \times 1)$ vector of unknown general factor loadings, $\mathbf{e_x}$ and $\mathbf{e_y}$ are a $(J \times 1)$ and a $(H \times 1)$ random vector of errors, respectively.

Let include model 2 into model 1 and considering the loading matrix $\mathbf{A}$ is restricted to the product $\mathbf{A} = \mathbf{BV}$ Vichi (2017), the HDFA model is defined

$$\mathbf{x} - \mu_{\mathbf{x}} = \mathbf{BV}(\mathbf{cg} + \mathbf{e_y}) + \mathbf{e_x} \tag{3}$$

Let rewrite the model 3 in matrix form

$$\mathbf{X} = \mathbf{gc'V'B} + \mathbf{E_x} \tag{4}$$

The variance-covariance structure related to the model 3 is

$$\Sigma_{\mathbf{x}} = \mathbf{BV}(\mathbf{cc'} + \Psi_{\mathbf{y}})\mathbf{V'B} + \Psi_{\mathbf{x}} \tag{5}$$

where

$$\Sigma_{\mathbf{y}} = \mathbf{cc'} + \Psi_{\mathbf{y}} \tag{6}$$

such that

$$\mathbf{V} = [\mathbf{v}_{jh} : \forall \mathbf{v}_{jh} \in \{0, 1\}] \tag{7}$$

$$\mathbf{V1}_H = \mathbf{1}_J \tag{8}$$

$$\mathbf{B} = diag(b_1, \ldots, b_J) \; with \; b_j^2 > 0 \tag{9}$$

$$\mathbf{V'BBV} = diag(b_{\cdot 1}^2, \ldots, b_{\cdot H}^2) \; with \; b_{\cdot h}^2 = \sum_{j=1}^{J} b_{jh}^2 > 0 \tag{10}$$

It is assumed that $\mathbf{y} \sim N_H(0, \Sigma_{\mathbf{y}})$ where $\Sigma_{\mathbf{y}}$ is the correlation matrix of the specific factors since they are standardised, and $\mathbf{e_x} \sim N_J(0, \Psi_{\mathbf{x}})$, where $Cov(\mathbf{e_x}) = \Psi_{\mathbf{x}}$ is the J-dimensional diagonal positive definite variance-covariance matrix of the error of model 1 and $Cov(\mathbf{e_x}, \mathbf{y}) = 0$. Furthermore, $\mathbf{g}$ is the random general factor with mean 0 and variance $\sigma_{\mathbf{g}}^2 = 1$ denoting the composite indicator related to a reduced set of specific factors. In addition, $\mathbf{e_y}$ is a non-observable $(H \times 1)$ random vector of errors. It is assumed that $\mathbf{g} \sim N(0, 1)$ and $\mathbf{e_y} \sim N_H(0, \Psi_{\mathbf{y}})$, where where $Cov(\mathbf{e_y}) = \Psi_{\mathbf{y}}$ is the H-dimensional diagonal positive definite variance-covariance matrix of the error of model 2. In addition it is assumed that errors in the two models are uncorrelated $Cov(\mathbf{e_x}, \mathbf{e_y}) = 0$; and errors and factors are uncorrelated, i.e., $Cov(\mathbf{e_x}, \mathbf{g}) = 0$ and $Cov(\mathbf{e_y}, \mathbf{g}) = 0$.

Suppose that a random sample of $n > J$ multivariate observations of $\mathbf{x}$ is observed, the maximisation of the log-likelihood with respect to $\mu_{\mathbf{X}}$ gives the sample mean, thus the reduced log-likelihood is as follows

$$\begin{aligned} L(\mathbf{x}_i, \mathbf{A}, \quad \Psi_{\mathbf{x}} \quad , \Psi_{\mathbf{y}}) = & \qquad (11) \\ = \quad & -\frac{nJ}{2} ln 2\pi - \frac{n}{2}\{ln|\mathbf{A}(\mathbf{cc'} + \Psi_{\mathbf{y}})\mathbf{A'} + \Psi_{\mathbf{x}}| + tr\{[\mathbf{A}(\mathbf{cc'} + \Psi_{\mathbf{y}})\mathbf{A'} + \Psi_{\mathbf{x}}]^{-1}\mathbf{S}\}\} \end{aligned}$$

where $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mu_{\mathbf{x}})'\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i - \mu_{\mathbf{x}})$

This is equivalent to the minimization of the discrepancy function

$$D(\mathbf{x}_i, \mathbf{A}, \Psi_{\mathbf{x}}, \Psi_{\mathbf{y}}) = ln|\mathbf{A}(\mathbf{cc'} + \Psi_{\mathbf{y}})\mathbf{A'} + \Psi_{\mathbf{x}}| + tr\{[\mathbf{A}(\mathbf{cc'} + \Psi_{\mathbf{y}})\mathbf{A'} + \Psi_{\mathbf{x}}]^{-1}\mathbf{S}\} \tag{12}$$

This is a discrete and continuous problem that cannot be solved by a quasi-Newton type algorithm, it is solved by a descendent coordinate algorithm. A general composite indicator should be composed by consistent and reliable specific composite indicators; thus we require that loadings must be positive during the estimation of $\mathbf{Y}$ and $\mathbf{g}$. So the discrepancy function 12 is minimised with respect to $\mathbf{B}_h = diag(\mathbf{b}_h)$ by

$$\widehat{\mathbf{b}}_h = \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{u}_{1h} (\lambda_{1h} - 1)^{\frac{1}{2}} \tag{13}$$

where λ_{1h} and $\mathbf{u}_{1h}$ are respectively the largest eigenvalue and the corresponding eigenvector of the variance-covariance matrix $\widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{S}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}}$ corresponding to variables identified by $\mathbf{v}_{\cdot h}$, that corresponds to h-th column of $\mathbf{V}$. It is important to notice that λ_{1h} and $\mathbf{u}_{1h}$ minimise the function

$$||\mathbf{X}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} - \sqrt{\lambda_{1h}} \mathbf{y}_h \mathbf{u}'_{1h}||^2 \tag{14}$$

where $\mathbf{X}_h$ is the centred data matrix. That can be solved by an Alternate Non-Negative LS algorithm, such that $\widehat{\mathbf{y}}_h$ is estimated by a step of a normal ALS while the estimations of $\widehat{\mathbf{u}}_{1h}$ consists . thus given $\widehat{\mathbf{u}}_{1h}$, $\widehat{\mathbf{y}}_h$ is computed by

$$\widehat{\mathbf{y}}_h = \mathbf{X}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \widehat{\mathbf{u}}_{1h} (\widehat{\mathbf{u}}'_{1h} \widehat{\mathbf{u}}_{1h})^{-1} \tag{15}$$

and given $\mathbf{y}_h$, $\mathbf{u}_{1h}$ is computed by

$$\widehat{\mathbf{u}}_{1h} = \begin{cases} \mathbf{X}_{h+} \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \widehat{\mathbf{y}}_h (\widehat{\mathbf{y}}'_h \widehat{\mathbf{y}}_h)^{-1} \\ 0, \qquad \text{otherwise} \end{cases} \tag{16}$$

where X_{h+} is the set of passive variables. Thus, this is an active set algorithm, where the H inequality constraints are active if the regression coefficient $\mathbf{u}'_{1h}$ in 14 will be negative (or zero) when estimated unconstrained, otherwise constraints are passive. The non-negative solution of 10 with respect to $\mathbf{u}_{1h}$ will simply be the unconstrained least squares solution using only the variables corresponding to the passive set, setting the regression coefficients of the active set to zero.

3. Application

Football teams' performances are complex phenomena, described by a huge quantity of information regarding sportive results and economic indices. The main goal of any Football club is to aim sport results, by taking under control the economic aspect. It is more and more important to find the way to measure football teams' performances in order to provide support for decision making. The number of statistics and measures related to sports is expanding every year and the need of build aggregated index to monitor the teams' behavior is even more important.

The Hierarchical Disjoint Non-Negative Factor Analysis has been applied on a dataset obtained from the financial statements filed by the Serie A football teams.

The indicators into the dataset come from different sources: Engsoccerdata, Opta and Transfermarkt. They are regularly updated and they are free.

In our application, we propose a hierarchically aggregated index that best represents the performances of the football teams in terms of sportive and economic conduct, via the statistical identification of reliable and unidimensional specific composite indicators, which are dimensions that measure specific concepts describing the main components of the football italian teams' performances.

In particular, we analyse the impact that all variables have on points made by football teams participating in the series A championship. Some variables are included into the analysis in order to enrich the information about teams. This approach guarantees good properties for the GCI such as (scale-invariance, non-compensability, non-negativity, reliability, unidimensionality, . . .).

References

OECD (2004). *The OECD-JRC Handbook on Practices for Developing Composite Indicators, paper presented at the OECD Committee on Statistics*.

Cavicchia, C., Vichi, M. (2019). Hierarchical Disjoint Non-Negative Factor Analysis. *Submitted manuscript*.

Vichi, M. (2017). Disjoint factor analysis with cross-loadings. *Advances in Data Analysis and Classification*, **11**(3), pp. 563–591.

The determinants of vaccination behaviour of general practitioners in South Tyrol: differences and similarities between Italian and German respondents

Giulia Cavrini[a], Andrea Lazzerini[b]
[a] Faculty of Education, Free University of Bozen/Bolzano, Italy
[b] Faculty of Statistics, University of Bologna, Italy

1. Introduction

Vaccinations (particularly for infants) have been decisive tools in the battle against infant morbidity and mortality for decades. Cases in point are the eradication of smallpox and the effective control of polio, results achieved due to the mandatory nature of these vaccinations (as it is in Italy) and through systematic vaccination campaigns in countries where polio is unfortunately still endemic. Despite this, in recent years there has been a huge disinformation campaign throughout western nations in which anti-vaccine campaigns have been extremely harmful and dangerous to the population, unfortunately with too often deadly results. Nevertheless, vaccines are among the most effective prevention tools available to clinicians and the success of an immunization program depends on high rates of acceptance and coverage. Moreover, for many decades in Italy and in other European countries, anti-vaccination practices have been spreading, as a reaction to supposed cases of permanent and disabling effects on children that have been attributed to vaccination (Jolley and Douglas, 2014). Children with exemption from school immunization requirements (a measure of vaccine refusal) are at increased risk of measles and pertussis and can infect others who are either too young to be vaccinated, cannot be vaccinated for medical reasons, or were vaccinated but did not have a sufficient immunologic response. Clinicians can play a crucial role in parental decision-making (Bean and Catania, 2013, Blume 2006, Collange *et al*., 2016). Healthcare providers are cited by parents as being the most frequent source of immunization information, including parents of unvaccinated children (Omer *et al*., 2009, Kundi e*t al*, 2016; Verger *et al*., 2015).

In Italy, in recent years, a number of surveys have been conducted to better understand the determinants of vaccination behaviours.

Of the Italian regions, the province of Bolzano has the lowest vaccination coverage and for this reason, an analysis of vaccination behaviour in this territory is required.

The general objective of the project, conducted in collaboration with the Health Authority of Bolzano, is to identify what the ideas behind this resistance are, what the arguments against vaccination are based on, and what influence doctors and health professionals may have on this lack of confidence in science. In particular, the main aims of this project are:

1. To identify the different factors that contribute to the rejection of vaccination.
2. To analyse the real information needs of the population.
3. To assess which tools to use to counteract ignorance and misinformation based on the need for information.
4. To evaluate, through separate analyses, differences between the two language groups.

2. Methods

Study setting and questionnaire.

In order to achieve the objectives, three different surveys have been planned:

1. Collection of the opinions of doctors and paediatricians.
2. Collection of the opinions of parents and prospective parents.
3. An opinion survey of the younger population - age group 18-24 years.

ISBN: 978-88-5495-135-8

The first survey was conducted in 2018 and involved 398 general practitioners (GP), free choice paediatricians and hospital paediatricians. General practitioners and paediatricians play a key role in the vaccination program. That is why we chose to interview them. The questionnaire consisted of 34 items and collected information about GP and paediatricians' beliefs on the safety, importance and utility of vaccines, their trust in the reliability of various sources of information about the benefits and risks of vaccines and their ability to convince parents to vaccinate children. The questionnaire was administered online with the collaboration of the Order of Doctors. The second and the third surveys will be conducted next autumn.

Statistical analysis.

In the data analysis phase, questionnaire variables that could take an integer value from 0 to 10 have been dichotomized into dummies that take value 1, if the original variable took value 9 or 10, and 0 otherwise. The only continuous variable i.e. the age of the respondent has been dichotomized using the median.

Differences between Italian and German speaking respondents in demographic features and survey responses were assessed using χ^2 test, where we report p-values.

Potential determinants that might be associated with the attitude of giving importance to vaccinations, considering vaccines safe and preference for mandatory vaccination rather than what is recommended were identified using multivariate logistic regressions with stepwise selection from the covariates with p-values < 0.1. Data analysis was performed with the software Stata 15.0.

3. Results

Demographics.

Of the 248 respondents who successfully filled out the questionnaire, without significant differences between Italian (66) and German speakers (182), about 70% are general practitioners (66.6% vs. 69.8%) while the remaining 30% are paediatricians, with the same proportion of the Italian/German speakers. As regards the gender of respondents, there is approximately the same percentage of females (51.6% vs. 45.6%) and males (48.4% vs. 54.4%). The median age is around 55 years for both the Italian and German speaking physicians.

Importance and safety:

German-speaking physicians tend to see less value in vaccination as a defence against infectious diseases (84.0% vs. 93.9%, $p < 0.05$) and even consider them less safe (83.3% vs. 63.0%, $p < 0.005$). The Italian speakers are supportive of the need to vaccinate health workers (71.2% vs. 56.9%, $p < 0.05$). Regarding the usefulness of vaccinating health workers, both the Italian and the German speakers agree without significant differences with tetanus vaccination (74.2% vs. 73.1%), hepatitis B vaccination (97.0% vs. 97.8%), flu vaccination (81.8% vs. 79.7%), MMR vaccination (77.3% vs. 81.3%) and varicella vaccination (60.1% vs. 62.1%) while they both believe hepatitis A vaccination to be useless but with a lower percentage for Italian speakers (28.8% vs. 42.9%, $p < 0.05$) and only the Italian majority consider the meningococcal B and C vaccinations useful (77.3% vs. 49.5%, $p < 0.001$; 72.7% vs. 47.2%, $p < 0.001$). In terms of contraindications with vaccines, both groups of respondents without significant differences believe the following to be false: the contraindication to vaccination with breast-feeding (72.7% vs. 66.5%), allergies to vaccines in family members (68.2% vs. 75.2%), down syndrome (84.8% vs. 76.9%), allergy to pollen (83.3% vs. 78.6%) and, for just over the half, with ongoing antibiotic therapy (56.1% vs. 54.9%). Both groups consider the contraindication of vaccines with previous allergy to the same vaccine to be true (9.1% vs. 11.0%). The only significant difference we register is contraindication for colds or not febrile illnesses, which the Italian respondents believe to be false in contrast with the belief of German speakers (78.8% vs. 48.9%). Both German and Italian speaking physicians recommend some

ISBN: 978-88-5495-135-8

vaccines to patients at risk for pathology (89.4% vs. 82.9%). Among those who are parents, all have vaccinated their own children except in one case in each of the populations and both the German and Italian speaking respondents have a similar number of responses to never have any doubts as parents on the issue of vaccination (71.2% vs. 72.9%).

Professional development, information and communication.

Most of both the Italian and German speakers consider it really useful to be informed on scientific articles that correlate vaccines with diseases (78.3% vs. 81.1%). The Italian speaking physicians feel more sure than German speakers about how and where to report an adverse reaction, though the difference between the two lingual groups is significant only in the first case (51.7% vs. 29.2%, $p < 0.01$; 48.3% vs. 37.3%). There is no significant discrepancy regarding the main sources of information on vaccines, where the public health service, ministry websites or foreign health institutions and congresses/meetings appear to be the most probable choices without great differences among them (27.9% vs. 33.9%; 36.1% vs. 23.2%; 24.6% vs. 25.0%) while the Internet, journals and pharmaceutical sales representatives tend not to be chosen (4.9% vs. 7.1%; 4.9% vs. 5.4%; 0.0% vs. 0.6%). Both German and Italian speaking respondents are not fully satisfied with communication on vaccines in the public health service (73.8% vs. 73.0%) but they say they would be willing to attend an informative event on vaccines (91.8% vs. 97.0%). With a significant gap, 9 out of 10 Italian speaking physicians against about 7 out of 10 of German speaking physicians would like to have more information on vaccines (90.0% vs. 73.5%, $p < 0.01$) and among them the Italian speakers prefer to get vaccines information through informative events and periodic mail while the German speakers predominantly through direct update (44.4% vs. 21.0%; 38.9% vs. 28.0%; 14.8% vs. 48.2%, $p < 0.001$).

Patient relationship and action.

There was significant disagreement between the two groups of respondents when asked if mandatory rather than recommended vaccination was more important. Most of the Italian speakers answered in the affirmative in contrast to the German speakers (62.9% vs. 42.4%, $p < 0.01$). Most of both the Italian and German speakers would like to have more information on vaccines to present to parents who are against vaccination (77.0% vs. 64.4%). There are no differences for the most likely reasons that lead parents not to vaccinate their children which are firstly the fear of adverse reactions and secondly the belief of the presence of toxic substances in vaccines (51.7% vs. 54.9%; 31.7% vs. 28.7%) while the reasons “baby too young”, “uselessness of vaccination” and “difficulty in accessing the vaccination” do not even represent 5% (2.7% vs. 1.6%; 1.3% vs. 3.3% ; 0.9% vs. 0.0%). Both German and Italian speaking physicians consider vaccination services to be accessible (91.4% vs. 80.8%) while those who do not think so would change the hours (40.0% vs. 19.5%), the waiting time (20.0% vs. 36.1%) and the place (0.0% vs. 8.3%). Both Italian and German speakers believe that they must give correct information to parents who do not vaccinate (88.8% vs. 90.0%) while a poor percentage say they would respect the decision (1.7% vs. 4.9%). Only the Italian speakers always try to convince a doubtful parent to vaccinate (85.0% vs. 47.5%, $p < 0.001$) and no German speakers against only few Italian speakers succeed in convincing them (8.8% vs. 0.0%, $p < 0.001$). In the experience of the two groups of respondents, the parents who refuse vaccines do not also always refuse traditional medicine (3.6% vs. 1.3%). Both Italian and German speaking respondents tend to consider the age set in the vaccination calendar appropriate (67.3% vs. 52.5%), and those Italian speakers who believe the age set in the vaccination calendar is precocious would firstly postpone vaccinations within the first six months of life and secondly defer the calendar while the German speakers would predominantly reduce the number of multiple-dose vaccinations (40.0% vs. 18.05%; 26.7% vs. 6.1%; 13.3% vs. 40.0%, $p < 0.05$).

Multivariate analysis.

ISBN: 978-88-5495-135-8

We report the results obtained in multivariate logistic regression models. We observe that respondents who are most likely not to believe in the safety of vaccinations are those who: do not consider the age set in the vaccination calendar appropriate (OR 0.27; 95% CI 0.13 – 0.57), are themselves parents and have doubts about vaccination (OR 0.24; 95% CI 0.08 – 0.69), do not give importance to vaccines (OR 0.24; 95% CI 0.08 – 0.07), do not consider vaccination indispensable for health workers (OR 0.36; 95% CI 0.17 – 0.77), speak German (OR 0.39; 95% CI 0.15 – 1.03), do not consider it useful to be informed about scientific articles that correlate vaccines with diseases (OR 0.42; 95% CI 0.16 – 1.08) and finally do not consider themselves informed about which office to report an adverse reaction to vaccines to (OR 0.50; 95% CI 0.23 – 1.11).

The attitude of considering a recommended vaccination more important rather than a mandatory one is more frequent in physicians who never (or only sometimes) try to convince a doubtful parent to vaccinate (OR 0.21; 95% CI 0.07 – 0.61), are under 55 years of age (OR 0.42; 95% CI 0.23 – 0.78), are hospital paediatricians (rather than general practitioners) (OR 0.33; 95% CI 0.13 – 0.78), speak German (OR 0.47; 95% CI 0.23 – 0.95), or consider themselves informed on which office to report an adverse reaction to vaccines to (OR 0.50; 95% CI 0.26 – 0.96).

Finally, respondents who give less importance to vaccines, have a higher probability of: not considering it useful to be informed about scientific articles that correlate vaccines with diseases (OR 0.24; 95% CI 0.10 – 0.54), speaking German (OR 0.30; 95% CI 0.09 – 0.93) and being general practitioners rather than hospital paediatricians) (OR 0.16; 95% CI 0.02 – 1.24).

4. Conclusions

At present, we only have preliminary results but already with these first data, we can see some differences in the behaviour of physicians between the two language groups.

German speaking physicians tend to see less value in vaccination as a defence against infectious diseases and even consider it less safe. Most of both the Italian and German speakers would like to have more information on vaccines to present to parents who are against vaccination. There was significant disagreement between the two groups of respondents when asked if mandatory rather than recommended vaccination was more important. Most of the Italian speakers answered in the affirmative in contrast to the German speakers.

References

Bean, S.J., Catania, J.A. (2013). Vaccine perceptions among Oregon health care providers. *Qual. Health Res.*, **23**, pp. 1251–1266.

Blume, S. (2006) Anti-vaccination movements and their interpretations. *Social Science & Medicine*, **62**, pp. 628-642.

Collange, F., Verger, P., Launay, O., Pulcini, C. (2016). Knowledge, attitudes, beliefs and behaviors of general practitioners/family physicians toward their own vaccination: A systematic review. *Human Vaccines& Immunotherapeutics*, **12**(5), pp. 1282-1292.

Jolley, D., Douglas, K.M. (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions, *PLOS One*, **9**(2), pp. 1-9.

Kundi, M., Obermeier, P., Helfert, S., Oubari, H., Fitzinger S., Yun, J.A., Brix, M., Rath, B. (2016). The impact of the parent-physician relationship on parental vaccine safety perceptions. *Current Drug Safety*, **10**, pp.16-22.

Omer S.B., Salmon D.A., Orenstein W.A., deHart M.P., Halsey N. (2009) Vaccine refusal, mandatory immunization, and the risks of vaccine-preventable diseases. *N Engl J Med*, **360**, pp. 1981-1988.

Verger, P., Fressard, L., Collange, F., Gautier, A., Jestin, C., Launay, O., Raude, J., Pulcini, C., Peretti-Watel, P. (2015). Vaccine hesitancy among general practitioners and its determinants during controversies: a national cross-sectional survey in France. *EBioMedicine*, **2**, pp. 891-897.

ISBN: 978-88-5495-135-8

Analysis of the financial performance in Italian football championship clubs *via* longitudinal count data and diagnostic test

Anna Crisci[a], Luigi D'Ambra[a]
[a] Department Economic, Management, Institutions, University of Naples Federico II, Naples, Italy.

1. Introduction

Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. Professional business operators consider football an important industry with enormous potential in terms of growth and also for the indirect benefits gained by investors and management due to the popularity of football teams. In the football world, major consulting companies provide statistical data relating exclusively to athletic performance and sports results. The recipients of such data can be placed in two main categories. The first concerns professional football players, sports clubs, coaches, sports directors, etc. Such information is sold, in some cases, for payment. The second category is represented by media outlets, which release statistical reports to fans and sports people. 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 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 longitudinal count data. In addition to the introduction, this paper consists of two further sections. In Section 2, the panel data approach is described while, in Section 3 a case study is shown.

2. The panel data

We often have data where variables have been measured for the same subjects (or countries, or companies, or whatever) at multiple points in time. These are typically referred to as Panel Data or as Cross-Sectional Time Series Data. With panel data you can include variables at different levels of analysis (i.e. students, schools, districts, states) suitable for multilevel or hierarchical modeling. Why do we use panel data? (Hsiao, 1985).

Benefits:

- They allow to identify the effects that are not identified in the cross-section data (Ben-Porath,1973).
- The panel allows to study the dynamics: while the cross-section allows you to estimate what proportion of the population is unemployed in a unit of time, the panel data show how this share varies over time;
- The panel data contain more information, more variability and therefore less collinearity among the variables and produce estimates more efficient, more precise parameters.
- They allow to control the effect of individual heterogeneity: i.e variables constant over time (individual heterogeneity) not observed (for which no data are available) (Baltagi and Levin, 1992).

Limits:

- Difficulty in the sample design and data collection.
- Distortion of the measurement errors.

ISBN: 978-88-5495-135-8

- Problem of selection, no answers nor dissensions
- Limited dimension of time series.

2.1Fixed and random effects

The fixed effects (FE) explore the relationship between predictor and outcome variables within an entity (persons, teams, company, etc.). Each entity has its own individual characteristics that may or may not influence the predictor variables. Each entity is different, therefore the entity's error term and the constant (which captures individual characteristics) should not be correlated with the others (Stock and Watson, 2012).

The fixed effect model is:

$$\boldsymbol{y}_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \alpha_i + \boldsymbol{\varepsilon}_{it} \quad (1)$$

where
α_i (i=1…n) is the unknown intercept for each entity (n entity-specific intercepts).
$\boldsymbol{y}_{it}$ is the vector of dependent variables where $i = \text{entity}$ and $t = \text{time}$.
$\boldsymbol{x}_{it}$ represent the vector of covariates.
$\boldsymbol{\varepsilon}_{it}$ is the vector of error terms.

In the random effects the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model. The crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are "stochastic or not". The random effect model is:

$$\boldsymbol{y}_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \alpha_i + \boldsymbol{\varepsilon}_{it} \quad (2)$$

where $\boldsymbol{\nu}_{it} = \alpha_i + \boldsymbol{\varepsilon}_{it}$ is the error of the random effect model.

The generally accepted way of choosing between fixed and random effects is running a Hausman H-test (Hausman, 1978). Statistically, fixed effects are always a reasonable thing to do with panel data (they always give consistent results) but they may not be the most efficient model to run. Random effects will give you better p.values as they are a more efficient estimator, so you should run random effects if it is statistically justifiable to do so. Under the null hypothesis, the random effects is correctly specified, so both the fixed and random effects model are consistent, while under the alternative hypothesis, the random effects are correlated with the regressors, so the random effects model loses its consistency.

3. 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 some financial indicators have on the points achieved by football teams. We consider the following independent variables: Depreciation Expense of multi-annual player contracts (DEM), Net equity (NE) and Revenue net of player capital gain (RNC) . In addition, we have considered, on the bases a bivariate descriptive analysis, also the square effect of DEM (DEM^2), given the non- linear relationship between Point and DEM. Finally, the interaction between DEM and NE (DEM*NE) also was considered.

In order to explore the panel data, figure 1, shows Point versus Year from 2010 to 2015; a line connects the five observations within each team. These lines represent a change over time.

ISBN: 978-88-5495-135-8

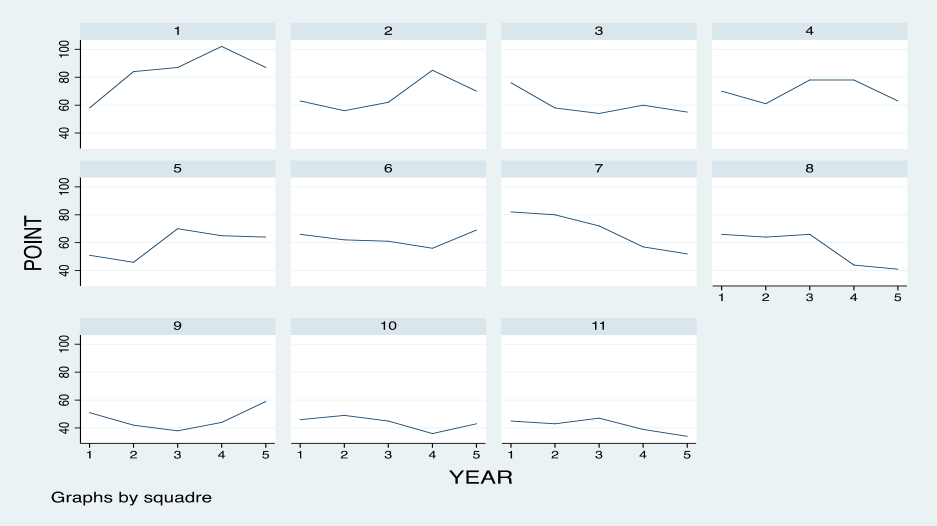


Figure 1: plot Point versus Year from 2010 to 2015

The fixed effects (FE) Poisson model, in table 2, shows a significant overall model (p.value = 0.0397), with only one statistically significant variable: the RNC.

Table 1: Fixed effects Poisson regression

Point	Coef.	Std.Err.	z	p.value
DEM	-0.426	2.864	-0.15	0.882
NE	-0.737	0.801	-0.92	0.358
RNC	0.288	0.104	2.75	0.006
DEM^2	-0.011	0.099	-0.11	0.914
DEM*NE	0.044	0.046	0.96	0.336

The output of the random effects (RE) Poisson model is shown in table 2:

Table 2: Random effects Poisson regression

Point	Coef.	Std.Err.	z	p.value
DEM	3.211	1.546	2.08	0.038
NE	-0.776	0.383	-2.02	0.043
RNC	0.316	0.060	5.25	0.000
DEM^2	-0.120	0.050	-2.3	0.017
DEM*NE	0.048	0.022	2.14	0.033
Cons.	-22.261	12.861	-1.73	0.083
/ln alpha	-7.3223	2.5236		
alpha	0.0006	0.0016		

In the random effects model we have all variables statistically significant. Finally, the Hausman H-test reveals that the random effects estimator is more appropriate (p.value= 0.2851, well above the critical value of 0.05).

Some final consideration should be made. The validity of the RE Poisson depends on very strong distributional assumptions. So, we would just stick to the FE regression. In particular, the choice of dealing with individual effects as fixed or random enough delicate. The fixed effects should be used to estimate the specific effects of the sample (i.e, an exhaustive sample countries, a sample of companies in a particular industry in which the selected sample is representative of the characteristics of the industry). By contrast, the random effects should be used for random samples and to make inference on the population. Then, in our case the choice

could be cast on the fixed effects model, as our entity can not really be thought of as random draws from a population. In fact, the inferences that we have drawn are conditioned to the individuals included in the sample as opposed to a random model where the individual characteristics become a component of the population and the inferences are then related to the same population.

References

Baltagi, B. H. and Levin D. (1992). Cigarette Taxation: Raising Revenue and Reducing Consumption. *Structural change and Economics Dynamic,* **3**(2), pp. 321-335.

Ben-Porath, Y. (1973). Labor-Force Participation Rates and the Supply of Labor. *Journal of Political Economy*, **81**(3), pp. 697-704.

Hausman, J. (1978). Specification Tests in Econometrics. *Econometrica*, **46**(6), pp. 1251-1271.

Hsiao, C. (1985). Benefits and Limitations of Panel Data. *Econometric Reviews*, **4**(1) pp.121-174.

Stock, H.J and Watson, M.W. (2012). *Introduction to Econometrics*, 3rd ed., Pearson Addison Wesley. Chapter 10.

ISBN: 978-88-5495-135-8

Cyberbullying: a threat for relationships and social health

Angela Maria D'Uggento[a], Nunziata Ribecco[a], Ernesto Toma[a], Ignazio Grattagliano[b]
Department of Economics and Finance, University of Bari Aldo Moro, Bari, Italy.
[b] Department of Education Science, Psychology, Communication Science, University of Bari Aldo Moro, Bari, Italy.

1. Introduction

The rapid growth of powerful digital and technological devices such as smartphones and tablets among young people is one of the main factors responsible for the attitude of new adolescents generation to replace personal face-to-face relationships with virtual ones, preferring messages, tweet in social groups and images posted on the web (Instagram, Facebook, etc).

Even though this lifestyle change facilitates and amplifies the contact opportunities among people, it might represent a new possibility for the phenomenon of bullying to spread, assuming an even more subtle and cowardly modality, the cyber-bullying.

Cyber-bullying consists in using technology to hurt, threaten, upset or harass someone else by means of aggressive messages, posts of personal information, pictures or videos. Intimidation and unpleasant comments may deal with a person's behaviour, sexual orientation, physical differences or any other kind of discrimination.

Cyber-bullying is even more dangerous and underhanded than other face-to-face acts of bullying as it's difficult to control and to stop and the victim can be tormented all day long, violating his/her privacy trough any device or computer. The cyber-bully is able to reach a potentially infinite public through the web, preserving his anonymity and without being physically reachable, then can strike even more aggressively and subtly with offenses and insults the victim who cannot defend himself.

Cyber-bullying is, undoubtedly, a serious problem mainly affecting young people, particularly adolescents and pre-adolescents, so adults, parents and educators must counter it with every means. In order to promote a greater awareness of the relational dynamics and the risks connected to the incorrect use of digital devices, to assess the phenomenon of cyber-bullying and to understand the psychological, social and cultural aspects connected to it, some researchers of the University of Bari carried out a field survey, administering a questionnaire to a large sample of high school students during 2018.

Bullying is a complex phenomenon, involving psychological and sociological aspects, often originating from a profound discomfort afflicting both the bully and the victim, therefore, it requires strategies capable of capturing and managing this discomfort.

2. Main results and discussion

The survey *Le determinanti sociali del cyberbullsimo* has been carried out in the first two months of 2018 with the aims of promoting a greater awareness of the risks connected to the phenomenon of cyber-bullying and to understand the underlying psychological, social and cultural aspects.

A questionnaire composed of 32 questions has been administered to a sample of 3,768 students attending several Apulian high schools. The questionnaire has one section on socio-demographic characteristics, the second relates to the use of technological means and is aimed at investigating the critical and conscious use of social networks and media, the third section concerns motivations and reactions of victims, spectators and protagonists of cyber-bullying.

A deeper analysis on those students who declared to have directly experienced cyber-bulling, as victims or harassers, has been carried out in order to draw the two corresponding

ISBN: 978-88-5495-135-8

profiles. Moreover, the main results of the survey were compared with those of similar surveys carried out in Italy.

The sample consists of 43.6% of female students and 56.4% of male students, almost exclusively Italians (97.2%), followed by Romanian and Albanian students (1.1%) and Chinese, to list the most numerous foreign nationalities. The average age was 16 years (15.78±1,507 years). The composition and characteristics of the household reflects the typical southern one, with the presence of both parents, the mother in 98.0% of cases and the father in 94.2%, and brothers and/or sisters (85.9%). The level of parents' education is mainly the high school, but 47.4% of mothers and 45.2% of fathers are graduate. As to working conditions, 83.4% of fathers and 44.5% of mothers are employed.

The "social environment" in which the students live is an important information for the analysis. The sample seems to show a positive relational framework, composed of serene teens, who grow up in the protective shell of the friends and family network. They feel confident that they can talk with friends (61.7%) and family (24.3%) when they have a problem, relying on solid friendship (89.3%) and on the comfort of both parents (83.3%).

The main part of a teenager's day is spent at school, in a relaxed climate, in fact about 98% of the respondents declares to have good relationships with classmates. However, if we move from relationships with the "outside world" and enter the "inner world", we find a quarter of the sample with little appreciation for their physical appearance and this could represent a potential risk factor of exposure to bullying.

As above said, the main research goal is to draw the profile of cyber-bullying victims and harassers to identify any predictors of the causes.

The results show that the victims are 358 (9.5% of the sample), who are women in the 10.9% of their reference group and men in 8.3%, with an average age of about 16 years, and mainly harassed outside school. If we analyse the responses of victims about their physical appearance and gender, we find that 11,0% of the girls declares to be less satisfied versus the 8,0% of the boys and the difference is statistically significant ($p=0.000$). These data are in line with national data that identify a predominantly female victim, not at all satisfied with their physical appearance and, therefore, more easily targeted by cyber-bullying because of their intimate frailties. In our sample, girls are more exposed as they spend much more hours on chat and social media. Male students, instead, use to meet themselves to play several sports, often in a very competitive environment, and this can be considered a conducive occasion to the practice of bullying.

Based on the results of a survey conducted by ISTAT (2015) and Ipsos - Save the children (2015), our data confirms that the interviewees indicate the reason why a boy or girl may become bullied as mainly due to his physical characteristics.

Given the importance that physical appearance has, especially in adolescence, it can be assumed that greater or lesser satisfaction with physical appearance makes potential victims of cyber-bullying. Consequently, crossing the answers on "Satisfaction with own physical appearance" and whether or not to have been a victim shows the presence of a significant relationship, with a double percentage of victims among the unsatisfied, in accordance with other national surveys. Moreover, the victims are harassed by bullies of the same gender ($p=0.000$).

To further investigate, we search for even differences in the type of acts suffered according to gender, but it does not appear to be any significant evidence on average. For both genders, the most frequent harassment suffered by victims are phone prank, insults and threats via mobile phone, exclusion from class groups/friends on main social media, insults and threats via social networks, receiving photos or videos on mobile phone and, finally, posting photos/videos without permission. In particular, the comparison of the data shows a greater frequency of incidents of exclusion from groups of friends in the case of females, as if to obtain a limitation of the social life of the chosen victim while the retaliation suffered by the

ISBN: 978-88-5495-135-8

males' is the privacy breach.

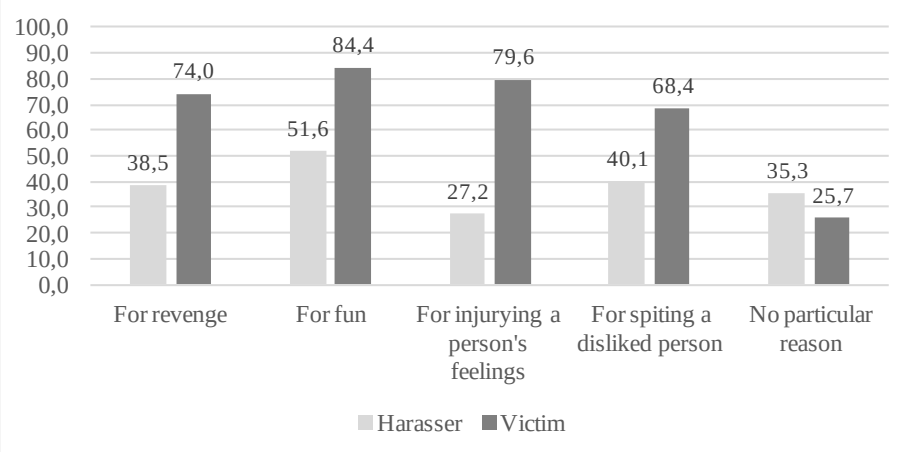
Is there a different reaction according to the gender of the victims? In general, the answer is negative; in any case, intimate suffering, trying to ignore their persecutors, hoping that they desist from despicable intentions are the main reactions, the latter being more typical for women. As shown in Table 1, about one-fifth of the responses relate to a request for help expressed to an adult, usually parents, while just over 10% of responses show a sense of impotence, despite to another cumulative 10% who report the adoption of some proactive self-defense actions.

Table 1. Main victims' reactions by gender

Victims' reactions	*% F*	*% M*	*% T*
I ignored what was happening, hoping that the episodes would not happen again	28,8	20,9	24,9
I told a parent or an adult who takes care of me	22,6	17,4	20,1
I didn't do anything, I felt helpless	11,3	11,0	11,2
I directly asked the harasser to stop	10,7	9,3	10,0
I told it to a friend	6,2	7,6	6,9
I tried to do to them what they had done to me	4,5	5,2	4,9
I blocked text messages / phone calls / emails from harasser	4,0	4,7	4,3
I told it to a teacher	2,8	4,7	3,7
I changed my mobile number	1,7	1,2	1,4
Other	7,3	18,0	12,6
Total	*100,0*	*100,0*	*100,0*

As opposed to the victims, we then analysed the profile of the harassers or that of spectators of some acts of cyber-bullying. Fortunately, they are a minority, about the 8.2% of the sample. Investigating the possible reasons why cyber-bullying is carried out, we find almost the same opinions expressed by those who have claimed to have committed these acts as protagonists/spectators or have suffered them as victims. According to both groups, in fact, it happens more frequently for having fun, for revenge or to spite a disliked person. The intention to injure a person's feelings seems to be more significant for the victims, as shown in Figure 1.

Figure 1. Most frequent reasons inspiring cyber-bulling acts according to victims and harassers.



Interestingly, asking if they acted as bully being alone or with others, it emerges that twice out of three actions took place along with peers or older friends, confirming the theory

ISBN: 978-88-5495-135-8

of acting as a gang. Finally, the first reaction of those who witnessed cyber-bullying as viewers was to console the victim (40.7%), then to intervene to stop the episodes (30.3%) while a further 17.6% did not give importance to the incident, revealing a certain superficiality.

The results of the research highlighted that bullying is a complex phenomenon often originating from a profound unease about the bully, the group as well as the victim and, therefore, requires strategies capable of capturing and managing this discomfort.

References

Grattagliano, I; Craig, F.; Lisi, A.; Pierri, G.; Stallone, V.; Margari, L.; Lecce, P. et al. (2018). Awareness of the offense and perception of the victim among juvenile sex offenders. *La clinica terapeutica*. Vol. 169-4. pp. 155-164.

Greco, R.; Grattagliano, I.; Toma, E.; Taurino, A.; Bosco, A.; Caffò, A.; Catanesi, R. (2017). Cyberbullying: a new form of bullying or a specific manifestation of violence on web? *Rassegna italiana di criminologia*. 11(1), pp. 76-82.

Greco, R.; Grattagliano, I.; Toma, E.; Taurino, A.; Bosco, A.; Caffò, A.; Catanesi, R. (2017). The role of internet and computer communication tools on quality of relationships between preadolescents. A pilot study. *Rassegna italiana di criminologia*. 11(1), pp. 67-75.

Istat (2015), Il bullismo in Italia: comportamenti offensivi e violenti tra i giovanissimi. Istat, Roma.

Ribecco, N.; D'Uggento, A. M.; Toma, E. (2018). Cyber-bullismo, generazioni connesse: un'esperienza a scuola. *Induzioni*, 56(1), pp. 39-59.

Smith, P.K.; Mahdavi, J.; Carvalho, M.; Fisher S.; Russell S.; Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 49:4, pp. 376-385.

Legge 29 maggio 2017, n. 71 *Disposizioni a tutela dei minori per la prevenzione ed il contrasto del fenomeno del cyber-bullismo*. (17G00085), GU Serie Generale n.127 del 03-06-2017.

MIUR, *Linee di orientamento per la prevenzione e il contrasto del cyber-bullismo*. Aprile 2015.

Osservatorio Nazionale Adolescenza (2017). *Nella rete della rete*. Report annuale. Roma.

Generazioni connesse. (2018). *Studio sugli effetti del Cyberbullismo in Italia*. Safer Internet Day2018. Università degli Studi di Firenze.

Save the children-IPSOS (2015). *I nativi digitali conoscono veramente il loro ambiente?* Safer Internet Day2015. Save the children Italia Onlus.

ISBN: 978-88-5495-135-8

Quantile Composite-based path modelling to handle differences in territorial well-being

Cristina Davino [a], Pasquale Dolce[b], Stefania Taralli [c], Domenico Vistocco[d]
[a] Department of Economics and Statistics, University of Naples Federico II, Naples, Italy;
[b] Department of Public Health, University of Naples Federico II, Naples, Italy;
[c] ISTAT[1], Ancona, Italy;
[d] Department of Political Science, University of Naples Federico II, Naples, Italy

1. Introduction

The Italian system of indicators on Equitable and Sustainable Well-being (Benessere Equo e Sostenibile - BES) proposed by the National Institute of Statistics represents a well-established reference database in the national and international debate on the research on alternative well-being measures. The main strengths of this set are represented by the broad coverage of all the components of this complex concept and the availability of information not only at the aggregate level but also at the provincial level (NUTS3 level) (Istat, 2019; Taralli et al., 2015). In this framework, it is possible to consider not only the levels of well-being but also the differences in their distribution thus highlighting differences in the territories. The paper proposes an advancement of work elaborated in Davino et al. (2018), where a hierarchical composite model was used to study relationships among components of the BES. The proposed hierarchical composite model allows us to synthesize individual indicators into single indexes, in order to construct composite indicators at a global and a partial level. Partial Least Squares path modeling (Lohmöller, 1989) and a recent method, called Quantile Composite-based path modeling (Davino and Esposito Vinzi, 2016), were used respectively to estimate average effects in the network of relationships among variables and to explore whether the magnitude of these effects changes across different parts of the variables distributions. The present contribution aims to deepen the study taking into account that living conditions are quite different according to unobserved or observed heterogeneity (for example according to the geographic area of the province).

2. A quantile composite-based model in the BES framework

Both Quantile Composite-based path modeling (QC-PM) and Partial Least Squares path modeling (PLS-PM) aim at computing proxies of complex constructs, which cannot be directly observed, as composites (i.e. weighted aggregates of the corresponding manifest variables, or MVs). If on one hand, PLS-PM is based on simple and multiple ordinary least squares (OLS) regressions, QC-PM introduces a quantile approach in the traditional PLS-PM algorithm using quantile regression (Koenker an Basset, 1978) and quantile correlation (Li et al., 2014) in the estimation steps. It follows that QC-PM can be considered as a complementary approach to PLS-PM able to highlight if and how the relationships between MVs and among composites change according to the explored quantile of interest.

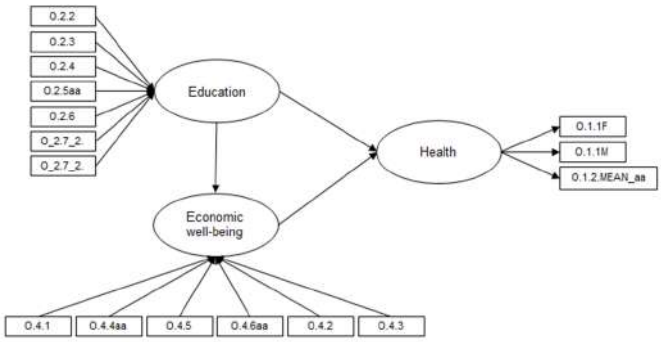
The paper discusses the potential of QC-PM considering a model based on three domains of the BES measured on Italian provinces within the BES framework: Education, EconomicWell-being and Health. The objective is to exploit the quantile regression's ability to explore the

[1]Disclaimer: The paper is the result of collaboration among the authors. Istat is not responsible for the contents.

ISBN: 978-88-5495-135-8

entire distribution of dependent variables to manage territorial differences.

Figure 1 (left-hand side) shows the specified network of relations. The underlying hypothesis, supported by literature and empirical studies, is that Economic well-being and Education affect Health. We consider both the direct effects on Health and the effect of Education on Economic well-being (human capital is a factor of economic growth). Thus assuming that the effect of Education on Health is also mediated by Economic well-being. Education, Economic Well-being and Health are the unobserved complex concepts that are measured as composites of the corresponding MVs (squares in left-hand side and description in right-hand side of Figure 1), taking into account the whole nomological network of dependence relationships.



Indicator (MV)	Description
O.1.1F	Life expectancy at birth (females)
O.1.1M	Life expectancy at birth (males)
O.1.2.MEAN_aa	Infant mortality rate
O.2.2	Percentage of people with at least upper secondary education level (25-64 years old)
O.2.3	Percentage of people having completed tertiary education (30-34 years old)
O.2.4	Proportion of new-graduates from upper secondary education enrolled for the first time at university by cohort of upper secondary graduates
O.2.5aa	Percentage of people not in education, employment, or training (Neet, 15-29 years old)
O.2.6	Participation in long-life learning (25-64 years old)
O_2.7_2.8	Level of literacy as scores obtained in the tests of functional numeracy and literacy skills of students in the II classes of upper secondary education
O_2.7_2.8_AA	Differences between males and females students in the level of numeracy and literacy
O.4.1	Ratio of total disposable income of households to the total number of residents.
O.4.4aa	Pensioners with pensions under 500 euros per month
O.4.5	Ratio of total net wealth of households to the total number of residents.
O.4.6aa	Rate of bad debts of bank loans to families
O.4.2	Average annual salary (employees)
O.4.3	Average annual per capita amount of pension income

Figure 1: A network of relationships among Education, Economic Well-being and Health (left-hand side) and description of the MVs (right-hand side).

3. Main results

Because of lake of space, Figure 2 shows only a part of the results, i.e. the impact played by Education and Economic well-being on Health (path coefficients in the PLS-PM jargon). Bars in each panel represent (from the top to the bottom) path coefficients measuring respectively the effects on the conditional average and on the conditional quartiles of Health. It is interesting to note how QC-PM results complement PLS-PM results: if on one hand Economic well-being is the most important driver of Health, on the other hand its effect decreases moving from provinces with good to worse health conditions. With regard to Education, the coefficients show exactly the opposite trend: the role played by a higher Education is greater in provinces with better health levels.

Differences in living conditions may be associated with the geographical location of the province (Davino et al., 2017). A possible source of heterogeneity could be, for example, the geographical area considering that Italian provinces are usually grouped into four areas: north-east (20%), north-west (23%), centre (20%) and south and islands (37%). Figure 3 shows the distribution of the three composites obtained by estimating the model in Figure 1 with PLS-PM, distinguishing the effect of the area. The three composites are highly correlated at a national level but differently at a local level. Moreover, a greater heterogeneity is beginning to emerge in the southern provinces, with similar structures for the north-east and north-west and the centre with an intermediate trend. The results of the QC-PM can provide a better definition of the characteristics of this heterogeneity.

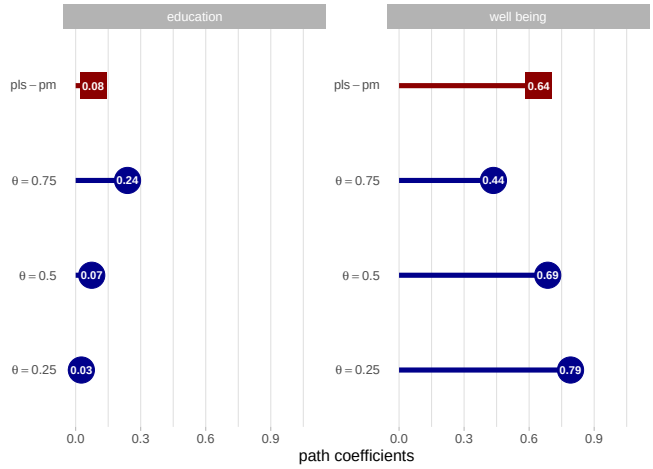


Figure 2: Path coefficients linking Education and Economic well-being to Health.

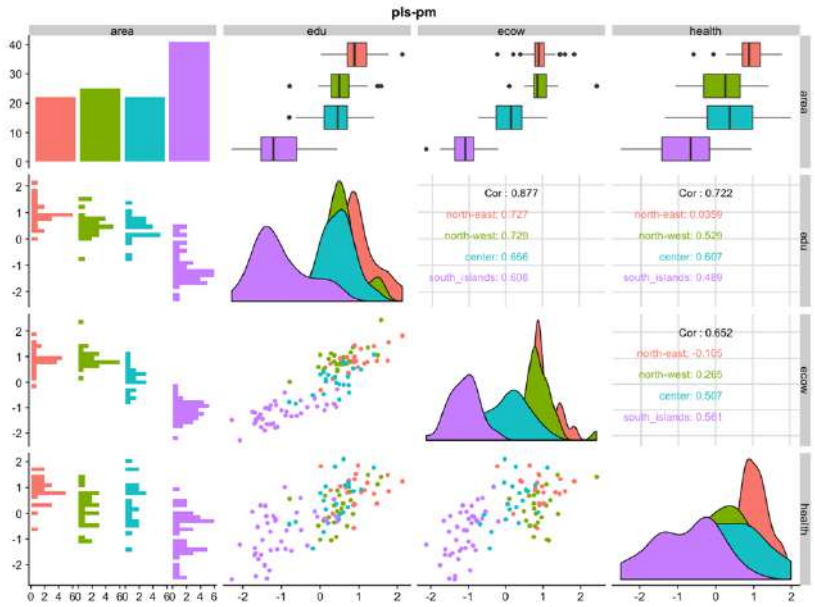


Figure 3: Education, Economic well-being and Health distributions according to the geographic area.

Focusing on the Health composite and on its three conditional quartiles ($\theta = [0.25, 0.5, 0.75]$), it is possible to analyse similarities and differences among the geographical areas at different health conditions. Figure 4 shows the distribution of the Health composite for each area (different panels) and for each model (rows in each panel). The density plot, the dot diagram and the boxplots allow to explore all the features of the distributions. In each line a segment joins the averages of the composite at the three quartiles. Taking into account that the global averages of the composites provided by the PLS-PM and by the three QC-PMs are equal respectively to 0, -0.47, 0.01 and 0.47, it is possible to note that the averages of the southern provinces distributions are always below the global average, while north-eastern provinces (and partially also the north-western ones) show an opposite behavior."

A further investigation of the results should include the analysis of the role played by the MVs which can suggest appropriate actions to policy makers. Moreover, the proposed model can be also exploited to provide conditional quantile predictions of the Health MVs given the explanatory blocks (Education and Economic well-being).

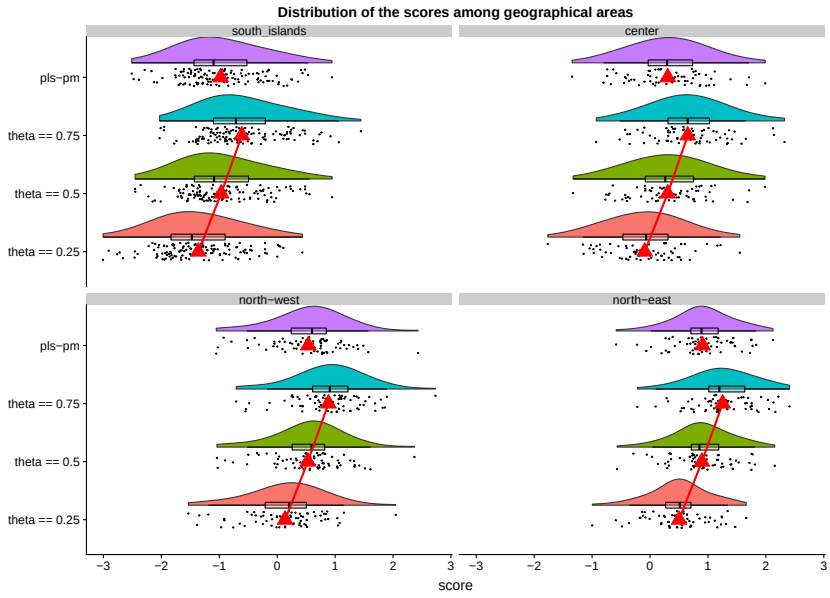


Figure 4: Distribution of the Health composite from a PLS-PM (top in each panel) and QC-PM estimated at the three quartiles, according to the geographic area.

References

Davino, C., Esposito Vinzi, V. (2016). Quantile composite-based path modelling. *Advances in Data Analysis and Classification*, **10**(4), pp. 491–520.

Davino, C., Dolce, P. Taralli, S. (2017). Quantile Composite-Based Model: A Recent Advance in PLS-PM. A Preliminary Approach to Handle Heterogeneity in the Measurement of Equitable and Sustainable Well-Being, in *Partial Least Squares Path Modeling. Basic Concepts, Methodological Issues and Applications*, eds H. Latan and R. Noonan, pp. 81-108, Springer International Publishing.

Davino, C., Dolce, P., Taralli, S., Esposito Vinzi, V. (2018). A Quantile Composite-Indicator Approach for the Measurement of Equitable and Sustainable Well-Being: A Case Study of the Italian Provinces. *Social Indicators Research*, **136**, pp. 999–1029, Dordrecht, Kluwer Academic Publishers.

Istat (2019). Misure del Benessere dei territori. Anno 2018, Roma, Istat, https://www.istat.it/it/benessere-e-sostenibilità/la-misurazione-del-benessere-(bes)/il-bes-dei-territori.

Koenker, R., Basset, G. (1978). Regression quantiles. *Econometrica*, **46**, pp. 33–50.

Li, G., Li, Y., and Tsai, C. (2014). Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association*, **110**(509), pp. 233–245.

Lohmöller, J.B. (1989). *Latent Variable Path Modeling with Partial Least Squares*. Physica-Verlag, Heildelberg.

Taralli, S., Capogrossi, C., Perri, G., (2015) Mesauring Equitable and Sustainable Well-being (BES) for policy-making at local level (NUTS3). *Rivista Italiana di Economia Demografia e Statistica*, **69**, pp. 95–106.

Modeling the joint effect of intensity and duration of alcohol drinking with bivariate spline models

Gioia Di Credico[a], Jerry Polesel[b], Luigino Dal Maso[b], Carlo La Vecchia[c], Francesco Pauli[a], Nicola Torelli[a], Valeria Edefonti[c] on behalf of the INHANCE Consortium

[a]Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy;

[b]Cancer Epidemiology Unit, Centro di Riferimento Oncologico CRO IRCCS, Aviano, Italy;

[c]Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milano, Italy;

1. Introduction

Spline functions, defined as piecewise polynomials with a fixed degree, whose joint points are called knots, are highly flexible tools to modeling non-linearity between a response and some continuous covariates [1]. In epidemiological studies, the number and position of knots usually have an important meaning. Therefore, special attention should be posed to techniques that allow to choose the number and position of knots. Here, we will follow one of the most recent approaches to variable selection in a Bayesian context. Estimating the positions of the knots is not easy and, for a fixed degree, regression coefficients and locations of knots have to be estimated simultaneously, turning the estimation into a non-linear optimisation problem.

The aim of the present work is to: 1. introduce a two-step Bayesian procedure within the semiparametric generalised linear model framework, to be applied in epidemiological studies where the effect of a continuous exposure on risk is under investigation; 2. show how this framework is applied in a bivariate context, where the aim is to modeling the joint effect of intensity and duration of alcohol drinking in cancer of the oral cavity.

2. Methods

The present work assumes the following model:

$$E[y_i] = g^{-1}(\eta_i), \qquad \eta_i = z_i\alpha + f(x_i), \quad \text{for } i=1,...,n,$$

where Y is the dependent variable, g is the link function and η is the linear predictor. Furthermore, Z is the covariate vector that enters linearly in the model, α is the vector of regression coefficients and X is a continuous variable affecting the response through a smooth function $f : R \rightarrow R$, described with a spline with few knots.

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. A typical example arises when truncated power basis of order one is used. In this case, positions of knots represent change points in the slope. Keeping a low number of knots alleviates numerical instability and slow convergence of the optimisation algorithm (Ruppert et al, 2003). Let then:

$$f(x)=\beta_0 +\beta_1 x+ \Sigma_k \gamma_k(x-\xi_k)_+, \; k=1, \ldots, K$$

where ξ_k is the position of the k-th knot, K is the total number of knots and the last term is the truncated linear function. The usual approach is, therefore, to: 1. choose the knot locations using standard criteria (Ruppert et al, 2003); 2. estimate models with a different number and location of knots and compare them through standard criteria, such as AIC, or GCV. This procedure often results in an insufficient ability to discriminate among all competing models.

Free-knot regression splines

A possible extension is to estimate knot locations together with the other regression coefficients. Within a maximum likelihood approach, exploration of the objective function surface

ISBN: 978-88-5495-135-8

could locate local maxima, and this may lead to solutions that strongly depend on the starting values. A Bayesian specification of the model and exploration of the posterior distribution, by using Markov Chain Monte Carlo simulations, could be much more effective.

No variable selection approach

A tricky approach to estimate both number and location of knots is to estimate several models with free knot locations and an increasing but fixed number of knots. A reasonable constraint of ordering of knots should be included in the prior structure:

$$\xi_k \sim \text{Unif}(\min(x), \max(x)), \text{ subject to } \xi_k \leq \xi_{k+1}, \text{ for } k = 1,...,K.$$

Diffuse priors on the regression and spline coefficients are chosen: α i.i.d. $\sim N(0,\sigma_\alpha)$ and β i.i.d. $\sim N(0,\sigma_\beta)$, where both σ_α and σ_β are selected such that the prior distribution is weakly informative. We will refer to this model as the *no variable selection (NVS) model*.

Models with an increasing number of knots are compared on the basis of diagnostic tools such as trace plots and Rhat to check convergence of parameters. Information criteria are used to choose the best model. Unfortunately, a large number of models have to be considered, with huge computational effort in high dimensional problems.

Stochastic search variable selection approach

Checks on simulated data show that the *NVS approach* performs well 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 overparameterised model;
- fit the final model on a restricted set of knots by simultaneously estimating knot locations, regression and spline coefficients.

In the first step, we estimate a model having more knots than reasonably warranted. This leads to an overparameterised 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, overparameterising the model may lead to chains that converge at different points. As an example, suppose that the true number of knots is two and we simulate two chains to fit the model with 5 ordered knots. It can happen that in the first chain the first and the second knot parameters, say ξ_1 and ξ_2, converge on the values of the true knots, while in the second chain the second and third knots parameters, ξ_2 and ξ_3, converge on the right values. Looking at the posterior results distinctly for each chain would let us to properly recognize the presence of two knots.

Since each knot location is uniquely linked to a spline coefficient, we evaluate the potential presence of a knot based on the analysis of the posterior distribution of the associated coefficient.

The concept underlying the proposed methodology is to perform variable selection on the basis functions, for this purpose we employ spike-and-slab priors. Several versions of this approach have been proposed in the literature, 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. 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 $\lambda_k \sim \text{Beta}(a,b)$, with $a = b$. Standard deviations of the two mixture components, σ_{sl} and σ_{sp}, are chosen to be respectively large and small.

Our method adapts the modified *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

ISBN: 978-88-5495-135-8

two components specified in the SSVS approach described above, while each element of the mixing proportion vector λ is now defined as: $\lambda_k|\xi_k \sim \text{Beta}(a,b_k)$, where a is a positive but very small value and $b_k : [\min(x); \max(x)] \rightarrow [a; 1 + a]$ 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 are defined as before.

Bivariate extension of the SSVSξ model

In this section we apply the proposed *SSVSξ* methodology to pooled data derived from the International Head and Neck Cancer Epidemiology (INHANCE) Consortium (Winn et al, 2015) to assess the joint effect of intensity and duration of alcohol drinking on the risk of cancer of the oral cavity. We therefore introduce the bivariate extension of the *SSVSξ* model. In detail, we specify a semiparametric logistic model for the Y that is explained by confounding factors Z (which enter linearly in the model) and variables X (alcohol intensity) and W (alcohol duration) which enter in the model as a bivariate linear spline function with truncated linear basis:

$$f(x,w) = \beta_0 + \beta_1 x + \beta_2 w + \beta_3 xw + \sum_{kx=1}^{K_x} \gamma_i^{(x)} \left(x - \xi_{kx}^{(x)}\right)_+ + \sum_{ky=1}^{K_w} \gamma_j^{(w)} \left(w - \xi_j^{(w)}\right)_+$$

$$+ \sum_{kx=1}^{K_x} \gamma_i^{(xw)} \left(x - \xi_{kx}^{(x)}\right)_+ w + \sum_{ky=1}^{K_w} \gamma_j^{(wx)} \left(w - \xi_{ky}^{(w)}\right)_+ x$$

$$+ \sum_{i=1}^{K_x} \sum_{j=1}^{K_w} \gamma_{kx,ky}^{(xw)} \left(x - \xi_{kx}^{(x)}\right)_+ \left(w - \xi_{ky}^{(w\text{å})}\right)_+$$

Prior distributions on the knot positions are still defined as Uniform distributions on the range of the related predictor, subject to ordered constraint. Similarly, priors on the regression coefficients α and on the spline coefficients β are defined as weakly informative T (3,10) and (3, 2.5). In addition, prior distributions on the spline coefficients γ and on the mixing parameters λ are the following ones:

$\pi(\gamma_{kx}|\lambda_{kx})=\lambda_{kx}N(0,100)+(1-\lambda_{kx})N(0,0.1)$, $\pi(\gamma_{kw}|\lambda_{kw})=\lambda_{kw}N(0,100)+(1-\lambda_{kw})N(0,0.1)$,
$\pi(\gamma_{2,kx}|\lambda_{2,kx})=\lambda_{2,kx}N(0,100)+(1-\lambda_{2,kx})N(0,0.1)$, $\pi(\gamma_{2,kw}|\lambda 2_{,kw})=\lambda_{2,kw}N(0,100)+(1-\lambda_{2,kw})N(0,0.1)$,
$\pi(\gamma_{3,kx,kw}|\lambda_{3,kx,kw}) = \lambda_{3,kx,kw}N(0, 100) + (1 - \lambda_{3,kx,kw})N(0, 0.1)$, and:
$\lambda_{kx}|\xi_{kx} \sim \text{Beta}(0.5,b_{kx})$, $\lambda_{kw}|\xi_{kw} \sim \text{Beta}(0.5,b_{kw})$, $\lambda_{2,kx}|\xi_{kx} \sim \text{Beta}(0.5,b_{kx})$, $\lambda_{2,kw}|\xi_{kw} \sim \text{Beta}(0.5,b_{kw})$, $\lambda_{3,kx,kw}|\xi_{kx,kw} \sim \text{Beta}(0.5, \min(b_{kx}, b_{kw}))$,
where $b_{kx} : [\min(x); \max(x)] \rightarrow [0.5; 1.5]$ and $b_{kw} : [\min(w); \max(w)] \rightarrow [0.5; 1.5]$.

Due to the high number of parameters to be estimated, we chose to fix the number of knots to 2 for both risk factors. We run 10 chains with 2,000 iterations each. Initial values for the knot location parameters are chosen uniformly spread on the linked predictor range; regression and spline parameters are initialised at 0. The mixing parameters are initialised at 0.9.

The INHANCE Consortium

The INHANCE Consortium aims at elucidating the aetiology of head and neck cancer through large-scale epidemiological studies all over the world (Winn et al, 2015). Within data version 1.5, 33 case-control studies provide information on alcohol intensity and duration. All of them collected information on cancer of the oral cavity in current drinkers (4839 cancer cases; 25,871 controls). Calculations were carried out using the open-source Stan program (Stan Development Team, 2017).

3. Results

Figure 1 shows the 2- and 3-dimensional representations of oral cancer risk for different combinations of intensity and duration of alcohol drinking.

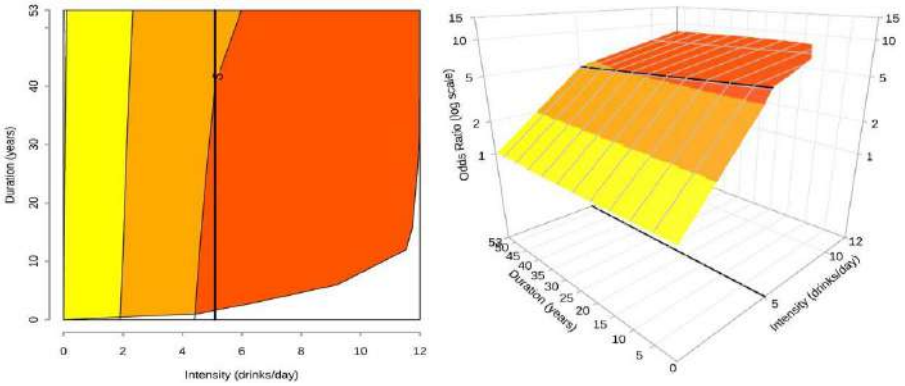


Figure 1: Contour and perspective plots representing the risk of oral cancer as related to alcohol drinking intensity and duration. On the grid, black thicker lines represent knot locations. Dark grey lines in contour plots (upper panel) indicate iso-risk curves at defined levels of risk.

After model selection, the model showed 1 knot for intensity (~5 drinks/day) and no knots for duration. At the highest levels (from 4 drinks/day onward, and any duration), the odds ratio reached 5.

4. Conclusions

The proposed methodology aims at estimating the number and position of knots in semiparametric regression models with linear splines. A well-known variable selection technique has been adapted in order to estimate the presence/absence of knots in possible overparameterised models. Once the number of knots is selected, the appropriate model can be fitted. Moreover, the method allows to inspect the marginal posteriors of knot locations.

In terms of computational complexity, a higher number of parameters has to be estimated, as compared to the *NVS* approach. On the other hand, this approach requires to estimate one model only to select the number on knots, while the *NVS* approach needs estimating a possibly large number of models, especially in the bivariate case. Lastly, to compute the WAIC or LOO criteria, in the *NVS* approach additional simulation steps remarkably increase the memory needed to store the simulations.

The methodology is designed for situations in which the number of expected slope changes is limited. This is reasonable in many epidemiological studies. When we expect a high number of knots, this methodology may be not appropriate, but despite this, the knots corresponding to the most evident slope changes are correctly identified in a simulation study, thus supporting the application of this methodology.

References

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge.

Winn, D.M., Lee, Y.C., Hashibe, M., et al. (2015). The INHANCE consortium: toward a better understanding of the causes and mechanisms of head and neck cancer. *Oral Dis*, **21**(6), pp. 685-693.

Stan Development Team. (2017). Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0; http://mc-stan.org.

Potential impact fraction for a continuous risk factor: assessing the burden of oral and pharyngeal cancer according to the adherence to the healthy eating index

Matteo Di Maso[a], Laura Tomaino [a], Monica Ferraroni[a], Carlo La Vecchia[a], Valeria Edefonti[a] and Francesca Bravi[a]
[a] Department of Clinical Sciences and Community Health, Branch of Medical Statistics, Biometry and Epidemiology "G.A. Maccacaro", University of Milan, Milan, Italy

1. Introduction

A generalization of the attributable fraction (AF) is the potential impact fraction (PIF) defined as the proportion of disease burden that would occur if the distribution of the risk factor in population was modified according to an alternative (or counterfactual) scenario (Ezzati et al., 2003). For a continuous risk factor X, the PIF is:

$$PIF_{X_{continuous}} = \frac{\int_{x=0}^{m} RR(x)\cdot P(x)\,dx - \int_{x=0}^{m} RR(x)\cdot P'(x)\,dx}{\int_{x=0}^{m} RR(X)\cdot P(X)\,dx}$$

where $RR(x)$ is the relative risk for the continuous risk factor X at exposure level x, $P(x)$ is the population distribution (or prevalence) of X at exposure level x, $P'(x)$ is the counterfactual distribution of X at level x, and m is the maximum exposure level. The denominator represents the exposure-weighted risk of disease in the population under the observed risk factor prevalence and the numerator represents the difference between this risk of disease and one obtained under the counterfactual scenario. Implicitly, the AF methodology assumes the *maximum prevalence reduction* scenario in which the whole population was shifted to the unexposed (or lowest exposure) level. Obviously, if the factor considered has a protective effect the shifting will be toward the highest exposure level. In practice, the *maximum prevalence reduction* scenario is not easily feasible and alternative public health scenarios should be considered. For instance, halve the risk factor prevalence for the whole population or changing the risk factor prevalence for only specific subgroups at higher risk could be more reasonable alternatives.

Using data from an Italian multicentre case-control study, we estimated PIFs for oral and pharyngeal cancer attributable to healthy eating index (HEI) adherence according to 6 counterfactual scenarios.

2. Methods

Data collection: Data came from a multicentre case-control study conducted between 1992 and 2009 in different Italian areas. Cases were 946 patients admitted to major hospitals in the study areas, with incident histologically confirmed cancer of the oral cavity and pharynx. The control group comprised 2492 cancer-free patients frequency-matched to cases by study area, study period, sex, and age. All study participants signed an informed consent in accordance with the recommendations of the Board of Ethics of the study hospitals.

Exposure assessment: Trained interviewers administered a structured questionnaire to collect information on socio-demographic characteristics, well-known and likely risk factors for oral and pharyngeal cancer, dietary habits, as well as relevant confounders. A reproducible and validated food frequency questionnaire (FFQ) was used to assess patients' usual diet (Decarli et al., 1996;

ISBN: 978-88-5495-135-8

Franceschi et al., 1993). The FFQ included information on weekly intake of 78-food (or recipe) items, according to the following sections: (i) milk, hot beverages and sweeteners; (ii) bread, cereals and first courses (including soups); (iii) second courses (e.g., meat, fish, and other main dishes); (iv) side dishes (i.e., vegetables and potatoes); (v) fruit; (vi) sweets, dessert and soft drinks. An additional section collected the lifetime consumption of alcoholic beverages. The FFQ-derived dietary data were used to calculate the HEI-2015 score for each study subject.

The HEI is an energy density-based measure for assessing diet quality, specifically the degree to which a set of foods is in compliance with the 2015-2020 Dietary Guidelines for Americans (DGA) (Krebs-Smith et al., 2018; Reedy et al., 2018). In particular, the HEI is composed of 9 adequacy components (i.e., total and whole fruits, total vegetables, greens and beans, wholegrain, milk/dairy products, total protein foods, seafood and plants, proteins, and fatty acids) and 4 moderation components (i.e., refined grains, sodium, added sugars, and saturated fats). The components are generally scored between 0 (non-compliance) and 10 (highest compliance) except for those further classified in two subcategories (e.g., total and whole fruits) that are scored between 0 and 5. Each component is scored on standard densities (e.g. component amount per 1000 kcal/day) according to DGA cut-offs and therefore the HEI corresponds to a continuous measure defined on the interval $[0,100]$ with higher HEI values indicating greater compliance to the DGA recommendations.

Risk estimation and counterfactual scenarios: We estimated odds ratio (OR) and 95% confidence interval (CI) for oral and pharyngeal cancer risk and HEI score adjusting for matching variables and confounders (i.e., years of education, body mass index, tobacco smoking, alcohol drinking, and total non-alcohol energy intake). We designed the following 6 scenarios: (1) *maximum prevalence reduction* (i.e., assigning the highest HEI value for all subjects); (2) *mild global prevalence intervention* (i.e., 10-point increment in the HEI for all subjects); (3) *strong prevalence intervention on low and lower-middle HEI subjects* (i.e., 1/3 increment in the HEI for subjects in the 1st and 2nd quartile of the HEI distribution); (4) *mild prevalence intervention on low and lower-middle HEI subjects* (i.e., 1/3 increment in the HEI for subjects in the 1st quartile of the HEI distribution and 1/4 increment for subjects in the 2nd quartile); (5) *strong prevalence intervention on lower- and upper-middle HEI subjects* (i.e., 1/3 increment in the HEI for subjects in the 2nd quartile of the HEI distribution and 1/4 increment for subjects in the 3rd quartile); (6) *mild prevalence intervention on lower- and upper-middle HEI subjects* (i.e.; 1/4 increment in the HEI for subjects in the 2nd and 3rd quartile of the HEI distribution).

3. Results

Overall, the HEI ranged from 35.8 to 88.7 with a median of 62.8 (interquartile range: IQR=6.6; figure 1). The OR of oral and pharyngeal cancer for 1-point increment in the HEI was 0.98 (95% CI: 0.97-0.99). The fraction of oral and pharyngeal cancer cases attributable to HEI under the *maximum prevalence reduction* scenario was 42.0%, 95% CI: 25.1%-59.9% (figure 1 top left panel). This figure was 18.8% (95% CI: 8.3%-26.5%) according to the *mild global prevalence intervention* scenario (figure 1 bottom left panel). Whereas the PIF estimates were 18.4% (95% CI: 10.0%-26.1%) and 16.4% (95% CI: 7.4%-23.2%) according to the *strong prevalence intervention on the low and lower-middle HEI subjects* scenario and the *mild prevalence intervention on the low and lower-middle HEI subjects* scenario, respectively (figure 1 top and bottom central panels). The fraction of attributable cases were 15.6% (95% CI: 8.0%-21.6%) under the *strong prevalence intervention on the lower- and upper-middle HEI subjects* scenario and 13.6% (6.6%-18.9%) under *mild prevalence intervention on the lower- and upper-middle HEI subjects* scenario (figure 1 top and bottom left panels).

4. Conclusions

A suitable epidemiological tool to address preventive strategies have to take into account the

ISBN: 978-88-5495-135-8

magnitude of risk factors and their prevalence in the population. The PIF fairly accomplish this because it considers both the risk of disease and the changing in the risk factor prevalence according to a counterfactual scenario. The well-known AF measure is a special case of the PIF under the *maximum prevalence reduction* scenario. Furthermore, the PIF offers the possibility of examining continuous risk factors. In conclusion, the need of epidemiologists and public health officers to translates risk factor prevalence and disease occurrence in useful numbers under more realistic and feasible intervention scenarios makes the PIF to be in a mounting epidemiological importance.

References

Decarli A., Franceschi S., Ferraroni M., et al. (1996). Validation of a food-frequency questionnaire to assess dietary intakes in cancer studies in Italy. Results for specific nutrients. *Ann Epidemiol*, **6**(2), pp.110-118.

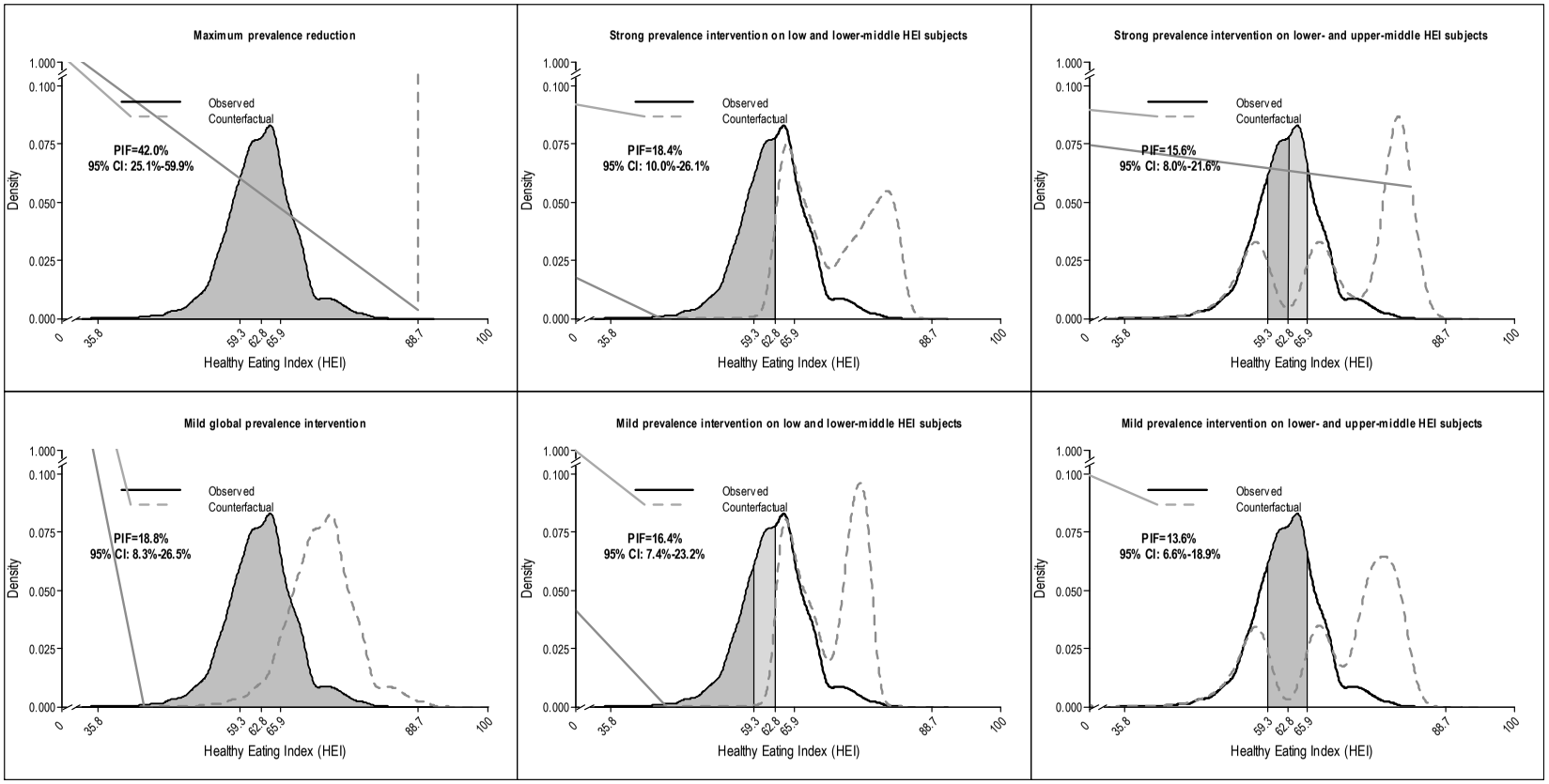
Ezzati M., Hoorn S.V., Rodgers A., et al. (2003). Estimates of global and regional potential health gains from reducing multiple major risk factors. *Lancet*, **362**(9380), pp. 271-280.

Franceschi S., Negri E., Salvini S., et al. (1993). Reproducibility of an Italian food frequency questionnaire for cancer studies: results for specific food items. *Eur J Cancer*, **29A**(16), pp. 2298-2305.

Krebs-Smith S.M., Pannucci T.E., Subar A.F., et al. (2018). Update of the Healthy Eating Index: HEI-2015. *J Acad Nutr Diet*, **118**(9), pp. 1591-1602.

Reedy J., Lerman J.L., Krebs-Smith S.M., et al. (2018). Evaluation of the Healthy Eating Index-2015. *J Acad Nutr Diet*, **118**(9), pp. 1622-1633.

Figure 1: Potential impact fraction (PIF) of oral cavity and pharyngeal cancer cases attributable to healthy eating index (HEI) adherence according to 6 scenarios.



Shadow areas represent target subjects involved in the prevalence changing intervention. Colour gradient represent the magnitude of the prevalence intervention.

CLEUP Coop. Libraria Editrice
ISBN: 978-88-5495-135-8

Comparing statistical models and machine learning algorithms in predicting football outcomes

Leonardo Egidi [a], Nicola Torelli [a]
[a] Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy

1. Introduction

From a statistical point of view, the outcome of a football match may be modelled according to two distinct perspectives. The *goals-based* approach implies modelling, via some suitable count distribution, the number of the goals scored and conceded by the teams competing in each match. In the literature we mainly recognize three types of goals-based Poisson models: double Poisson (Maher, 1982; Baio and Blangiardo, 2010; Groll and Abedieh, 2013; Egidi et al., 2018); bivariate Poisson (Dixon and Coles, 1997; Karlis and Ntzoufras, 2003); Poisson difference/Skellam (Karlis and Ntzoufras, 2009). Once a model has been estimated, the derivation of the so-called *three-way* process (home win, draw, away win) can be simply obtained by aggregating the estimated probabilities. The *results-based* approach consists of modelling directly the three-way process, by use of ordered probit (Koning, 2000) or logit (Carpita et al., 2015, 2019) regression models. This second framework is obviously nested within the first one: the result of a football match is established from the goals scored and conceded, while knowledge of the simple three-way result says nothing about the number of the goals scored by the two teams. In the last years, the widespread popularity of large datasets—provided, for instance, by the Kaggle platform (`https://www.kaggle.com/`)—promoted the use of Machine Learning (ML) tools such as Classification and Regression Trees (CART) and Random Forests as new results-based procedures (Schauberger and Groll, 2018; Groll et al., 2019).

In this paper we develop a comparison between some statistical models and some results-based Machine Learning algorithms, to explore predictive performance for future matches using the results of the FIFA World Cup 2018, hosted in Russia. Although not conclusive, we believe our comparison review may be beneficial for future scholars to discern between goals-based and results-based models.

2. Models and Machine Learning algorithms

21 Results-based

Let $z_n \in \{1, \mathrm{X}, 2\}$ denote the observed categorical result for the n-th match, $n = 1, \ldots, N$, where $\{1, \mathrm{X}, 2\}$ hereafter denotes the three-way process for the home team win, the draw and the away team win, respectively. One way to model this process is using a multinomial logistic regression, where each possible outcome is associated with a probability, in turn modelled with some predictors. Alternatively, we may end up to select some ML procedures (Friedman et al., 2001).

Precisely, in this paper we consider a simple Bayesian multinomial model and five ML algorithms: Random Forest, Classification and Regression Trees (CART), Bagged CART, Multivariate Adaptive Regression Splines (MARS) and Neural Network, according to their standard use as provided by the `caret` package (Kuhn, 2019);

ISBN: 978-88-5495-135-8

22 Goals-based

Goals-based models rely on the assumption that the goals scored by the teams in each match follow a discrete distribution, usually two independent Poisson or a bivariate Poisson accounting for positive correlation. Let (x_n, y_n) denote the observed number of goals scored by the home and the away team in the n-th game, respectively. A general bivariate Poisson model is the following:

$$\begin{aligned}(X_n, Y_n|\lambda_{1n}, \lambda_{2n}, \lambda_{3n}) &\sim \mathsf{BivPoisson}(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}) \\ \log(\lambda_{1n}) &= \theta + \text{att}_{h_n} + \text{def}_{a_n} + \frac{\gamma}{2} w_n \\ \log(\lambda_{2n}) &= \theta + \text{att}_{a_n} + \text{def}_{h_n} - \frac{\gamma}{2} w_n \\ \log(\lambda_{3n}) &= \beta_0,\end{aligned} \tag{1}$$

where the case $\lambda_{3n} = 0$ reduces to the double Poisson model. $\lambda_{1n}, \lambda_{2n}$ represent the scoring rates for the home and the away team, respectively, where: θ is the common baseline parameter; the parameters att_T and def_T represent the attack and the defence abilities, respectively, for each team T, $T = 1, \ldots, N_T$; the nested indexes $h_n, a_n = 1, \ldots, N_T$ denote the home and the away team playing in the n-th game, respectively; the predictor $w_n = (\text{rank}_{h_n} - \text{rank}_{a_n})$ is the difference of the FIFA World Rankings (`https://www.fifa.com/fifa-world-ranking/`)—expressed in FIFA ranking points divided by 10^3—between the home and the away team in the n-th game, multiplied by a parameter $\gamma/2$. This last term tries to correct for the well-known phenomenon of *draw inflation* (Karlis and Ntzoufras, 2003), favouring the draw occurrence when teams are close in terms of their FIFA rankings.

In a Bayesian framework, attack and defence parameters are usually assigned some noninformative prior distributions (Baio and Blangiardo, 2010) and imposed a sum-to-zero constraint to achieve identifiability.

3. Predictive performance

An usual way to compare statistical models and ML algorithms relies on predictive accuracy on out-of-sample data. We consider here the dataset containing the results of all the 64 tournament's matches (48 of the group stages, and 16 of the knockout stage) for the FIFA World Cup 2018 hosted in Russia. The value of the FIFA ranking difference w incuded in the models was considered on June 7th, only a bunch of days before the tournament takes place.

However, the choice of the *training set* and the *test set* is of crucial importance and is likely to affect the predictions. We decided to train our statistical models/ML techniques on distinct portions of matches from the group stage, where teams are more heterogeneous in terms of their FIFA rankings and actual strenghts. To assess predictive performance between statistical models and ML algorithms in predicting football outcomes, we compare the double Poisson and the bivariate Poisson model (goals-based), fitted by `rstan` package (Stan Development Team, 2018), with the Bayesian multinomial model (`rstan`) and the five ML procedures listed in the previous section (results-based).

As motivated above, we propose three different prediction scenarios (train and test sets are separated by the symbol $\rightarrow$):

1. 75% of randomly selected group stage matches $\rightarrow$ Remaining 25% group stage matches
2. Group stage matches $\rightarrow$ Knockout stage

3. Group stage matches for which both the teams have a Fifa ranking greater than 1 → Knockout stage.

Train	75% group	100% group	rank > 1
Test	25% group	knockout	knockout
Random forest	0.67	0.25	0.75
Bagged CART	0.67	0.31	0.75
CART	0.58	0.31	0.50
MARS	0.58	0.38	0.50
NN	0.67	0.25	0.75
Multinomial	0.42	0.62	0.62
Double Pois.	0.58	0.56	0.56
Biv. Pois.	0.58	0.50	0.50

Table 1: Prediction accuracy for the selected methods, according to three prediction scenarios.

Table 1 shows the accuracy in the predictions for the eight methods and the three scenarios. Random Forest, Bagged CART and Neural Networks perform quite well in the first and in the third scenario, whereas the multinomial model performs better in the second scenario. Goals-based statistical models tend to perform generally worse than results-based procedures in each of the considered scenarios; however, their performance definitely dominates the five ML algorithms when the training set is the largest (second scenario). As already argued, the choice of the training and the test set can dramatically change the predictive performance of the alternative models. It is worth noting that a statistical model seems to better learn from the entire group stage dataset (second column in Table 1), while ML algorithms clearly over-perform statistical models only when considering the specific sub-sample of the strongest teams that are more likely to go further in the competition (third scenario). Statistical models could therefore be adopted when the teams involved in a competition are more heterogeneous.

Obviously it should be noted that the results presented here are only preliminary also considering the small size of the dataset here analysed. A full appreciation of the different performances is out-of the scope of the current paper and should be definitely considered for future research, also aimed at defining a strategy for stacking models in this specific application case.

References

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

Carpita, M., Ciavolino, E., and Pasca, P. (2019). Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling*, 19(1):74–101.

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

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Egidi, L., Pauli, F., and Torelli, N. (2018). Combining historical data and bookmakers' odds in modelling football scores. *Statistical Modelling*, 18(5-6):436–459.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

ISBN: 978-88-5495-135-8

Groll, A. and Abedieh, J. (2013). Spain retains its title and sets a new record–generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports*, 9(1):51–66.

Groll, A., Ley, C., Schauberger, G., and Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports (Ahead of print)*.

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

Karlis, D. and Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, 20(2):133–145.

Koning, R. H. (2000). Balance in competition in dutch soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):419–431.

Kuhn, M. (2019). *caret: Classification and Regression Training*. R package version 6.0-84.

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

Schauberger, G. and Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, 18(5-6):460–482.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2.

The effects of attitude towards Statistics and Math knowledge on Statistical anxiety: A path model approach

Rosa Fabbricatore[a], Carla Galluccio[a], Cristina Davino[b], Daniela Pacella[a], Domenico Vistocco[a], Francesco Palumbo[a]
[a] Department of Political Science, Università di Napoli Federico II;
[b] Department of Economics and Statistics, Università di Napoli Federico II.

1. Introduction

Academic well-being is an important task to achieve at all educational stages. Specifically, the well-being of university students and their academic performances are negatively affected by stress and anxiety. The discomfort that some students feel in regard to Math or Statistics has a great impact on them. In literature this is referred to as statistical anxiety (SA). SA can be defined as *"the feeling of anxiety encountered when attending a statistics course or doing statistical analyses"* (Cruise, Cash & Bolton, 1985, p.92). In particular, since Statistics has been introduced in many university curriculum programs, including many humanities courses, such as psychology, political sciences or sociology, in recent years statistical anxiety has been widely studied. Students who attend non-mathematics programs consider Statistics as a burd (Sesé, Jiménez, Montaño & Palmer, 2015) and exhibit higher SA levels. They are made weary by anything related to mathematics and believe that Statistics is not important for their degree programs and careers. SA negatively affects students' statistics examinations: higher levels of SA lead them to lower performance (Galli, Chiesi & Primi). Due to SA great impact on students' academic well-being and performance, several studies have focused on this topic and classified SA antecedents into situational factors (e.g. math skills, previous statistical experience), dispositional factors (e.g. attitude toward statistics, self-concept and self-efficacy) and demographic factors (e.g. gender, age). In a study with undergraduate psychology students Chiesi and Primi (2010) examined the students' grade of achievement in a preliminary statistics course taking into account cognitive factors (e.g. math background) as well as non-cognitive factors (e.g. attitude towards statistics measured before and after the introductory course) and SA. They showed that SA directly depends on math knowledge and pre-course attitude, which, in turn, influence post-course attitude along with SA. Finally, students' grade of achievement was explained by post-course attitude and math knowledge. Sesé et al. (2015) confirmed the influence of the math background on SA and attitude towards Statistics. On the other hand, the math background only had an indirect effect on students' performance, through attitude. Moreover, several authors showed that gender was one of the most important demographic variables: some studies have highlighted that females experienced higher level of SA than males (see among the others Baloğlu, Deniz & Kesici, 2011). However, other authors did not report any significant differences between genders (e.g. see Baloğlu, 2003). To measure the construct of SA, the Statistical Anxiety Rating Scale (STARS) was used. This psychometric instrument is one of the most common and widely used tool to assess SA (Cruise et al., 1985).

Data collected in the context of the ALEAS (Adaptive LEArning in Statistics; https://aleas-project.eu/wordpress/) ERASMUS+ project were used in this work to explore the antecedents of SA in undergraduate students enrolled in the psychology course at Federico II University of Naples. In addition to the variables discussed above, in this study we also considered as antecedents of SA the high school final mark and the past experience with Statistics.

ISBN: 978-88-5495-135-8

In particular, we tested the following hypotheses:

Hypothesis 1. Gender, high school final mark, math comprehension and math background affect both the attitude towards Statistics and the levels of statistical anxiety;

Hypothesis 2. Past experience with Statistics and attitude towards Statistics predict statistical anxiety.

2. Material and methods

Participants and procedure

The participants were N = 100 undergraduate students, enrolled in the psychology course at Federico II of Naples and thus involved in an introductory statistics course. Sample's age ranged from 18 to 27 (mean = 19.53, sd = 1.4). The students involved were predominantly female (81%) and came from different types of high schools (24% scientific studies, 38% humanistic studies, 38% others). High school final mark's median was 86 over 100. At the beginning of the course researchers administered the questionnaire in the classroom and answers were collected in a paper-and-pencil form.

Questionnaire

The questionnaire was structured in three sections. The first included questions about demographic variables, type of degree, high school final mark, math comprehension, and past experience with Statistics. Math comprehension was evaluated using the single 5-point (1 = 'Strongly disagree' to 5 = 'Strongly agree') Likert-type item '*During math lessons I can understand even the most difficult concepts*', whereas past experience with Statistics was measured asking students if they were ever enrolled in a statistical course before (dichotomous item). In the second section attitude towards Statistics and SA were assessed using the STARS (Cruise et al., 1985). Students answered on a 5-point Likert scale ranging from 1 = 'Strongly disagree' to 5 = 'Strongly agree' (for attitude) and from 1 = 'No anxiety' to 5 = 'Very high anxiety' (for anxiety). Both subscales scored Cronbach's alpha = 0.92. The last section, aimed at evaluating the students' math background, included twenty multiple choice questions; 7 of these, concerning operations and set theories, were selected from the scale for Mathematical Prerequisites for Psychometrics (Galli et al., 2008); the others, about relations and fractions, were selected from university entrance exams. Cronbach's alpha was 0.61.

Statistical analysis

Statistical analyses were performed using the R statistical software. Since Cronbach's alpha ensured the internal consistency reliability of the measures, each variable was defined as the score sum of the corresponding item set. Math background scores were computed using 2PL IRT model (Bartolucci, Bacci & Gnaldi, 2015). To test our hypotheses path analysis based on maximum likelihood estimation (Duncan, 1966) was carried out using the `lavaan` package. Goodness-of-fit was evaluated using the following fit indices: the Chi-square (χ^2), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR) [1].

[1]Good model fit was defined by the following criteria: RMSEA values of about 0.08 or below, SRMR values less than about 0.08, CFI values of about 0.95 or above, and TLI values above about 0.90.

3. Results and discussion

Participants showed, in general, medium-low levels of SA, which are mainly related to subscale "Test and Class", and little negative attitude towards Statistics, as shown in Table 1. It is worth noting that since the STARS is not a diagnostic scale there is no threshold value to discriminate pathologically high levels of anxiety. About math comprehension, only 14 out of 100 stated that they can understand even the most difficult concepts during Math lessons. Finally, most participants (85%) claimed that they had never studied Statistics before.

STARS	No. of item	Range	Q1-Q3	M (SD)	M/No. of item
Anxiety					
Interpretation	11	11-55	20-30	25(7.9)	2.27
Test and Class	8	8-40	24-32	27.4(6.3)	3.42
Fear of Asking for Help	4	4-20	7-10	8.6(2.9)	2.15
Attitude					
Worth of Statistics	16	16-80	36-45	40.7(10.23)	2.54
Computation Self-Concept	7	7-35	17-23	20.55(5)	2.94
Fear of Statistics Teachers	5	5-25	10-15	12.84(3.6)	2.57

Table 1: Quartiles, Means and Standard Deviations of STARS subscale scores.

Our hypothesized model is shown in Figure 1 including standardized regression coefficients. All fit indices pointed to a good fit of the model: χ^2 = 1.083 ($p = 0.3$), CFI = 0.999, TLI = 0.986, RMSEA = 0.029, SRMR = 0.017. As it can be seen in Figure 1, results show that math background not affect attitude towards Statistics nor statistical anxiety. Math comprehension has a negative impact on both attitude towards Statistics ($\beta = -0.45$, $p < 0.01$) and SA ($\beta = -0.34$, $p < 0.01$), whereas high school final mark variable only affects SA ($\beta = 0.20$, $p < 0.05$). Contrary to our hypotheses, past experience with Statistics and gender have no significant effect on SA.

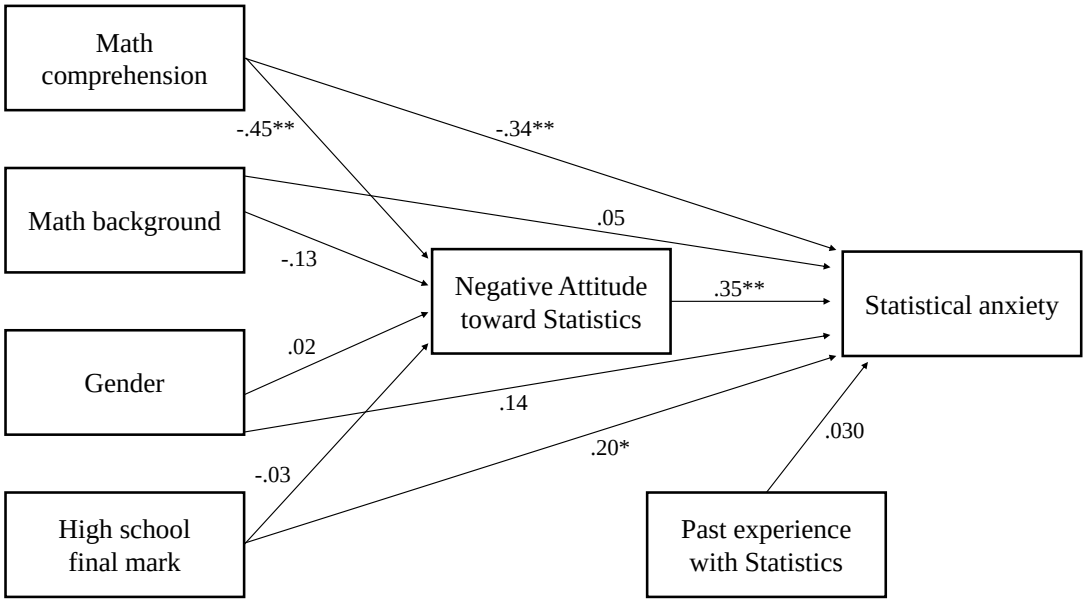


Figure 1: Path analysis diagram with standardized regression coefficients. *$p < 0.05$; **$p < 0.01$.

Our results confirmed evidences from the literature. Believing that Statistics is useless, not feeling comfortable with studying it and not being sure of one's mathematical knowledge are all factors that increase the SA. Negative attitude towards Statistics was the most influential antecedent of SA, followed by math comprehension. The results with regard to gender and math background did not confirm our hypotheses: females did not experience higher levels of statistical anxiety than males, and lack of math background did not affect attitude towards Statistics nor SA. Moreover, what is interesting is the effect of the high school final mark on SA regardless of the type of high school attended: in fact, SA increases according to the high school final mark. We believe that clever students have generally higher levels of anxiety, regardless of the subject taken into consideration, and this may be related to their own perfomance expectation. This relationship should be further studied with a larger sample to obtain more robust estimates. We believe, however, that the number of subjects participating in the project will get higher, which will increase the size of the collective, thus increasing the efficiency of our estimates. Finally, the relevance of the effect of SA on academic performance and well-being led researchers to explore the efficacy of different teaching techniques on reduction of students' SA. Among these, there are the use of real-life data, active learning activities, and humorous cartoons (Lesser & Pearl, 2008). Also ALEAS ERASMUS+ project takes part in this scenario. In fact, the aim of the project is to provide students with a technological platform for self-guided learning of statistics, in order to offer personalised learning paths so as to influence their SA in a positive way.

References

Baloğlu, M. (2003). Individual differences in statistics anxiety among college students. *Personality and Individual Differences*, **34**(5), pp. 855–865.

Baloğlu, M., Deniz, M. E., & Kesici, Ş. (2011). A descriptive study of individual and cross-cultural differences in statistics anxiety. *Learning and Individual Differences*, **21**(4), pp. 387–391.

Bartolucci, F., Bacci, S., & Gnaldi, M. (2015). *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Chapman and Hall/CRC, Boca Raton (FL).

Chiesi, F., & Primi, C. (2010). COGNITIVE AND NON-COGNITIVE FACTORS RELATED TO STUDENTS'STATISTICS ACHIEVEMENT. *Statistics Education Research Journal*, **9**(1), pp. 6–26.

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *American Statistical Association Proceedings of the Section on Statistical Education*, **4**(3), pp. 92–97.

Duncan, O. D. (1966). Path analysis: Sociological examples. *American journal of Sociology*, **72**(1), pp. 1–16.

Elias, H., Ping, W. S., & Abdullah, M. C. (2011). Stress and academic achievement among undergraduate students in Universiti Putra Malaysia. *Procedia-Social and Behavioral Sciences*, **29**, pp. 646–655.

Galli, S., Chiesi, F., & Primi, C. (2008). The construction of a scale to measure mathematical ability in psychology students: An application of the Rasch Model. *TPM (Testing Psicometria Metodologia)*, **15**(1), pp. 1–16.

Lesser, L. M., & Pearl, D. K. (2008). Functional fun in statistics teaching: Resources, research and recommendations. *Journal of Statistics Education*, **16**(3), pp. 1–11.

Sesé, A., Jiménez, R., Montaño, J. J., & Palmer, A. (2015). Can attitudes towards statistics and statistics anxiety explain students' performance. *Revista de Psicodidáctica*, **20**(2), pp. 285–304.

Personal and familial determinants of gambling risk among adolescent Italian students

Luigi Fabbris[a], Alessandra Andreotti[b], Bruno Genetti[b], Paolo Vian[b], Claudia Mortali[c], Luisa Mastrobattista[c], Adele Minutillo[c], Roberta Pacifici[c]

[a] Department of Statistical Sciences, University of Padua, Italy.

[b] Explora Center for Research and Statistical Analysis, Vigodarzere (PD), Italy.

[c] Centro Nazionale Dipendenze e Doping, Istituto Superiore di Sanità, Rome, Italy.

1. Introduction

Gambling affects both adult and adolescent populations worldwide. The increasing trend of gambling practice among adolescents represents a social concern in as much as gambling may constitute a ground-breaking occurrence of an adult dependence. The Italian literature examining circumscribed samples of adolescents and young adults discovered relationships between individual psychological characteristics, gambling behaviours, drinking attitudes and substance use (Canale et al., 2015; Buja et al., 2017). This literature is generally consistent with international mainstream literature which suggests the progressive diffusion and complexity of the phenomenon of gambling, whose origin may be grounded in children's playing and which could interact, in particular during adolescence, with other transgressive habits of youth.

This paper discusses the social and personal dimensions that may influence the risk of youth gambling and highlights attitudinal, behavioural, family- and peer-related characteristics that may influence gambling beyond situational and demographic descriptors.

2. Data and methods

We analysed the data collected in a national survey representative of Italian students aged 14 to 17 conducted in 2017. The sampling procedure followed a three-stage probability proportional to size (PPS) structure with stratification at the first, second and third stage units. The first sampling stage was that of municipalities; the second stage was that of schools within sampled municipalities; and the third stage was that of classes within sampled schools. All students aged 14 to 17 of the sampled classes were included in the sample. The survey was conducted through a computer-assisted web-interview (CAWI) technique (with the presence of the data collector in the classroom), using a nonreplicable, unique and anonymous access identification system.

The students' propensity to gamble was assessed through an Italian-validated version of the South Oaks Gambling Screen-Revised Adolescent (SOGS-RA) scale (Winters et al., 1993; Colasante et al., 2014), which identifies three levels of gamblers: social, at-risk, and problem gamblers. In order to determine the possible relationships between students' propensity to gamble, on the one hand, and students' attitudes and behaviours and various external stimuli, on the other, we applied a multilevel ordinal logistic model (McCullagh, 1980; Agresti, 2002). The outcome variable, *Y*, represented the level of gambling of the concerned students. It could assume four levels: 0, which corresponds to non-gamblers; 1, corresponding to social gamblers; 2, for at-risk gamblers; and 3, for problem gamblers. The predictors considered for the analysis were clustered into blocks, as follows: 1) *control variables*, that is, the nonmodifiable personal background variables, such as gender, age, nationality and geographical area; control variables were forced into the model independent of their statistical significance; 2) *attitudes and behaviours of students*, which includes the personality traits and the comorbidity descriptors; 3) *family lifestyle exposure*, which includes the family context and the family gambling culture and experience; and 4) *peer group exposure*, which includes certain variables concerning friends, such as the presence of gambler friends and the presence of problem gambler friends.

ISBN: 978-88-5495-135-8

The model sought to account for the hierarchical structure of the data considering the attitudes and behaviours of students (first-level variables), the family context (second-level variables), and peer group exposure (third-level variables). Given that *Y* is ordinal, the probability of outcome *Y* for respondent *h*, $P(y_h \leq i \mid \boldsymbol{X}_h)$ (h=1, …, n), with n being the student sample size, can be estimated with a linear random-intercept model with covariates (Goldstein, 2010):

$$P(y_h \leq i|\boldsymbol{X}_h) = \gamma_0 + \sum_j^3 \gamma_j + \beta_0 \, \boldsymbol{z}_h + \sum_j^3 \boldsymbol{\beta}_j \, \boldsymbol{x}_h + \varepsilon_h \quad (h = 1, \dots, n) \tag{1}$$

where: γ_0 is the intercept across all students; γ_j is the random effect associated with the hierarchical level j (j = 1, .., 3); β_0 is the measure of the association between *Y* and the control variables; $\mathbf{z}_h$ is a vector of control variables observed at individual *h*; $\boldsymbol{\beta}_j$ is a vector of parameters posited to measure the association between the variables at level j and the criterion variable; $\boldsymbol{x}_h = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3]$ is a vector of predictors inherent to the three levels of the social structure, and ε_h stands for the sum of the residuals at all levels, both the individual one and the higher level ones, $\sum_j^3 \gamma_j$. The analyses were performed using the statistical software R (R Core Team, 2019).

3. Results

The analysed data comprised 15,602 students, attending 201 secondary schools in 97 municipalities; 49.1% of students were male and 50.9%, female, with a mean age equal to 15.5 years. The percentage of gamblers aged 14 to 17 years in the sample was 28.2%. Of these, 77% were social gamblers, 12.5% were at-risk gamblers, and 10.5% were problem gamblers. Out of the overall sample, 21.7% were social gamblers, 3.5% were at-risk gamblers, and 3.0% were problem gamblers. The multivariate model (Table 1) including only the intercept and the control variables revealed an R^2 value of 0.068, while for the final model, R^2 = 0.116. The analysis revealed that attitudes of students and certain family-related stimuli together influenced the propensity to gamble beyond the biological and situational characteristics of youngsters. In particular, all the control variables were significant for at least the social gambling, even when all other predictors were considered in the model.

*Being male w*as strongly significant at all gambling grades. These results are in line with the literature (Kang et al., 2019; Andrie et al., 2019). The odds ratio (OR) of being a social gambler for males was 3.5 times that of female counterparts. Furthermore, males were 3.2 times and 3.7 times more likely than females to be at-risk and problem gamblers, respectively. When these values were compared with those obtained from simply crossing gender with gambling grades, the univariate risk of being a social gambler decreased (OR = 2.7) and that of being at-risk or problem gamblers increased (OR = 6.4 and 8.1, respectively) than values including multivariate estimates. This means that being a male significantly interacted with personal and family characteristics of students in determining the possible grades of gambling risk, although the direction of interaction was heterogeneous. Another relevant variable was *age*. Even if all students were adolescents of approximately the same age, the older the students, the higher were the risks. This result suggested that the longer people stayed exposed to the risk of gambling, the more likely they were to gamble, although the risk of becoming a problem gambler mildly depended on social exposure. *Residence area* was also important for all gambling grades. In the Centre and South of Italy, students appeared more prone than North-West counterparts, and in the North-East of Italy, they appeared to be less prone than North-West counterparts. This result highlights the social origin of gambling at all exposure grades. Finally, *nationality* - which also indicates the social origin of risk - was significant but only for social gambling. Such an observation may mean that social gambling has roots in social practices that stem from group culture.

ISBN: 978-88-5495-135-8

Table 1: Ordinal logistic model results

Variable	*Social gambler*	*At risk gambler*	*Problem gambler*	*%*
Intercept	-4.11 ***	-7.081 ***	-8.145***	-
Male *vs. Female*	1.25 ***	2.064 ***	2.107***	49.1
16-17 years *vs. 14-15 years*	0.29 ***	0.288 ***	0.216*	51.1
Nationality Italian *vs other*	0.212 *	0.004	-0.240	94.8
Area North East *vs. "North West"*	-0.295 ***	-0.010	0.005	19.9
Area Centre *vs. "North West"*	0.202 **	0.615 ***	0.700***	18.7
Area South & Islands *vs. "North West"*	0.259 ***	0.650 ***	0.836***	39.3
School delay 1 year *vs. none*	0.131 *	0.226 *	0.265*	11.0
School delay 2 years or more *vs. none*	0.016	0.290	0.199	2.1
"I'm a loser" – very adequate *vs. other*	-0.127	0.242	0.476*	2.7
Act with impulse of moment–not adequate *vs. other*	0.731 ***	0.961 ***	1.016**	9.3
Have concentration problems: very adequate *vs. other*	0.012	0.484 ***	0.726***	13.8
Goes out at night on weekdays *vs. does not go out*	0.586 ***	0.60 ***	0.651***	50.6
Uses coffee almost every day *vs. sometimes, never*	0.403 ***	0.314 ***	0.206 *	39.5
Uses vitamins almost every day *vs. sometimes, never*	0.135 *	0.312 *	0.239	8.8
Plays sport 2-3 times/week *vs. <2 times per week*	0.100 *	-0.099	-0.099	39.9
Plays sport 4+ times/week *vs. <2 times per week*	0.375 ***	0.111	0.073	36.1
Uses smartphone for Internet use *vs does not use*	0.561 ***	0.314 **	0.367 *	85.5
Uses tablet for Internet use *vs does not use*	0.310 **	0.056	0.102	6.4
Know the mother education level *vs does not know*	0.151	0.327 *	0.307	92.5
Father primary education level *vs. more than primary*	0.348 **	0.555 ***	0.548 *	3.0
Father secondary education *vs. other than secondary*	0.223 ***	0.048	0.104	29.1
Father high education level *vs. other than high*	0.146 **	-0.066	-0.057	39.6
Credit card use *vs. does not use*	0.393 ***	0.279***	0.324***	27.9
Very dissatisfied of relationship with siblings *vs. other*	0.304	0.620**	0.403	1.4
Mother played/plays *vs. didn't play or does not play*	0.745 ***	0.754***	0.690***	4.7
Does not ask parents for doubts gambling *vs. asks*	-0.364 ***	0.149 *	0.393**	67.8

Significance: *** p-value ≤ 0.001, ** p-value ≤ 0.01, * p-value ≤ 0.05.

Concerning students' attitudes, the *difficulty to master own impulses* correlated with gambling at all grades and the *feeling of being a loser* also correlated with being a problem gambler. The awareness of a student's incapacity to dominate the stimulus to gamble, whose consequences are known to be systematically negative, was admitted as if it were an external force to which the student could not offer resistance. The feeling of being a loser appeared to be more a consequence than a cause of gambling: If this feeling could be analysed in depth, we could understand if this sensation came from the awareness of the inability to excel in other fields, where gambling becomes a sort of provisional refuge against unhappiness. Some behaviours, such as *concentration problems, school failures*, and *use of smartphone for Internet access* accompanied gambling practices. Others indicated an adolescent search for excess, such as *going out at night at weekends, daily use of coffee and vitamin supplement*. The literature associates gambling with the use of alcohol and substances; however, we did not find any corroboration of this association. It may be conjectured, though, that socially rejected behaviours would not come to light through a direct questionnaire, which suggests that the 'light' behaviours we discovered may indeed mask harder abuses. Surprisingly, *being active in sports* correlated positively with social gambling. This result may mean that sports was a way for youngsters to spend time together with peers. If true, it indirectly implies that even social gambling was a way for young people to socialise. The family characteristics

ISBN: 978-88-5495-135-8

correlated with gambling at all grades were the following: *the possibility given to students to use credit cards, the involvement of their mother in gambling practices* and *the questioning of parents for doubts about gambling*. The first two variables revealed that family disposition towards gambling was transferred to offspring. The latter one, revealed that parents were not a counselling reference for youth in serious difficulties. Other family descriptors correlated with social and at-risk categories are *the (low) education of father, problematic situations with siblings* and *the knowledge of the mother's educational level*.

4. Discussion

Male propensity to gamble displayed significant and large enough differences compared to females so as to suggest consideration of separate and gender-based models in order to improve the model fit with the available data. In addition, age seems distinctive enough to suggest consideration of estimating a separate model for the age 16 and 17, possibly crossed with gender, as compared with that of 14 and 15. One result appears particularly relevant to us is the social origin of gambling. Future analyses should consider leaving out social origin altogether in order to allow more meaningful correlations to be accounted for. We also demonstrated that social gambling was a distinct problem from at-risk or problem gambling. This may mean that future analyses should focus more on the two latter categories, including social gambling with the absence of problems.

References

Agresti, A. (2002). *Categorical Data Analysis*, Second edition. Hoboken, NJ: Wiley Interscience.

Andrie, E.K., Tzavara, C.K., Tzavela, E., Richardson, C., Greydanus, D., Tsolia, M. and Tsitsika, A.K. (2019). Gambling involvement and problem gambling correlates among European adolescents: Results from the European Network for Addictive Behavior study. *Social Psychiatry and Psychiatric Epidemiology*: 1-13 (doi: 10.1007/s00127-019-01706-w).

Buja, A., Lion, C., Scioni, M., Vian, P., Genetti, B., Vittadello, F., Sperotto, M., Simeoni, E. and Baldo, V. (2017). SOGS-RA gambling scores and substance use in adolescents. *Journal of Behavioral Addictions*, **6**(3), pp. 425-433.

Canale, N., Vieno, A., Griffiths, M.D., Rubaltelli, E. and Santinello, M. (2015). Trait urgency and gambling problems in young people: The role of decision-making processes. *Addictive Behaviors*, **46**, pp. 39-44.

Colasante, E., Gori, M., Bastiani, L., Scalese, M., Siviliano, V. and Molinaro, S. (2014). Italian adolescent gambling behaviour: Psychometric evaluation of the South Oaks Gambling Screen Revised for Adolescents (SOGS-RA) among a sample of Italian students. *Journal of Gambling Studies*, **30**(4), pp. 789-801.

Goldstein, H. (2010). *Multilevel Statistical Models*, Fourth edition, Chichester, West Sussex: Wiley.

Kang, K., Ok, J.S., Kim, H. and Lee, K.S. (2019). The gambling factors related with the level of adolescent problem gambler. *International Journal of Environmental Research and Public Health*, **16**(12) (doi:10.3390/ijerph16122110).

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (Methodological)*, **42**(2), pp. 109–142.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Winters, K.C., Stinchfield, R.D. and Fulkerson, J. (1993). Toward the development of an adolescent gambling problem severity scale. *Journal of Gambling Studies*, **9**(1), pp. 63-84.

Winters, K.C., Stinchfield, R.D. and Kim, L.G. (1995). Monitoring adolescent gambling in Minnesota. *Journal of Gambling Studies*, **11**(2), pp. 165-183.

ISBN: 978-88-5495-135-8

A functional data analysis of Google Trends on health and wellness

Francesca Fortuna[a], Giulia Caruso[a], Tonio Di Battista[a]
[a]Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University G. D'Annunzio of Chieti-Pescara, Italy

1. Introduction

Nowadays Internet search engines, represent a popular source of information. Through Google Trends, instead, it is possible to analyze the popularity of search queries in Google Search across different regions and languages, identifying both spatial and temporal patterns (Nuti et al, 2014).

Google is often used also to obtain information related to health and wellness (Vance et al, 2009, Lambert and Loiselle, 2007, Pandey et al, 2013). In particular, a topic of primary importance in this framework is represented by the theme of work-life balance, that is the relation between work and the rest of life. It owes its importance to the significant impact on both the physical and mental health of the individual and the community (Haar et al., 2014; Bohle et al, 2004).

For this reason, we decided to explore the functionalities of Google Trends using work-life balance as keywords.

Google Trends allows to interact with Internet search data, providing insights into population behavior and health-related phenomena (Nuti et al., 2014). We analyzed Google Trends data in a functional framework (Ramsay and Silverman, 2005), mostly because this approach yields an efficacious statistical analysis in those cases in which the number of variables is higher than the number of observations, as in the case of search queries (Fortuna, et al., 2018).

This paper is organized as follows. In Section 2 we discuss Google Trends data in a functional framework, whereas in Section 3 we expose the functional unsupervised clustering methods that can be employed, namely the k-means algorithm and the hierarchical technique (Caruso et al, 2018, Caruso et al, 2019, Caruso and Gattone, 2019, MacQueen, 1967).

Finally, in Section 4, we synthetize some conclusions.

2. Clustering of Google Trends data in a functional framework

Google Trends provides real time trend data, regarding interest as operationalised by Internet search volume. It shows how often search terms are entered in Google, with regards to the total search volume over time, since 2004, and across different geographical locations. Thus, search query index represents a relative search volume index, which is normalized by the highest query share of a specific keyword, over the time series. The normalization process returns an index between 0 and 100 (Choi and Varian, 2012).

Since Google Trends data continuously flow from the server of a web site, they can be seen as functions in a continuous domain, rather than scalar vectors (Fortuna et al., 2018). Despite the continuous nature of functional data, in real applications, sample curves are observed with error in a discrete set of sampling points of the domain. Specifically, let $y_j(t_{jl})$, j=1,…,n; l=1,…,L, be a functional variable observed in a discrete set of l sampling points in the temporal domain T . Let us also assume that $y(t) \in L^2(T)$, where $L^2(T)$ is the Hilbert space of square integrable functions. One usual solution to reconstruct the functional form of the n samples, starting from the discrete observations, is to assume that sample paths belong to a finite-dimension space, spanned by a basis, so that they can be expressed as follows:

ISBN: 978-88-5495-135-8

$$y_j(t) = \sum_{b=1}^{B} a_{jb}\varphi_b(t), \qquad j = 1, \dots, n$$

where a_{jb} is the b-th basis coefficient for the j-th functional observation $y_j(t)$, and $\varphi_b(t)$ represents the b-th basis functions.

Functional queries may present a variety of distinctive patterns corresponding to different shapes and variation, which can be identified by clustering the functions (Sangalli et al., 2010; Fortuna et al., 2018; Tarpei, 2007). Specifically, a set of homogeneous clusters in L^2 can be identified by determining a partition of the space according to the minimal distance. To this end, an L^2 metric in function space has been applied, combined with a k-means algorithm for finite dimensional data (Forgy, 1965). The k-means clustering algorithm (Forgy, 1965; MacQueen, 1967) is an iterative procedure, which alternates a step of centroid calculation, in which a relevant functional representative (the centroid) for each cluster is identified, and a step of cluster assignment, in which all curves are assigned to a cluster, and in particular to the cluster whose centroid is nearer according to a specific distance.

Specifically, the k-means algorithm finds a partition of the functional space into K clusters, by minimizing the sum of squared error criterion between the cluster center and the functions belonging to the cluster, as follows (Jain and Dubes, 1988; Tan and Witten, 2015; Caruso et al., in print; Caruso et al., 2019):

$$\begin{aligned} J(C) &= \min \sum_{k=1}^{K} \sum_{y_j(t) \in C_k} d^2\left(y_j(t), \varphi_k(t)\right) \\ &= \min \sum_{k=1}^{K} \sum_{y_j(t) \in C_k} ||y_j(t) - \varphi_k(t)||^2 \\ &= \min \sum_{k=1}^{K} \sum_{y_j(t) \in C_k} \left(\int |y_j(t) - \varphi_k(t)|^2 dt \right)^{\frac{1}{2}} \end{aligned}$$

where the centroid is computed by averaging the functions across the replications (Ramsay and Silverman, 2005):

$$\varphi(t) = \sum_{y_j(t) \in C_k} \frac{y_j(t)}{n_k}$$

Where n_k is the number of functions in the k-th cluster, with $\sum^{K}_{k=1} n_k = n$.

3. Concluding remarks

This paper explores the functionalities of Google Trends using work-life balance as keywords. In this context, our main aim is to provide practitioners with additional methodological tools for the analysis of Google trends data. Specifically, the use of the functional approach allows to deal with the high dimensionality of these data and to identify functional clusters, able to reflect the dynamics of the search process over the whole temporal domain.

References

Bohle P., Quinlan M., Kennedy D., Williamson A. (2004). Working hours, work-life conflict and health in precarious and "permanent" employment. *Rev. Saúde Pública* 38 suppl. São Paulo.

Caruso G., Di Battista T. and Gattone S.A. (in print). A micro-level analysis of regional economic activity through a PCA approach. In Decisions economics: Complexity of decisions and decisions for complexity; Bucciarelli E., Chen S., Corchado J. M. Eds.; Springer International Publishing, Advances in Intelligent Systems and Computing.

Caruso G., Gattone S.A., Fortuna F., Di Battista T. (2018). Cluster Analysis as a Decision-Making Tool: A Methodological Review. In Decision Economics: In the Tradition of Herbert A. Simon's Heritage; Bucciarelli, E., Chen, S., Corchado, J. M., Eds.; Springer International Publishing, Advances in Intelligent Systems and Computing; **618**, pp. 48-55.

Caruso G., Gattone S.A., Balzanella A., Di Battista T. (2019). Cluster analysis: an application to a real mixed-type data set. In Models and Theories in Social Systems; Flaut, C., Hošková-Mayerová, Š.,Flaut, D., Eds.; Springer International Publishing, Studies in Systems, Decision and Control; 179, pp. 525-533.

Caruso G., Gattone S.A. (2019). Waste management analysis in developing countries through unsupervised classification of mixed data. Social sciences.

Choi H.Y., Varian H. (2012). Predicting the Present with Google Trends. *Economic Record*, **88**, pp. 2-9.

Forgy, E. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, **21**, pp. 768-769.

Fortuna, F., Maturo, F., Di Battista, T. (2018). Clustering functional data streams: Unsupervised classification of soccer top players based on Google trends. *Quality and Reliability Engineering International*, **34**(7), pp. 1448-1460.

Haar J. M., Russo M., Suñe A., Ollier-Malaterre A. (2014). Outcomes of work–life balance on job satisfaction, life satisfaction and mental health: A study across seven cultures, *Journal of Vocational Behavior*, **85** (3), pp. 361-373.

Jain, A., Dubes, R. (1988) Algorithms for Clustering Data. Englewood Cliffs, New York: Prentice Hall.

Lambert, S. D., Loiselle, C. G. (2007). Health Information-Seeking Behavior. *Qualitative Health Research*, **17**(8), pp. 1006–1019.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability; University of California Press.

Nuti, S.V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R.P., Chen, S.I., et al. (2014). The Use of Google Trends in Health Care Research: A Systematic Review. PLoS ONE 9(10): e109583, https://doi.org/10.1371/journal.pone.0109583.

Pandey A., Hasan S., Dubey D., Sarangi S. (2013). Smartphone Apps as a Source of Cancer Information: Changing Trends in Health Information-Seeking Behavior. *Journal of Cancer Education*, **28**(1), pp. 138-142.

Ramsay J.O., Silverman B.W.(2005). *Functional Data Analysis*, Springer, 2nd edition

Sangalli, L., Secchi, P., Vantini, S., Vitelli, V. (2010). K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, **54**, pp. 1219-1233.

Tan, K., Witten, D. (2015). Statistical properties of convex clustering. *Electronic Journal of Statistics*, **9**, pp. 2324-2347.

Tarpey T. (2007). Linear transformations and the k-means clustering algorithm: applications to clustering curves. *Journal of the American Statistical Association*, **61** (1), pp. 34–40.

Vance K., Howe W., Dellavalle R. P. (2009). Social Internet Sites as a Source of Public Health Information. *Dermatologic Clinics*, **27**(2), pp. 133-136.

Measuring health inequalities: Some application in Marche region

Alberto Franci[a,b], Pietro Renzi[b]
[a] Department of Economics, Society, Politics, University of Urbino, Urbino, Italy.
[b] Department of Economics, Science and Law, University of the Republic of San Marino

1. Introduction

The term "health inequality" generically refers to differences in the health of individuals or groups. Any measurable aspect of health that varies between individuals or social-relevant groupings can be called a health inequality. Absent from the definition of health inequality is any moral judgement on whether observed differences are fair or just. By contrast, a health inequity or disparity is a specific type of health inequality that denotes an unjust difference in health. It is arguable that when health differences are preventable and/or unnecessary, allowing them to persist is unjust.

The monitoring of inequalities in health is an important public health task. Interest in health inequalities amongst EU countries and their regions, as well as amongst the various social clusters in the EU population, is growing.

As a consequence, the search for the best appropriate "summary measure" of health inequalities, that can be observed individually or in terms of groups of individuals, is a task that occupies a lot the researchers involved in related fields.

Lately, in the EU, it has been recognized that a more focused effort is required. Of course, it is natural to suggest and construct methodologies, or indices, that are suitable for assessing trends in mortality, morbidity and self-perceived health. However, the selection of an appropriate indicator or measurement methodology to evaluate and monitor health inequality across the EU-27 countries is a demanding task. This is because each available indicator has advantages and disadvantages. Simple indicators are usually comprehensive but may not have some specific desirable characteristics. Other indicators are more technical and difficult to understand, apply and/or interpret; but can be of more assistance in explaining significant components of the concept of "health inequality". Complex indicators can also be very useful for breaking down the factors and issues relating to inequality. Based on the above, it is reasonable to state that the main goals of our study are the following:

- to present the principal models depicting the wider determinants of health;
- to propose appropriate measurement methods in the form of indicators that can estimate and capture the level of inequality in a population.
- to present some important results of the measures proposed to assess health inequalities in the Marche region, and in other contexts, using existing and available data.

2. Principal models able to explain the social determinants of health

The primary social determinants of health are the circumstances in which people are born, grow up, live, work and age; and the systems that are in place to deal with and prevent illness.

The concept of social determinants of health may be useful to explain how social inequalities can be transformed into health inequalities. Age, sex and inheritance are clearly important factors, but researchers have highlighted that other such determinants include: socioeconomic factors (i.e. education, job status, family/social support, income, community safety…); physical environment; health behaviours (i.e. tobacco use, diet and exercise, alcohol use, sexual activity…); access to care; and quality of care.

The most important models that present the determinants of health are the following:

- Whitehead and Dahlgren model (1991);

ISBN: 978-88-5495-135-8

- Pathway model (or of the causes of the causes) by WHO Commission of social determinants of health (2008);
- Duran and Pérez-Stable titled of "Relationship between health determinants and health disparity outcomes" (2019).

3. Presentation of the most suitable summary measures for monitoring health inequalities

The distribution of a health variable can be described in terms of various statistical measures: its central tendency, dispersion, range, etc. This are univariate measures. Frequently, the term "health inequality" is use incorrectly from a statistical point of view, as the objective is to quantify the relationship between a variable (e.g. gender, race, a socioeconomic characteristic etc) and health, to determine the projected impact of the distribution of this variable on the health of the population. This, therefore involves bivariate measures.

To better identify the main measures, we can take into consideration the following schemas:

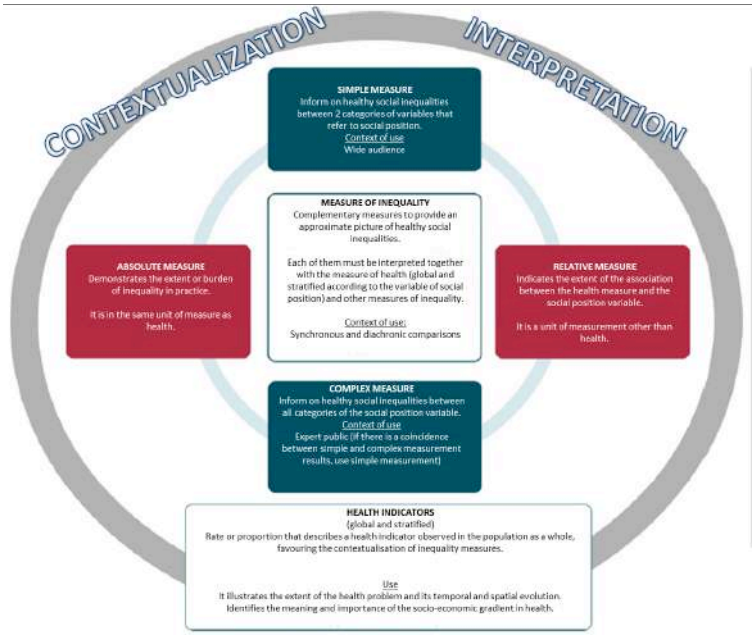


Figure 1: Institut national de santé publique di Québec (2017) - How to use SSISSQ to study social inequalities in health, Version 1, Bureau d'information et d'études en santé des populations

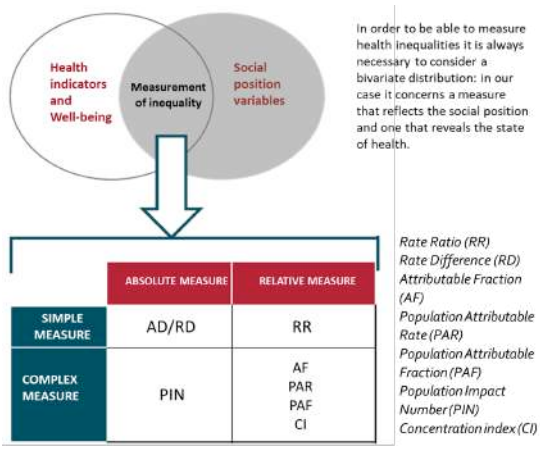


Figure 2: Institut national de santé publique di Québec (2017) - How to use SSISSQ to study social inequalities in health, Version 1, Bureau d'information et d'études en santé des populations

ISBN: 978-88-5495-135-8

4. Methodology

Data to populate the model were derived from the following sources: Health for all (WHO, 2018); the Italian National Institute of Statistics (ISTAT) (2019), which is the main producer of official statistics in Italy, provided national mortality, life-expectancy and cause of death statistics; Passi (Progress of Health Companies in Italy; 2018), Passi d'Argento Population Surveillance System (2018); sentinel events using physician databases; hospital discharge forms; CENSIS (2018) databases; and databases of the Local Health Authority investigated. The data used was for primarily 2017.

As it can be seen in Table 1 and Figure 3 the measures used in our research are absolute difference and rate ratio for some of the data collected, because they are more useful for wide audience, while concentration index was used to compare health status between the higher and the lower socioeconomic categories because the context of use is for a expert public.

5. Results

	Regional value	Value in lowest social group (low educational level)	Value in highest social group (high educational level)	Absolute/Rate difference
General health status				
Life expectancy in men (years)	81,1	80,3	83,4	3,1
Life expectancy in women (years)	87,1	86,6	89,8	3,3
Healthy life years in men	40,5			
Healthy life years in women	38,27			
% of the population that assess their health as good or very good	70%	65,9%	74,7%	8,8%
Accessibility of care				
Breast cancer screening (% women aged 50–69)	71,8%	65,90%	77,80%	11,90%
Autonomous breast cancer screening	29%			
Organized breast cancer screening	51,60%			
Cervix cancer screening (% women aged 25–64)	79,7%	76%	83,4%	7,4%
Autonomous cervix cancer screening	27%			
Organized cervix cancer screening	56,20%			
Delayed contacts with health services because of financial reasons	15,3% - 7,6%*			

* divided into people with limitations and without limitations

Table 1a: Summary of socioeconomic inequalities for selected indicators in the Marche region

	Regional value	Value in lowest social group (low educational level)	Value in highest social group (high educational level)	Rate difference
Appropriateness				
% of adult diabetes patients (aged 25+)	3,8%	5,4%	2,3%	3,1%
Health promotion				
% of the population that reports to smoke daily	19%	17,4%	20,5%	3,1%
% of the adult population considered as being obese (BMI ≥30)	9,5%	15,2%	10,1%	5,1%
% of the adult population considered as being overweight or obese (BMI≥25)	31,5%	35,8%	25,8%	10%
% of the population reporting to eat at least 200 g vegetables and 2 fruits per day	77,3%			
% of the population reporting to practice at least 30 min of physical activity per day	9,1%	7,5%	16,6%	9,1%

Table 1b: Summary of socioeconomic inequalities for selected indicators in the Marche region

Through the use of simple measures, social disparities are observed in general health status, accessibility and appropriateness of care, and health promotion (See Table 1).

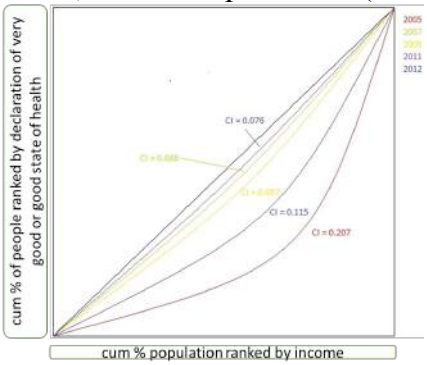


Figure 3: Concentration index: application in the Marche region

Figure 3 shows the Concentration Curves for the Marche region relating to the 5 years where there was a statistical significance in the results of 95% or higher. It suggests that the Concentration Index calculated for the region is relatively stable. It is therefore inferred that perceived good and very good health in Marche tends to be concentrated among those who have a higher socioeconomic status. This serves to emphasise that tackling health inequalities remains a very important issue.

6. Conclusions

The issue of social inequalities and health is a key public health issue. Its inclusion is essential for health policies to be effective at different levels (national, regional, local), as well as for the implementation of health improvement programmes and action for health education among the population. The approach and use of models and principal indices, described in this paper, have been successfully applied to estimate (the significant) health inequalities in the Marche region, particularly in relation to rate ratios and the concentration index.

References

CENSIS (Study Center for Social Investment) (2018). www.censis.it. Accessed 31 July 2019.

Duran, D.G., Pérez-Stable, E.J. (2019). Novel Approaches to Advance Minority Health and Health Disparities Research, *American Journal of Public Health* 109, S8_S10, https://doi.org/10.2105/AJPH.2018.304931.

ISTAT (2019). Information and services for users. https://www.istat.it/en/information-and-services. Accessed 31 July 2019.

O'Donnell, O., van Doorslaer, E., Wagstaff, A., Lindelow, M. (2008). Analyzing Health Equity Using Household Survey Data: A Guide to Techniques and Their Implementation. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/6896 Accessed 12 March 2019

Passi (Progress of Health Companies in Italy) (2018). La sorveglianza Passi. https://www.epicentro.iss.it/passi Accessed 31 July 2019.

Passi d'Argento Population Surveillance System (2018) La sorveglianza Passi d'Argento. https://www.epicentro.iss.it/passi-argento Accessed 31 July 2019.

Popay, J., Escorel, S., Hernández, M., et al. (2008). Understanding and tack-ling social exclusion. Final report to the WHO Commission on Social Determinants of Health from the Social Exclusion Knowledge Network. Lancaster (UK): WHO; 2008, p. 207.

Public Health Agency of Canada (2017). Pan-Canadian health inequalities reporting initiative - summary measures. http://publications.gc.ca/collections/collection_2018/aspc-phac/HP35-79-2017-eng.pdf. Accessed 12 March 2019.

Spinakis, A., Anastasiou, G., Panousis, V., Spiliopoulos, K., Palaiologou, S., Yfantopoulos, J. (2011). Expert review and proposals for measurement of health inequalities in the European Union - Full report, European Commission Directorate General for Health and Consumers. Luxembourg. https://ec.europa.eu/health/sites/health/files/social_determinants/docs/full_quantos_en.pdf. Accessed 12 March 2019.

Whitehead, M., Dahlgren, G. (1991). What can we do about inequalities in health. *The Lancet*, 338, pp. 1059-1063.

World Health Organization (2018). European Health Information Gateway: European health for all database (HFA-DB). https://gateway.euro.who.int/en/datasets/european-health-for-alldatabase/. Accessed 31 July 2019.

ISBN: 978-88-5495-135-8

Socioeconomic inequalities and cancer risk: the challenges and opportunities of worldwide epidemiological data consortia

Carlotta Galeone[a], Rossella Bonzi[a], Federica Turati[a], Claudio Pelucchi[a], Carlo La Vecchia[a]

[a] Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy

1. Introduction

Consolidated evidence shows that morbidity and premature mortality from noncommunicable - or chronic - diseases (NDCs, i.e., cardiovascular diseases, diabetes, chronic respiratory diseases and cancers) is higher in low-income and middle-income countries, and, in high-income countries, in people with lower social economic position (SEP). Among NCDs, cancer is the second leading cause of death worldwide with incidence, survival and mortality rates subjected to large variations across countries and, within countries, across social groups. High-income countries show much higher incidence rates of several cancers than most low- and middle-income countries. However, within almost all countries, mortality rates are, to a disproportionate extent, higher for groups of the population with low SEP due to poorly designed health systems or limited or even inhibited access to preventive interventions, early detection, diagnosis, treatment or and palliative care.

Over the last few decades, social epidemiology started to develop as a solid epidemiology branch, with the aim to understand how social experiences influence population health, with a main feature being the study of SEP in relation to NDCs.

At the same time, the advent and growing of collaborative and interdisciplinary research frameworks along with the proliferation of multi-institutional research consortia markedly affected cancer epidemiology. The uniquely large data set on which consortia are based permit to define and quantify, with a degree of accuracy higher than ever before, the main effects of each risk factor of interest and to adequately address associations in subgroups of the population, as well as interaction between environmental, genetic and socio-economic factors. A general feature and strength of consortia is the easy and prompt communication among members for an interconnected sharing of knowledge and study results, aiming at maximizing the efficiency to understand, prevent, treat, and relieve the risk and the incidence of diseases on the population at a global level.

2. Socioeconomic position and cancer risk evaluation in worldwide epidemiological data consortia

SEP is a complex concept, which involves several dimensions including education, work experience and household income, access to material resources, prestige and social position. These dimensions are associated, even though each of them accounts for different aspects of the socioeconomic stratification. In a broader sense, speaking of socioeconomic status involves referring to the most common forms of inequality.

The assessment of socioeconomic position in the epidemiologic research is usually performed throughout a series of indicators, traditionally education, occupation and income, being their specific use often and strictly dependent on data availability. Individual, or, better, household income, which may be a useful indicator in particular for women or those who may not be the main earners in the household, reflects the material component of people everyday life. However, income is the SEP indicator mostly subjected to changes, also on a short-term basis, it is age-dependent and with the highest non-response rate in epidemiological investigations when compared to other SES measures. Occupation reflects the privileges related to social standing, material resources and job-related risk factors. Occupation based indicators of SEP are widely used in the epidemiologic research due to their large availability in many data sources and the ability to adapt measures from one to several individuals belonging to the same family or unit. Occupation indicators clearly cannot be assigned to currently unemployed or retired

ISBN: 978-88-5495-135-8

people, housekeepers, students, and people with informal, unpaid or illegal jobs. Also, the definition of occupation related to SEP may vary according on individual birth date and geographical location, which consequently represents an issue in terms of international comparisons. Education reflects the intellectual assets of individuals besides the socioeconomic conditions in childhood and adolescence and it represents people potential opportunity to access to higher-level jobs and earnings. Educational attainment is a widely used indicator of SEP. The strength of using education as a proxy for SEP in the adult population is the evident reduction of the likelihood of reverse causation (e.g., whether poor health may be cause or consequence of low SEP), which always represents a big issue of other standard SEP measures.

Large data consortia as The Stomach Cancer Pooling (StoP) Project and the International Head and Neck Cancer Epidemiology (INHANCE), in which the University of Milan is proactively involved, allowed investigators to address the effects of education and household income, the main SEP determinants, on the onset and evolution of gastric and head and neck cancer, respectively, confirming the existence of a strong association between low SEP and the disease.

The statistical analyses are carried out through pooled analyses of individual-level data, after central collection and validation of the original datasets. The individual-level data approach allows harmonization of information and analyses, consistency of adjustment terms and multivariate models, and investigation of heterogeneity and interaction between covariates. The large amount of data allows to investigate cancer subsites and histological subtypes, and to evaluate subgroups of the populations and possible interactions among risk factors. Brief descriptions of the StoP and INHANCE consortia and their results on SEP are here reported.

The StoP Project is a consortium of epidemiological studies on gastric cancer established in 2012; the University of Milan is among the founders of the project. Up to date, the consortium includes 33 studies for a total of 13,000 gastric cancer cases and 31,000 controls. The main aim of the StoP Project is to examine several lifestyle, environmental and genetic risk factors in association with gastric cancer, taking advantage of a large data set with original information from various geographic areas.

StoP recently published results on education, household income and gastric cancer. Data on education level were available from 25 studies; seven studies provided data on household income. Education data were standardized across studies following the International Standard Classification of Education from the United Nations Educational, Scientific and Cultural Organization (UNESCO), ISCED 2011. Education level was classified as: (i) low (no education, early childhood and primary education (ISCED 0–1)); (ii) intermediate (secondary education and postsecondary non-tertiary education (ISCED 2–4)); (iii) high (tertiary vocational and higher education, and education leading to a university degree (ISCED 5–6)). Household income was estimated by standardizing available study questionnaires data; comparable income levels were grouped into 4 categories, i.e. low, lower middle, upper middle and high

Analyses of the StoP data showed that SEP, measured through education level and household income, is a strong determinant of gastric cancer. Subjects with intermediate and high education levels had, respectively, about 30% (pooled odds ratio, OR, 0.68, 95% confidence interval, CI, 0.55-0.84) and 40% (pooled OR 0.60, 95% CI, 0.44–0.84) decreased risks of GC compared to those with lower education attainment. Analyses accounted for a number of lifestyle and dietary habits, which may confound the associations of SEP with gastric cancer, including tobacco smoking, race/ethnicity and the intake of alcohol, fruit and vegetables. Strong inverse associations were observed for both cardia e noncardia gastric cancer, as well as for diffuse and intestinal subtypes. In addition, the inverse association between education level and gastric cancer risk was evident regardless of infection with *Helicobacter Pylori* (*HP*), and in subgroups defined by age, sex, cigarette smoking and alcohol drinking. In analyses by geographic area, strong inverse associations were reported by studies from Europe and Asia, while combined results from the three North American studies indicated a non-significant inverse association. Conversely, Central/South America studies (mainly Mexican studies) did not find any relation between education level and gastric cancer, raising concerns about the reliability of education as a SEP proxy in such countries. When household income was used as proxy of

ISBN: 978-88-5495-135-8

SEP, a 35% reduced risk of gastric cancer was observed for subjects in the highest versus the lowest income category (OR 0.65, 95% CI, 0.48–0.89).

The INHANCE consortium, established in 2004 as a collaboration among international research groups, includes over 35 international studies including 30,000 patients with head and neck cancer and 40,000 controls without these cancers. The primary goal of the consortium is to address the associations of head and neck cancer with a number of environmental factors, including tobacco smoking and alcohol drinking (i.e., the most relevant risk factors for the disease), anthropometric characteristics and nutritional factors).

The INHANCE consortium conducted a detailed analysis on the association between low educational status and household income and head and neck cancer. ISCED 97 protocol was used to categorize education levels. Education level was classified as follows: (i) low (ISCED 0–1); (ii) intermediate (ISCED 2–4); (iii) high (ISCED 5–6). Household income data, available from 10 studies (mostly from USA), were standardized as far as possible by grouping comparable levels based on the strata used in the original study, starting from category 1 (lowest income level) to category 5 (highest level).

The analyses on education, based on 31 case-control studies and almost 24,000 head and neck cancer patients and 32,000 controls, indicated that subjects with low education had a more than two-fold increased risk of head and neck cancer compared to those with high education. When accounting for smoking, alcohol and selected dietary factors, the association was attenuated but still significant, with an over 30% elevated risk among subjects with low versus those with high education (pooled OR 1.34, 95% CI, 1.04–1.73). In addition, the risk remained increased by over 50% in subjects who never smoked or used other type of tobacco and never drank alcohol (OR 1.61, 95% CI, 1.13–2.31). This suggests that the association of head and neck cancer with low education level is not totally attributable to these detrimental behaviors, although some degree of residual confounding could not be excluded. In addition, part of the association observed with low education could be explained by *Human Papilloma Virus* (HPV) infection. The association with low education level was observed for all head and neck cancer subsites and was somehow stronger in North and Central/South American populations as well as in higher income inequality countries. The analyses on household income gave results in line with those on education, with an over two-fold increased risk for the lowest *vs* the highest category of income. Again, the association was attenuated, but still evident, after allowance for smoking and alcohol, with an over 50% increased risk among subjects with the lower monthly income.

The above-mentioned results are a clear and complete example of the crucial role of social epidemiology in understanding socio-structural factors related to health and disease. In an era of fast inter-diffuse communication and data-sharing, large collaborative groups and data consortia are among the most effective strategies to create new social epidemiological useful evidences. In particular, data analyses of large epidemiological consortia found that SEP is strongly related to a number of cancers. Reduction of socioeconomic inequalities both at national and international level are advocated to decrease the burden of cancers in deprived populations.

References

Conway, D. I., Brenner, D. R., McMahon, A. D., Macpherson, L. M. D., Agudo, A., et al. (2015). Estimating and Explaining the Effect of Education and Income on Head and Neck Cancer Risk: Inhance Consortium Pooled Analysis of 31 Case-Control Studies from 27 Countries. *International Journal of Cancer* **136**(5), pp. 1125-39.

Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J. W., Davey Smith, G. (2006). Indicators of Socioeconomic Position (Part 1). *Journal of Epidemiol Community Health* **60**(1), pp. 7-12.

Geyer, S. Hemstrom, O. Peter, R. Vagero, D. (2006). Education, Income, and Occupational Class Cannot Be Used Interchangeably in Social Epidemiology. Empirical Evidence against a Common Practice. *Journal of Epidemiol Community Health* **60**(9), pp. 804-10.

Honjo, K. (2004). Social Epidemiology: Definition, History, and Research Examples. *Environ Health*

Prev Med **9**(5), pp. 193-9.

Ioannidis, J. P., Schully, S. D., Lam, T. K., Khoury, M. J. (2013). Knowledge Integration in Cancer: Current Landscape and Future Prospects. *Cancer Epidemiol Biomarkers Prev* **22**(1), pp. 3-10.

Niessen, L. W., Mohan, D., Akuoku, J. K., Mirelman, A. J., Ahmed, S., Koehlmoos, T. P., Trujillo, A., Khan, J., Peters, D. H. (2018). Tackling Socioeconomic Inequalities and Non-Communicable Diseases in Low-Income and Middle-Income Countries under the Sustainable Development Agenda. *Lancet* **391**(10134), pp. 2036-46.

Pelucchi, C., Lunet, N., Boccia, S., Zhang, Z. F., Praud, D., et al. (2015). The Stomach Cancer Pooling (Stop) Project: Study Design and Presentation. *European Journal of Cancer Prevention* **24**(1), pp. 16-23.

Rota, M., Alicandro, G., Pelucchi, C., Bonzi, R., Bertuccio, P., et al. (2019). Education and Gastric Cancer Risk-an Individual Participant Data Meta-Analysis in the Stop Project Consortium. *Internatonal Journal of Cancer*.

Shavers, V. L. (2007). Measurement of Socioeconomic Status in Health Disparities Research. *Journal of Natl Med Assoc* **99**(9), pp. 1013-23.

UNESCO. (1997). International Standard Classification of Education: Isced 1997. Paris: UNESCO Institute for Statistics.

UNESCO. (2012). International Standard Classification of Education: Isced 2011. Montreal: UNESCO Institute for Statistics.

Vaccarella, S., Lortet-Tieulent, J., Saracci, R., Conway, D.I., Straif, K., Wild, C.P. (2019). Reducing Social Inequalities in Cancer Evidence and Priorities for Research. *IARC Scientific Publication No. 168*.

A data analytics framework: medical prescription pattern dynamics

Ilaria Giordani[a], Gaia Arosio[b], Ilaria Battiston[a], Francesco Archetti[a,b]
[a] Department of Computer Science, Systems and Communication, University of Milano Bicocca, Milano, Italy;
[b] Consorzio Milano Ricerche, Milano, Italy;

1. Introduction and motivation

Healthcare data is defined as that information used to provide, manage and/or report the services used across the entire healthcare system. Its origin is the encounter between a patient and a provider, who will record the service rendered, the conditions of the service, patient information and clinical information.

Prescription pattern monitoring studies (PPMS) exploit medical information to improve the prescribing practices and thus the standards of medical treatments at all levels of healthcare. Results and insights are obtained through healthcare analytics, a powerful tool which helps to uncover hidden relationships in the data as well represent and visualize them.

Data analytics can give a major contribution to PPMS, tackling health and socioeconomic challenges: by means of insights originated from analytical results health authorities can plan informed actions, promote appropriate use and reduce the abuse/misuse of monitored drugs.

Big data can assess the appropriateness of prescribing through the existing classification systems, comparing drug consumption patterns within specified time ranges, focusing on specific products and defined geographical areas.

This paper presents a data analytics framework composed by different statistical computational modules, summarising relevant insights obtained by the analysis of data related to about 1.500 general practitioners (GPs) and around 1.015.000 patients, during the period 2000 to 2018. The available database was collected and managed complying with current privacy regulations, within the project "Territorial Analysis on Local Antibiotic Resistance through Data Analytics Techniques", by Consorzio Milano Ricerche, Milano (Italy).

This research is specifically focussed on the doctor-patient relationship: person-centred care concerning diagnoses and prescriptions, analysing their changes according to habits, physiological aspects, time and geographical area of both interested parts. After collecting the first batch of results and checking them with domain experts, information with unusual patterns is outlined, and further examination is made on a subset of features. The two main risks encountered while doing analysis are information loss and inappropriate prescribing, that compromise the quality of statistics. Data may be incomplete, biased or filled with noise: another goal of analytics is to contrast incompleteness and incorrectness, obtaining coherent and clear results. The first step to take, having a deep understanding of the data, is recognising the extent and impact of the progressive information loss, to define the final dataset according to the specific target of the analysis. After a quality assessment of the whole database, the sample was reduced to about 720.000 patients, balanced in terms of gender, age, diagnosis and prescriptions. Starting from the explorative and descriptive analytics modules, the data analysis process allows the reconstruction of patterns of diagnosis and prescriptions. Prescriptions are additionally grouped and measured along with diagnoses, to identify eventual linkage between most popular ones, and focus on discrepancies. Global analytics relies on the "patient journey" a component of the framework that can reconstruct the patient's history and relationships with primary healthcare, identifying patterns and changes. This approach can be useful to extract a cohort of patients beginning their treatment, and analyse the variations of first-time prescriptions, especially for chronic illnesses. The biggest risk is again the loss of information: the impact of data cleansing is heavy, and the obtained results might not give an insightful

ISBN: 978-88-5495-135-8

perspective. Descriptive statistics is again applied to the resulting dataset, extracting the most common co-prescriptions which show the heavy impact of antibiotics. This suggests the issue of antibiotic resistance, one of the major challenges to public health despite guidelines and calls by government agencies which threatens the effective prevention and treatment of an ever-increasing range of infections caused by bacteria, parasites, viruses and fungi. Microorganisms exposed to antimicrobial drugs develop the ability to defeat substances designed to kill them, making infections persist in the body due to the unsuccessful action of agents.

The analytics framework is composed of two main modules on one side exploratory analysis, including in particular clustering performed by k-means, also to capture the process dynamics thru time series analysis and clustering of trajectories on the other side network analytics provides a very powerful model to analyse connected data because its algorithms focus on the relationships between nodes to infer the organization and dynamics of complex systems. Also, simple network characterization like degree and centrality indexes allows to identify communities of similar nodes and their dynamics. Community formation is common in all types of networks and its characterization can uncover structures like hubs and hierarchies, find nested relationships and infer similar behaviours.

Defined time frames in limited temporal windows are selected to make a detailed trajectory analysis related to prescriptive appropriateness compared to antibiotic resistance, without going into details of pathologies. Having information about AICs (authorisation to commerce codes) allows to give a new insight not only on ethical matters, but also on trends of expenditures. Prescription patterns are analysed highlighting the heterogeneous trends, using data related to the last 10 years (2000-2017) to avoid dispersion of information and reconstruct coherent time series. The set of records regarding antibiotics in the selected time range consists in approx. 8 millions of prescriptions, belonging to almost 700.000 patients. This allows to identify external causes for unusual trends, such as marketing campaigns, aggressive advertisement and pressure to use a specific drug rather than an equivalent one. Trends are similar to ATC ones: the two sets of values have in fact a Pearson coefficient of 0.88. Prescription patterns are analysed to group doctors according to their prescription habits, using a subset of 372 general practitioners, performing clustering to identify main features of communities and observing their variation in terms of number of members through time. Comparisons are made selecting two snapshots of data according to different years, and visualising outcomes to understand whether individuals have shifted cluster(general practitioners have changed habits). After identifying the clusters and linking each one with a specific prescription patten, it is possible to discriminate between GPs with changing and constant behaviours among time. All the amounts of prescriptions for each antibiotic have to be considered to extract similarity and scaled to normalise numbers. Other attributes such as patients' phenotype and total amount of prescriptions are illustrative, and will be attached to clustering results, to offer further information to be compared after having a general idea of each doctor's group. The cluster analysis for each antibiotic displays how general practitioners can be assigned to 4 clusters suggesting clearly possible factors of influence on antibiotic resistance. Additional analysis have been performed on variations through time, understanding the relationship between number of patients and number of prescriptions to check whether general practitioners are effective over-prescribers.

2. Graph databases: an enabling technology of network analytics

A graph G =<E; V> is an abstract data type showing connections (edges E) between pairs of vertices (V). Nodes identify entities and their properties, while relationships are joining attributes between tables (patients, drugs, GP) which can be further specialized with additional characteristics.

Graph databases are data management systems allowing persistent representation of entity and relationship in a graph structure, implementing the Graph Model efficiently down to the storage level. Neo4j a graph platform specifically optimized to map, analyse, store and traverse networks of connected data to reveal invisible contexts and hidden relationships. In Neo4j, everything is

ISBN: 978-88-5495-135-8

stored in the form of an edge and node: any node and edge can have any number of attributes and be labelled. Unlike other databases, relationships take first priority. A graph database is purpose-built to handle highly connected data, providing great performance, flexibility and frictionless development.

Queries allow to match pattern of nodes (and their attributes) and relationships (and their attributes) in a graph. Queries are written using Cypher, a declarative graph query language that allows for expressive and efficient querying and updating of the graph.

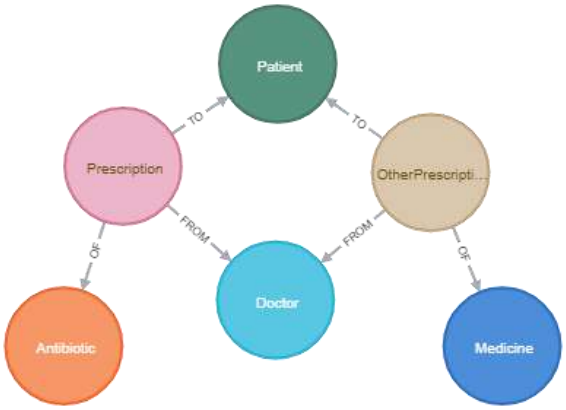


Figure 1: The graph database obtained with Neo4j library APOC

Another Neo4J library is "Graph Analytics". It returns implemented and parallel version of solvers of common network problems e.g. Centrality/Importance indexes (which determine the importance of distinct nodes in the network), Pathfinding & Search (which finds the optimal paths or evaluates route availability and quality), Community Detection (which detects group clustering or partition options) and Heuristic Link Prediction (which estimates the likelihood of nodes forming a relationship).

3.Visualization of results

To give an instance of the representation power of Neo4J, we show (Figure 2) how the ease of inferring and representing links in a graph is perfectly suited for coupling prescription.

Figure 2 depicts the co-prescription graph: orange nodes are antibiotics, blue other medicines, the number of co-prescriptions is represented thru the weight of the link. The two antibiotics at the center are the two most prescribed. Along with the relation one can visualize the features of each node and link. The isolated component represents an antibiotic (antimalarial) given to treat arthritis coupled with a corticosteroid for rheumatism (co-prescribed about 5000 times).

Figure 3 displays the graph of inter-GPs relationships based on a "prescriptive similarity" parameter. Analysing the graph with the Graph Analytics Neo4J library exhibits the communities of GPs with similar prescriptive habits.

ISBN: 978-88-5495-135-8

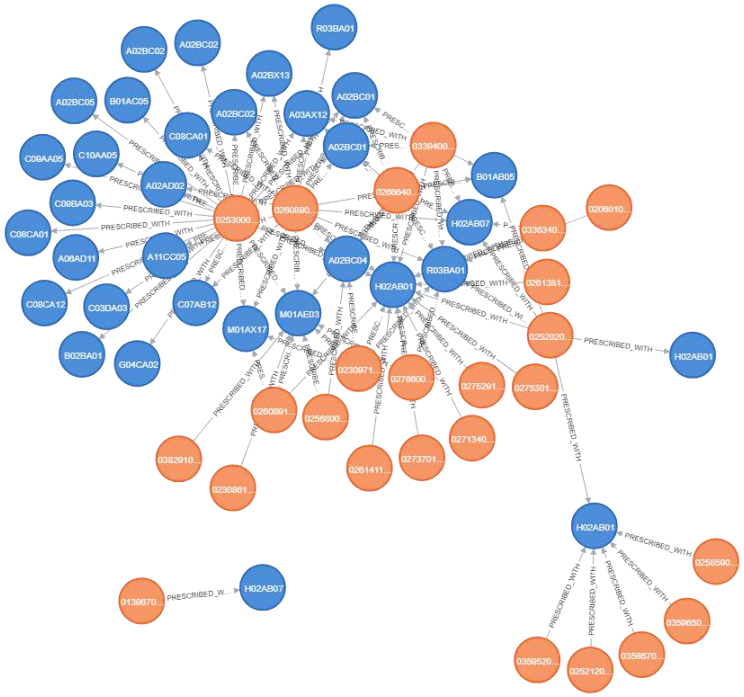


Figure 2: Co-prescription graph

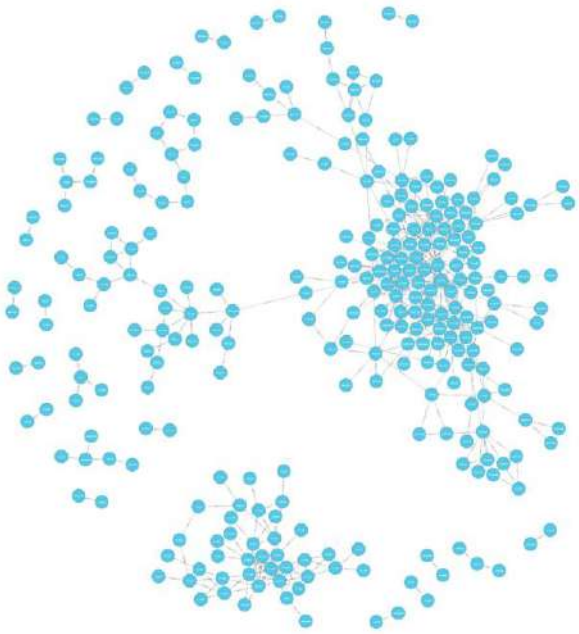


Figure 3: Relationships graph between GPs

References

Candelieri A., Giordani I., Archetti F. (2017). Automatic Configuration of Kernel-Based Clustering: An Optimization Approach, in *Learning and Intelligent Optimization. Prooceeding of LION 2017 conference* eds. R. Battiti, D. Kvasov, Y. Sergeyev Lecture Notes in Computer Science, Springer, Cham, pp. 34-49.

Needham, M., Hodler, A. E. (2019*). Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O'Reilly Media.

Newman, M. (2018). *Networks*. Oxford University press.

Applying network modelling to uncover the relationships among well-being dimensions

Laura Giuntoli [a], Giulio Vidotto [a]
[a] Department of General Psychology, University of Padova, Padova, Italy

1. Introduction

Individual well-being can be interpreted as the amalgam of the different components proposed in the literature by the hedonic stream, measuring positive emotional states, and the eudaimonic stream, that proposed several dimensions of optimal psychological functioning such as self-acceptance, meaning in life and positive relationships (Deci & Ryan, 2008). According to the analogy with mental illness proposed by Keyes (2002), mental health may be operationalized as an emerged condition based on the concept of a syndrome of symptoms of individuals' subjective well-being, that is individuals' perception and evaluation of their life in terms of affective states and psychological and social functioning. Thus, well-being can be conceived as a network of interacting and self-reinforcing symptoms, not an underlying entity that produces symptoms (i.e., positive individual traits and experiences). For instance, sense of competence may activate other positive psychological features (e.g., self-acceptance and engagement), likely in circular, self-reinforcing ways. In this view, hedonic and eudaimonic dimensions of well-being are not mere passive psychometric indicators, but are active causal ingredients of mental health.

Probably, the latent variable approach, that is based on items' local independence, is not the best instrument to test such a theoretical model of well-being. By contrast, network psychometrics (Epskamp, Rhemtulla, & Borsboom, 2017) is a data-driven approach that does not bring into play latent factors and allows the model structure to spontaneously emerge from the relationships among indicators. Network modelling is a powerful tool to simultaneously illustrate the strength of the interconnections among individual indicators and the presence of clusters identifying specific well-being domains.

2. Strength decomposition

A network is composed of a set of nodes, that represent any kind of entity, and a set of edges, that represent any kind of relationship which connect the nodes. A psychological construct can be represented by a psychometric network in which the nodes are questionnaire items, while the edges connecting them are partial correlations (i.e., the covariation displayed between two nodes when the effects of all the other nodes in the network are partialled out).

Edge weights estimation is typically achieved by means of a regularization algorithm (e.g., GLASSO; Epskamp & Fried, 2018) to estimate a sparse network in which the spurious partial correlations are shrunk to zero.

Edges connecting nodes can differ in their strength, indicating if a relationship is strong (visualized as thicker edges) or weak (visualized as less saturated edges), and positive (green edges) or negative (red edges). The properties of a network structure can be summarized by the centrality metrics: Strength is the sum of a node's edge weights; Closeness is the inverse of the sum of the shortest paths between a specific node and all the other nodes; Betweenness is the total number of shortest paths that pass through a node. Centrality metrics indicate how

closely a node is interconnected with all the other nodes in the network, meaning that it could be considered a particularly important indicator of the construct.

Our focus in this report lies on the decomposition of node strength centrality to numerically summarize the structural relationships observed in a network. Given that the strength computation in certain cases uses weigths' absolute values, the term expected influence (EI) is used to make clear that the compuation takes into account the negative or positive signs of the edge weight's values. The global EI of a node V_i is defined as:

$$EI_i = \sum_{j=1}^{N} w_{ij}, \qquad \forall \quad V_j \; where \; j \neq i$$

where N is the total number of nodes in a network, and w_{ij} is the element in the weights matrix that corresponds to the row i, namely the node of interest V_i, and to the column(s) j, representing all the other nodes (V_j) in the network except V_i.

Bridge-EI accounts for the influence of a node on the rest of the network except the nodes' community it belongs to (C_k). Bridge-EI formula can be defined as:

$$bridge\text{-}EI_i^k = \sum_{j=1}^{N} w_{ij}, \qquad \forall \quad V_j \notin C_k \; where \; j \neq i.$$

Dealing with a multi-dimensional network it is also important to consider the influence of a node with respect to specific communities, thus we present other two new measures of nodes' importance that decompose EI and bridge-EI.

EI represents the total effect of a node on all the other nodes in the network, thus it can be decomposed in bridge-EI (i.e., the effect of a node on the network except the nodes belonging to its own community), and in the strength of a node within the nodes belonging to the same community, that we call *within-EI*. The within-EI of a node V_i is defined as:

$$within\text{-}EI_i^k = \sum_{j=1}^{N} w_{ij}, \qquad \forall \quad V_j \in C_k \; where \; j \neq i.$$

When in a network can be distinguished more than two communities the importance of a node can be decomposed separately for each nodes' communities. Instead of considering globally bridge-EI, the strength of the connections of a node V_i with a specific community other than its own can be named *between-EI* and we define it as:

$$between\text{-}EI_i^{k_1 k_2} = \sum_{j=1}^{N} w_{ij}, \qquad \forall \quad V_j \in C_k \; where \; V_i \in C_{k_1} \; \& \; C_{k_1} \neq C_{k_2}$$

where the superscript k_1 indicates the belonging community of the node V_i, whereas the superscript k_2 indicates a specific community of nodes other than the community of node V_i. Thus, bridge-EI corresponds to the sum of all the between-EI indices, while EI corresponds to the sum of within-EI plus bridge-EI.

3. Application

The data were selected from an existing database of 2392 respondents (Giuntoli & Vidotto, 2019) who compiled the Diener's questionnaires (Diener et al., 1985; 2010), namely the Flourishing Scale (FS), the Satisfaction With Life Scale (SWLS) and the Scale of Positive and Negative Experience (SPANE). For illustrative purpose we selected three items for each of four

dimensions of well-being: *flourishing* (F), *life satisfaction* (LS), *positive affect* (PA) and *negative affect* (NA).

The graphical visualization of the network of the selected items (Figure 1) is based on the Fruchterman-Reingold algorithm that places nodes with stronger and more connections close to each other. Strength centrality metrics are shown in Table 1.

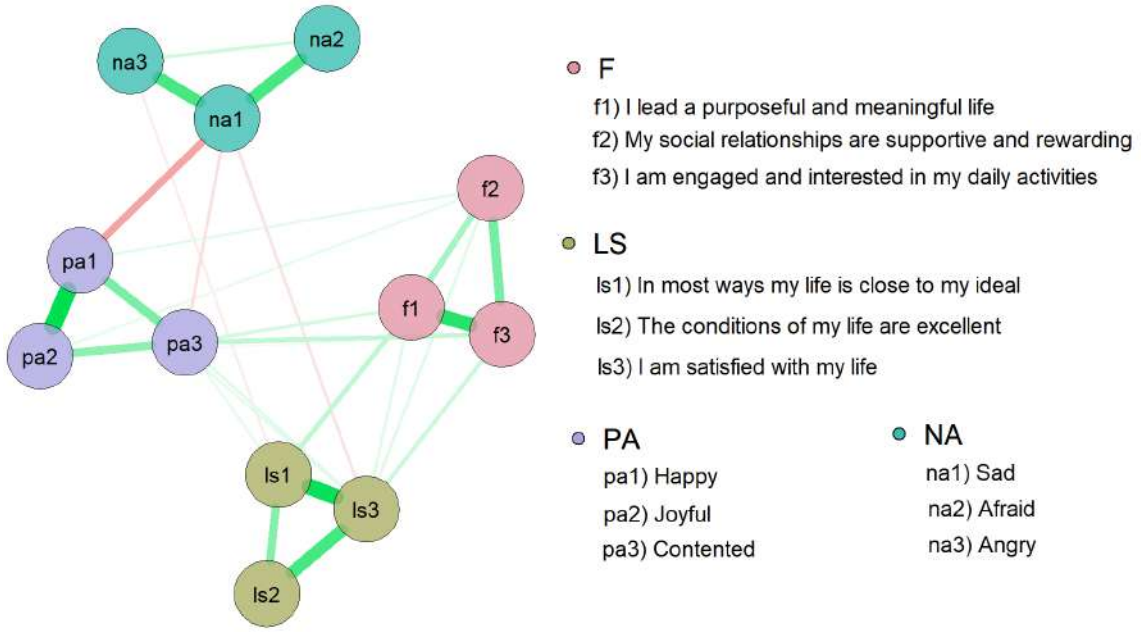


Figure 1. Estimated network model of well-being. The items are represented by nodes colored according to their respective dimension

Node	Within-EI				Between-EI				Bridge-EI	EI
	F	LS	PA	NA	F	LS	PA	NA		
f1	0.604					0.182	0.086	0.000	0.268	0.872
f2	0.445					0.063	0.109	0.000	0.172	0.617
f3	0.705					0.103	0.159	0.000	0.262	0.967
ls1		0.735			0.124		0.051	-0.042	0.133	0.868
ls2		0.617			0.000		0.000	0.000	0.000	0.617
ls3		0.860			0.224		0.116	-0.062	0.278	1.138
pa1			0.775		0.057	0.057		-0.199	-0.085	0.690
pa2			0.740		0.087	0.000		0.000	0.087	0.827
pa3			0.490		0.209	0.110		-0.071	0.248	0.738
na1				0.730	0.000	-0.062	-0.271		-0.333	0.397
na2				0.457	0.000	0.000	0.000		0.000	0.457
na3				0.459	0.000	-0.042	0.000		-0.042	0.417

Table 1. Nodes metrics describing the network structure by means of strength decomposition.

The well-being items form four tightly connected communities of nodes, with the items clustering together according to their theoretical dimension. Indeed, the thicker edges are those connecting the items belonging to the same dimension. These intradimensional relationships are summarized by the within-EI metric that can be interpreted as an index of internal consistency given that it summarize the connectivity of a node toward the nodes belonging to the same community.

The analysis of expected influence by means of strength decomposition is able to reveal important structural relationships among well-being items. First, the overall EI of a node denotes the amount of its interconnections with the other nodes in the network. Consequently, overall EI can be interpreted as the conceptual centrality of a node in the definition of well-being. Second, by means of within-EI it can be quantified the degree of which a node cluster together with other nodes pertaining to the same conceptual dimension of well-being. Third, between-EI is useful to identify which nodes serve as bridge between separate well-being domains. To provide an example of analysis of a specific node we describe the strength decomposition of node ls3 ("I am satisfied with my life") that showed the highest value of overall EI. Notably, item ls3 represents a superordinate focal node given that its semantic content refers to a global judgement on life overall and possibly encompasses judgements on subordinate life aspects represented by the other well-being dimensions. Through the analysis of strength decomposition it can be noted that most of the ls3 strength is determined by its within-EI (0.860). Between-EI summarizes the ls3 strength on the other dimensions that are (in descending order) 0.224 for F, 0.116 for PA and -0.062 for NA.

4. Discussion

We illustrated a simple procedure to summarize the structural relationships among sample items that depict a well-being construct. The starting point was a partial correlation matrix of selected items. Following the current state of the art in network psycometrics we applyed a regularization algorithm (GLASSO) to obtain a sparse weight matrix. Whereas latent variable approaches rely on the assumption of local indipendence (i.e., the observed variables are conditionally independent of each other given that the latent variable explains why they are related to one another), because of the absence of latent variables network psychometrics is particularly useful when we are interested in considering the cross-dimensional interconnections among items. Through strength decomposition it is possible to summarize the information contained in the weigths matrix by summing up the elements of the matrix according to the items' putative dimensions. The strength decomposition acts as an item analysis and it is a potentially useful method to simultaneously explore the internal consistency of the items assigned to a dimension (within-EI) and to check for cross-dimansional associations of specific items (between-EI).

References

Deci, E. L., & Ryan, R. M. (2008). Hedonia, eudaimonia, and well-being: An introduction. *Journal of happiness studies*, **9**(1), pp. 1–11.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, **49**(1), pp. 71–75.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2010). New Well-being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings. *Social Indicators Research*, **97**(2), pp. 143–156.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, **23**(4), pp. 617–634.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*, **82**(4), pp. 904–927.

Giuntoli, L. & Vidotto, G. (2019). Exploring Diener's multi-dimensional conceptualization of well-being through network psychometrics (Version 1). *SageSubmissions*.

Keyes, C. L. M. (2002). The Mental Health Continuum: From Languishing to Flourishing in Life. *Journal of Health and Social Behavior*, **43**(2), pp. 207.

Emotional Text mining and health psychology: the culture of organ donation in Spain

Francesca Greco[a], Silvia Monaco[b], Michela Di Trani[b], Barbara Cordella[b]
[a] Department of Social Sciences and Economics, Sapienza University of Rome, Rome, Italy.
[b] Department of Dynamic and Clinical Psychology, Sapienza University of Rome, Rome, Italy.

1. Introduction

According to literature, people's behaviours depends not only on their rationale thinking but also, and sometimes most of all, on their emotional and social way of mental functioning (Carli, 1990; Moscovici, 2005). That is, people consciously categorize reality and, at the same time, unconsciously symbolize it emotionally (Fornari, 1976). These two thinking processes are the product of the double-logic way of functioning of the mind (Matte Blanco, 1981) which allows people to adapt to their social environment.

The unconscious processes are social and culturally determined, as people generate interactively and share the same emotional meanings. If the conscious process sets the manifest content of a text, that is what is said, the unconscious process can be inferred through how it is said, i.e. the words chosen to narrate and their association within the text. We consider 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 (Salvatore and Freda, 2009; Grasso et al., 2016; Greco, 2016).

This paper presents the application of Emotional Text Mining (ETM) in the field of health psychology and, in particular, on organ donation. ETM is an unsupervised text mining procedure aiming to profile media discourses that can be considered a proxy of the culture setting people's choice to donate. It allows the identification of the representation, and the cultural symbolization of organ donation. Organ donation is an extremely important issue in our society as patients on the waiting lists for transplantation die every year due to the lack of organs. The Spanish healthcare system is a best practice in Europe, reaching 48 donors per million people. A best practice compared to the Italian one in which only 28 donors for million people make this choice.

In order to understand the cultural elements influencing the choice to donate, this paper applies ETM to the Spanish media and compare the results to an Italian study performed by the authors and presented at the ESA/RN27 Mid-Term Conference, held in Catania in 2018 (Monaco et al., 2018).

2. Methods

The ETM (Cordella et al., 2014; Greco, 2016; Greco and Polli, 2019) is a text mining procedure that, by means of its bottom-up logic, allows for a context-sensitive text mining approach on unstructured data. ETM is a non-supervised text mining methodology, based on a socio-constructivist approach and a psychodynamic model, statistically simulating the inverse process of the mental functioning (Greco and Polli, 2019). According to this approach, sentiment is not only the expression of a mood, but also the evidence of a latent and social thinking process that sets people interactions, behavior, attitudes, expectations, and communication. It allows for the detection of the symbolic matrix and the representations of an entity, e.g. organ donation. These elements are connected between them, as the symbolic matrix generate the representations and the representation sets the social interactions (Moscovici, 2005).

In order to explore the culture on organ donation of the Spanish media, we collected all the articles published in the last ten years, from 2009 to April 2019, containing the multiword "donacion de organos" from the *El Mundo* and *El Pays* online archives, two of the most

ISBN: 978-88-5495-135-8

widespread Spanish newspapers. The sample of 342 articles were collected in a large corpus of 220.615 tokens. First, two lexical indicators were calculated in order to assess whether it was possible to statistically process data: the type-token ratio and the percentage of hapax (TTR= 0.092; Hapax%= 51.5).

Then, data were cleaned and pre-processed with the software T-Lab (Lancia. 2017) and keywords selected. In particular, we used stem as keywords instead of type, filtering out the multiword organ donation and those of the high and low rank of frequency (Greco. 2016). All the texts were segmented in context units (CU), and on the CU per keywords matrix, we performed a cluster analysis with a bisecting k-means algorithm (Savaresi and Boley, 2004) limited to ten partitions, excluding all the CU that did not have at least 2 keywords co-occurrence. Three clustering validation measures were taken into account in order to identify the optimal solution: the Calinski-Harabasz, the Davies-Bouldin and the intraclass correlation coefficient (ICC). To finalize the analysis, a correspondence analysis on the keywords per clusters matrix (Lebart and Salem, 1994) was made in order to explore the relationship between clusters, and to identify the emotional categories setting the symbolic matrix and the representation of organ donation.

3. Main results

The results of the cluster analysis show that the keywords selection criteria allow the classification of 98,2% of the CU and the optimal solution was six clusters. The correspondence analysis detected five latent dimensions, and the first three factors explain the 72,0% of the inertia. The result interpretation in reported in Table 1.

Table 1: Correspondence analysis interpretation (between brackets are reported the cluster coordinates in each factor)[1]

Cl	UC	%UC	Factor 1 Perspective	Factor 2 Intervention	Factor 3 Transplantation	Factor 4 Organ donator	Factor 5 Promoter
1	598	12.85	Community (0.362)	Social (0.392)			Health System (0.748)
2	1161	24.95	Personal (-0.453)			Dead (0.475)	Health System (0.229)
3	540	11.60	Personal (-0.671)	Medical (-0.345)	Waiting list (0.493)	Alive (-0.691)	Media (-0.309)
4	1034	22.22		Social (0.622)	Transplantation (-0.200)	Alive (-0.264)	Media (-0.309)
5	676	14.53	Community (0.957)	Medical (-0.466)	Waiting list (0.309)		Media (-0.210)
6	645	13.86	Personal (-0.193)	Medical (-0.628)	Transplantation (-0.782)		

The first factor identifies the perspective, distinguishing the individual interest and reasoning in making the donation choice from the community one. The second factor reflects the health intervention, characterized by the medical procedure and the social intervention, focused on actions and initiatives about organ donation. The third factor distinguish two moment of the donation process: the act of donating, composed of the will to donate allowing the transplantation, and the period preceding that choice, implying the need to sign in the waiting list. The fourth factor focuses on the donor, who can remain alive, in case of sperm/ovulum donation leading to a successful generative project that constitutes the family, or who can die in case of vital organ donation. The last factor is about promotion, through a media promotional campaign or the reorganization of the Health Care System management.

According to the factorial space interpretation the sic cluster highlighted six representation of

[1] The coordinates >0.2 and >-0.2 are not reported in the table as they were not considered for the cluster interpretation.

ISBN: 978-88-5495-135-8

organ donation. The first cluster represents the *Spanish practice* showing the Spanish innovative model for the implementation of organ donation. There is a meaningful aspect of national identity in this representation, where institutions play an important role. The second cluster reflects the *human solidarity* which is possible thanks to the personal choice facing the idea of the death. It implies the active positioning and choice of the donor. The third cluster represents the possibility to donate remaining alive focusing on the sperm/ovulum donation that support other couple choices to create a new family through the act of generating a child. This cluster as the previous one focuses on the individual choice. The fourth cluster is the *Promotional Campaign*, which reflect the idea that organ donation requires information and social support. It seems that Spain wants take the responsibility to be a promoter of innovation in the public context. The fifth cluster represents donation as a *Social Challenge* related to the national pride in being the European best practice. The challenge highlights also the will to improve donation practice in the future, leading to a virtuous circle where citizens are driven to donate in order to participate to the national excellence. Finally, the sixth cluster focuses on transplantation, the surgical act of relocating a tissue, a body structure, or an organ from one body to another.

4. Conclusion

The Spanish culture is characterized by six different representation of organ donation. While one representation (*the Human Solidarity* - cluster 2) connects the donation to death, all the other representations highlight positive elements associated to life. This could be a relevant factor, which could explain the high rate of the Spanish donors, who focus more on positive elements related to this practice, as the generativity, the social challenge, and the pride of being a reference point in Europe. In the Italian culture, the symbolic space is reduced only to two factors explaining: the difference between the individual perspective and the social one, and the distinction between life and death. The Italian culture seems to lack of the positive elements characterizing the Spanish culture, focusing mainly on the death issue evoked by organ donation Finally, the use of the ETM methodology seems to support both: the understanding of a complex social phenomenon, as organ donation, highlighting the elements supporting a deep understanding of the social factors influencing people interaction and the choice to donate, and allowing the comparison among cultures.

References

Carli, R. (1990). Il processo di collusione nelle rappresentazioni sociali. *Rivista di Psicologia Clinica*, 4, pp. 282-296.

Cordella, B., Greco, F. Raso, A. (2014). Lavorare con Corpus di Piccole Dimensioni in Psicologia Clinica: Una Proposta per la Preparazione e l'Analisi dei Dati. *Actes JADT 2014. 12es Journées internationales d'Analyse Statistque des Données Textuelles*, eds E. Née, M. Daube, M. Valette and S. Fleury, Lexicometrica, Paris, pp. 173-184.

Fornari, F. (1976). *Simbolo e codice: Dal processo psicoanalitico all'analisi istituzionale.* Feltrinelli, Milano.

Greco, F. (2016). *Integrare la disabilità: Una metodologia interdisciplinare per leggere il cambiamento culturale.* Franco Angeli, Milano.

Grasso, M., Cordella, B., Pennella, A.R. (2016). *Metodologia dell'intervento in Psicologia Clinica*. Carocci, Roma.

Greco, F., Polli, A. (2019). Emotional Text Mining: Customer profiling in brand management, *International Journal of Information Management*. DOI 10.1016/j.ijinfomgt.2019.04.007

Lancia, F. (2017). *User's Manual : Tools for text analysis*. T-Lab version Plus 2017.

Lebart, L., Salem, A. (1994). *Statistique Textuelle*. Dunod, Paris.

ISBN: 978-88-5495-135-8

Matte Blanco, I. (1975). *The Unconscious as Infinite Sets: An Essay in Bi-logic*. Duckworth, London.

Monaco S., Greco F., Di Trani M., Cordella B. (2018), The culture of organ donation in the italian newspapers. *ESA-RN27 Book of Abstract, Catania (Italy), 4-6 October, 2018*, pp. 39-40.

Moscovici, S. (2005). *Le rappresentazioni sociali*. Il Mulino, Bologna.

Savaresi, S.M., Boley, D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, 8(4), pp. 345-362.

Validation of a food insecurity scale through structural equation models

Elena Grimaccia[a], Alessia Naccarato[b]
[a] Istat, and Department of Economics, University of Roma Tre, Rome, Italy.
[b] Department of Economics, University of Roma Tre, Rome, Italy.

1. Introduction

Food insecurity is at the basis of sustainable development, and its measurement is still a policy and academic issue. FAO developed the Food Insecurity Experience Scale (FIES) that presents important advantages: FIES refers to individuals, making it possible to identify the characteristics of food insecure people, and allows to analyse food insecurity also in rich and developed countries (Nord et al. 2016). FIES has been included among the indicators used to monitoring Goal 2 of the Sustainable Development Goals. Experiential measures capture cross-cultural aspects of food insecurity (Coates 2013), but studies on the validation of households' food insecurity scales are still restricted to the appraisal of a few aspects of reliability and validity (Marques et al. 2014). The aim of the study is to test the internal consistency, convergent and construct validity of FIES, and it is the first to use CFA, based on the structural equation model (SEM) methodology. The methodological properties of FIES were evaluated on a sample of 150 thousand people, all over the world. Two latent constructs were identified and analysed, while external validity has been evaluated by a micro econometric analysis of the relationship of FIES, extreme poverty and other factors upon food insecurity.

2. Methods

FIES data were surveyed in 147 countries all over the world, and they provided the first nationally representative data on food insecurity at the individual level for a very large number of countries. Surveys were conducted on samples representative of the male and female resident population aged 15 and over (Gallup 2017). The dataset was, then, composed by 150,000 adults, and included the FIES's eight items together with other meaningful social, economic and demographic characteristics of respondents. A measure has high reliability if it produces similar results under analogous conditions (Mohajan 2017). The analysis of frequency distributions of the eight items allowed to perform the first evaluation of the scale: whether the items were ordered according to the foreseen ascending order of severity. The reliability of the FIES was evaluated using Cronbach's coefficient alpha (Cronbach 1951), with the convention of the value 0.70 indicating a minimally reliable scale. CFA is a type of structural equation modelling that deals with the relationships between observed measures or indicators (test items, in our case) and latent factors (Brown and Moore 2013). A factor is an unobservable variable that influences observed variables and which accounts for their correlations. The linear coefficient that represents the effect of the latent factor on each item must be estimated, taking into account the relations among all the considered items. To evaluate the validity of FIES, we used the results of exploratory factor analysis (EFA) to postulate the relationship pattern a priori, then tested the hypothesis statistically (Thompson 2004) through CFA.

3. Results

The results of the analysis of frequencies show that indeed more respondents answered affirmatively to the items indicating less severe food insecurity than those more severe (Table 1).

For the FIES, Cronbach's alpha was 0.927, therefore we could deduce that FIES's internal consistency was excellent. The application of EFA (Principal Component Analysis -Varimax

ISBN: 978-88-5495-135-8

rotated method, SYSTAT software) showed that the two factors explained almost 80% of the total variance (Table 1). Analysing the corresponding eigenvalues, the best choice has been to consider two factors (Costello and Osborne 2005).

Table 1: Exploratory factor analysis – Frequencies and rotated loading matrix (VARIMAX, gamma =1.0000)

Items	**Questions**	**Frequencies**	**Factor Loadings**	
	During the last 12 months, was there a time when	(% yes)	**1**	**2**
HEALTHY	Q1. You were worried you would run out of food because of a lack of money or other resources?	33.3	0.843	0.297
FEWFOOD	Q2. You were unable to eat healthy and nutritious food because of a lack of money or other resources?	31.4	0.843	0.305
WORRIED	Q3. You ate only a few kinds of foods because of a lack of money or other resources?	32.9	0.819	0.293
ATELESS	Q4. You had to skip a meal because there was not enough money or other resources to get food?	21.6	0.658	0.555
SKIPPED	Q5. You ate less than you thought you should because of a lack of money or other resources?	26.2	0.515	0.683
WHLDAY	Q6. Your household ran out of food because of a lack of money or other resources?	20.7	0.157	0.865
HUNGRY	Q7. You were hungry but did not eat because there was not enough money or other resources for food?	18.1	0.388	0.806
RUNOUT	Q8. You went without eating for a whole day because of a lack of money or other resources?	12.2	0.478	0.718
'Variance' Explained by Rotated Components			3.192	2.954
Percent of Total Variance Explained			39.905	36.921

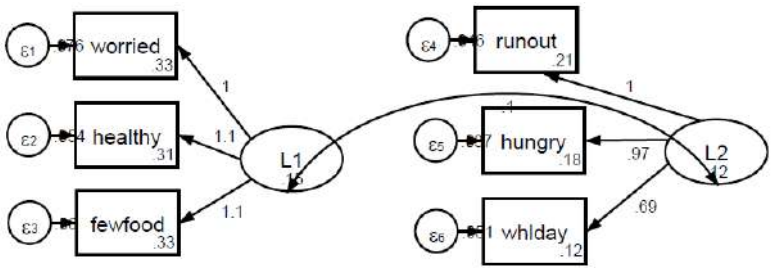
Source: Authors' elaborations on FIES data

The two factors referred to different features of food insecurity: the first component was related to perceptions and personal evaluations (being worried about not having enough food, not eating nutritious and healthy food or eating less food than desired); the second was about actual experiences, such as not eating for a whole day, feeling hungry or running out of food. EFA, then, showed a dissimilarity in respondents who effectively ate less, and those who perceived a form of food insecurity but had actually eaten. These results suggested that the share of the sample that answered positively to the first subscale composed of the 'perceived' items did not answer positively to the questions related to actually experienced food insecurity. This denotes that FIES was indeed built coherently. In order to point out the most relevant variables, as suggested in the literature (Costello and Osborne 2005), because of the high correlation between items, we have fixed a threshold for the factor loadings equal to 0.7. Therefore, the two variables 'ateless' and 'skipped' cannot be considered in determining the latent construct, because they present a lower communality on both axes that is below the threshold value. In order to verify the results of EFA, a CFA has been applied, considering two latent constructs: the first related to personal and perceived aspects of food insecurity, composed by the first three items: 'worried', 'healthy' and 'fewfood', and the second latent factor given by the items 'runout', 'hungry' and 'whlday', related to more quantitative issues of individual food insecurity. Therefore, a two factors model was tested using a SEM for two latent variables:

$$\begin{cases} worried = const_w + \beta_w L_1 + \varepsilon_w \\ healthy = const_h + \beta_h L_1 + \varepsilon_h \\ fewfood = const_f + \beta_f L_1 + \varepsilon_f \\ \\ runout = const_r + \beta_r L_2 + \varepsilon_r \\ hungry = const_h + \beta_h L_2 + \varepsilon_h \\ whlday = const_{wh} + \beta_{wh} L_2 + \varepsilon_{wh} \end{cases} \quad (1)$$

where L1 represented the latent factor related to the first subscale (perceived aspects of food insecurity) and L2 was the latent factor for the second subscale (actual experiences of food insecurity). The coefficients estimated in the model (STATA software) with two factors with interaction presented higher values compared to the model with one latent construct (Figure 1).

Figure 1: SEM diagram and coefficients (two factors model with interaction)



Source: Authors' elaborations on FIES data

The two factors model presented better goodness of fit statistics than a single scale, because it presented lower values for the Root mean squared error of approximation, the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), and values higher than 0.95 for the Comparative Fit Index and the Tucker-Lewis Index.

Convergent validity refers to the degree to which two measures of constructs that theoretically should be related are indeed related. In our case, a more established measure available at the micro level to be analysed together with the FIES score, is "extreme poverty" (less than 1.25 US dollars a day), which is an objective situation that could determine difficulties in accessing to enough food. The relation between FIES and extreme poverty has been verified: the odds ratios of an ordered logistic model, where the FIES score is the dependent variable and extreme poverty is the independent one - supposing that the causal relation is that a lack of money prevents people from accessing food - showed that not being extremely poor significantly decreased the probability of having one or more symptoms of food insecurity. External validity measures whether causal relationships can be generalised to different measures, persons, settings and times (Mcleroy 2008). We evaluated the external validity by analysing the FIES scale together with other meaningful covariates. In Grimaccia and Naccarato (2019), an extensive analysis of the factors related to the experience-based individual food insecurity is presented, according to the following model:

$$fies = \alpha c + \beta_1 \text{ gender} + \beta_2 \text{ age} + \beta_3 \text{ age square} + \beta_4 \text{ location} + \beta_5 \text{ extreme poverty} + \beta_6 \text{ marital status} + \beta_7 \text{ number of children} + \beta_8 \text{ education} + \beta_8 \text{ Region} + \varepsilon \quad (2)$$

In the present study, the analysis has been replicated for the subscales identified above, in order to verify if also the two scales based on a reduced number of selected items presented significant relations with meaningful covariates. The dependent variable for the first subscale was "perceived FIES", that was computed by summing up the affirmative answers to the first three items. The same multivariate set-up has been estimated for the second subscale of the "actual experience FIES". This measure was computed by summing the affirmative answers to the last three items of the FIES. The estimations of the model through an ordered logistic regression showed that the two subscales indeed work well in representing the relations with economic, social and demographic factors that have an impact on 'perceived' and 'actual experienced' food insecurity. Education appeared as the most important driver against food insecurity, measured by all the three measures. Household with children and widowed or divorced people appeared to experience food insecurity at a larger extent than other kind of families, both according to the "perceived" and the "actual" measures. Specifically, an additional child in the household was associated with a 0.88 percentage point higher probability of experiencing food insecurity in general, while the perceived food insecurity increased by 0.94 points and the probability of actual events by 0.64 points. Interestingly, the two sub-scales model presented a higher value of the coefficient of determination R^2, and lower values of the AIC and BIC Indexes. These results suggested that shorter and more specific subscales could work as well as a single scale to measure individual food security, and at the same be even more informative, measuring separately two different aspects of individual food security: the perceived and the actually experienced.

4. Conclusions

This study presents original results obtained through CFA applied at the global level, since it is the first to assess the reliability and validity of FIES using methodologies based on individual data, referring to 150,000 interviews, considering all the diverse issues related to the reliability, dimensionality, cumulability, and internal and external validity of FIES at the micro level across the globe. This is highly important because the scale is used at the global level to monitor development targets (SDG2). The FIES presented a good level of reliability and internal consistency, so it can be employed successfully, and the measure of food insecurity associated with a respondent can be calculated by the number of positive responses to the items. However, the cumulability of the scale was not perfect. Moreover, the results of the EFA and CFA models indicated that two items could be deleted without losing fundamental information, in order to reduce the statistical burden on respondents and to improve the cumulability of the scale. The subscale measuring 'perceived' aspects of food insecurity and the subscale related to 'actual" activities could be measured separately. Through the application of a CFA including the covariance of the two subscales, we verified that the 'subjective' and 'objective' constructs are indeed related, and future research should be devoted to building a composite index of individual food insecurity, with the aim of combining the two subscales. This would allow for a more precise measure of the two latent constructs and provide a single measure of experience-based food insecurity.

References

Brown, T. A., Moore M. T. (2013). Confirmatory Factor Analysis, in *Handbook of Structural Equation Modeling*, eds. Rick H. Hoyle. The Guilford Press, pp. 361-379.

Coates, J. (2013). Build it back better: Deconstructing food security for improved measurement and action. *Global Food Security* **2**, pp. 188-194.

Costello, A. B., Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10(7). ISSN 1531-7714.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, pp. 297-334.

Flora, D. B., Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, **9**(4), pp. 466-491.

Gallup (2017). *Worldwide research methodology and codebook*. Gallup.

Grimaccia, E., Naccarato, A. (2019). Food Insecurity Individual Experience: A Comparison of Economic and Social Characteristics of the Most Vulnerable Groups in the World. *Social Indicators Research*, **147**(1), pp. 391-410.

Marques, E. S., Reichenheim, M. E., de Moraes, C. L., Antunes, M. M., Salles-Costa, R. (2014). Household food insecurity: a systematic review of the measuring instruments used in epidemiological studies. *Public Health Nutrition*, **18**(5), pp. 877-892.

Mcleroy, K. (2008). The importance of external validity. *American Journal of Public Health*, **98**(1).

Mohajan, H. (2017). Two Criteria for Good Measurements in Research: Validity and Reliability. *Annals of Spiru Haret University*, **17**(4), pp. 56-82.

Nord M., Cafiero, C., Viviani, S. (2016). Methods for estimating comparable prevalence rates of food insecurity experienced by adults in 147 countries and areas. *Journal of Physics: Conference Series* 772.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC, US: American Psychological Association.

ISBN: 978-88-5495-135-8

A Mixture Model with Discrete Variables for Depression Diagnosis in Infertile Couples

Maria Iannario[a], Domenico Vistocco[a], Maria Clelia Zurlo[a]
[a] Department of Political Sciences, University of Naples Federico II, Naples, Italy.

1. Introduction and motivation

Infertility is a major psycho–social crisis as well as being a medical problem. The factors that predict psycho–social consequences of infertility may vary in different gender, education level, socio-economic status. The primary purpose of this study was to investigate the relationship between socio-demographic characteristics and levels of depression and anxiety in infertile couples by exploring the role of each partner and of the related perceived levels of depression and of quality of dyadic adjustment.

The perception of depression and/or anxiety are typically evaluated through latent components. This paper analyses these components by means of a mixture model for ordinal rating responses, allowing for uncertainty in answering. In responding to rating questions, indeed, an individual may give answers either according to her/his feeling or to her/his level of indecision, typically motivated by a response style. Since ignoring the uncertainty may entail misleading results, we define the distribution of the ordinal responses via a mixture model which weights both components in answering. The study allows also to model the actor/partner interdependence in case of categorical dyadic data (Kenny et al., 2006) by presenting an alternative approach with respect to the current used methods.

The effectiveness of the model is attested through the analysis of a cross-sectional study of infertile couples. The research aims to test and to evaluate the effects that some aspects linked to the couple's relationship in infertile dyads have on depressive experience for both partners. It points out the role played by marital adjustment perceived by the wives in the definition of depressive symptoms of husbands, and vice versa. Specifically, the study is designed to measure *interdependence within interpersonal relationship*, that is when one person's emotion, cognition, or behaviour affects the emotion, cognition, or behaviour of the partner (see Kelley and Thibaut (1978), among others). One of the consequences is that observations of the two individuals are linked or correlated such that knowledge of one person's score provides information about the other individual's score.

2. The survey: design and data

Data stem from a survey conducted in medically assisted procreation centers in a period of about two years, from 2014 to 2016. The sample concerns 206 infertile couples who attended clinics for treatment of their infertility problems. The average age of the couples is 34 years. The 31.5% of the sample has a female infertility problem, in 27.7% of cases the lack of a baby can be attributed to man. The 24.8% has a mixed diagnosis, however the 16.0% does not know the reason of the infertility. The questionnaire included, among others, the following scales: Dyadic Adjustment Scale, the Edinburgh Depression Scale and the State-Trait Anxiety Inventory for the evaluation of the perceived levels of psychological disease. For further details, see Zurlo et al (2017, 2018). The measurement of the status of depression has been performed by means of the second scale. Figure 1 depicts the scores of the couples interviewed: the two columns refer to female (left) and male (right), while the rows correspond to the diagnosis. In particular *both*

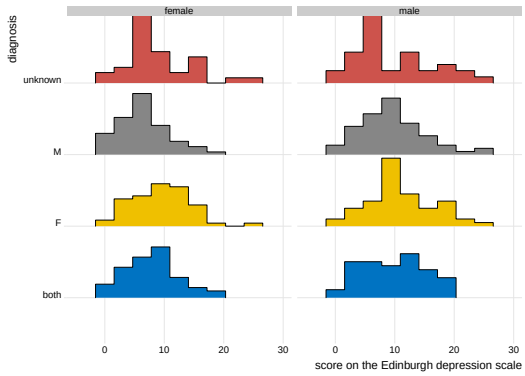


Figure 1: Score on the depression scale.

refers to the cases of mixed diagnosis, *unknown* to the case in which the reasons of the infertility are unknown, while *F* and *M* to the cases of female and male infertility, respectively. The scores have been discretized into 5 equally spaced classes (5 indicates *the worst* depressive condition) for the proposed mixture model. The assessment referred to dyad scores (Cook, 1998; Kenny et al., 2006) implies that they are nonindependent observations; thus, it is necessary to treat the dyad rather than the individual as the unit of the analysis (Kenny, 1995). The presence of nonindependence is determinated by measuring the association between the scores of the dyad members which in this paper, has been provided by means of the tests in Table 1 referred to Lang and Iannario (2013). Here, the Pearson omnibus test (X^2), the row-means Cochran-Mantel-Haenszel test (Q_1^2), the 2-moment score test (X_2^2), and the 3-moment score test (X_3^2) are reported.

Table 1: Tests of Independence for depressive condition in dyads

Approach	**Test Statistic**	*Observed Value*	**df**	**p-value**
Omnibus	X^2	42.019	16	0.000392
Restricted-Alternative	Q_1^2	10.098	4	0.038810
Relaxed-Null	X_2^2	20.945	8	0.007296
	X_3^2	26.660	12	0.008647

Commonly used statistical procedures (e.g., ANOVA and multiple regression), implemented for the analysis of this kind of data, assume independent (uncorrelated) observations in the dependent variable. Consequently, the scores of two *linked* individuals would be treated as if they were completely independent observations or analysed as the sum or the average of the the two individual scores and treat it as a *dyad score* in the analysis by presenting the complain noted in Christensen and Arrington (1987).

In the present paper we take both individual and dyadic factors into account by using a bidirectional view which would predict as each person influences the other. We treat the ordinal score of the partner as predictor variable (explanatory covariate) of the dependent Y of the other member of the couple. Thus, actor effects are estimated controlling for partner effects and viceversa in two separate models obtained by means of the implementation of the CUB mixture (Piccolo, 2003). The model allows to take into account the feeling and the uncertianty expressed by each member of the couple with respect to the depression status. The order of these two analyses does not matter. Couple is the unit of analysis, so the independence assumption is not violated. The interests are the magnitude of actor and partner effects in each analysis and their statistical significance. Notice that a partner effect for each partner must be statistically significant to support the hypothesis that influence is bidirectional. It may be tested by means of the common local tests of validation. Additionally, the model also includes other contextual variables or other factors that are not personal characteristics of either partner.

3. Model based results

The 'best' estimated models are reported in Table 2. Here, the direction of the arrows indicates the effects (positive or negative) that the covariates exert on feeling component, in relation to depression. Notice that that there are no significant covariates in estimating the component of uncertainty. The latter can be considered as the result of a number of converging factors that summarize the interest/disinterest of respondents to the items considered. It is believed that the interviewed couples have replied with pleasure to the questionnaire, mainly with the objective to help future patients to understand and consciously address the difficulties related to a possible cycle of medically assisted procreation. Moreover, the presence of a specialized interviewer minimized the errors from lack of understanding of questions.

Table 2: Significant covariates for the feeling component

Response variable	**Significant covariates**	**BIC index**
Male depression [1]	Male anxiety ↑, Male diagnosis ↑, Female depression ↑	*435.78*
Male depression [2]	Male anxiety ↑, Female diagnosis ↓, Female depression ↑	439.03
Female depression [1]	Female anxiety ↑, Female education ↓,	*512.82*
	Male depression ↑, Male dyadic satisfaction ↑	
Female depression [2]	Female anxiety ↑, Female work ↓,	521.10
	Male depression ↑, Male dyadic satisfaction ↑	

The model for male depression [1] (in Table 2) explains male depression as a direct function of personnel anxiety, level of female depression and the situational variable (a dummy) concerning "male diagnosis" (male responsible for infertility). According to the latter variable, it is interesting to note that the estimated coefficient has a negative sign ($\gamma_2 = -0.559$), that is, the risk of depression in infertile men tends to increase when he finds out to be the cause of the lack of a son. Notice that the feeling is measured by $log(1 - \xi_i) = -\boldsymbol{x}_i\boldsymbol{\gamma}$ in the CUB model where $\boldsymbol{x}_i$ represents the information set extracted from the matrix $\boldsymbol{X}$ and $\boldsymbol{\gamma}$ is the parameter vector.

The second model for male depression [2] differs from the first one because of the presence of the situational variable (dummy) concerning "female diagnosis". The signs and the values of the estimated coefficients are almost the same, except for the dichotomous variable "diagnosis of infertility" ($\gamma_2 = 0.461$). The positive sign indicates, in fact, that the husband tends to be less depressed when he knows he is not the cause of infertility.

According to the estimated female model [1], the risk of female depression tends to increase if accompanied by a personnel anxiety component and a depressive situation in the partner. Furthermore, the perception of satisfaction for the latter directly influence the level of female depression, the less the husband is satisfied, the more the emotional life of the woman weakens. Model [1] emphasizes also the role of education for women: depression reduces with high qualified female; whereas model [2] indicates the significant role of job which reduces the perceived levels of depression.

Among the selected models for male and female depression, the Bayesian Information Criterion (BIC) suggests the choice of models marked with [1] for both the male and female ordinal variables.

By further inspecting the female depression model [1] (estimation results for the feeling component are in Table 3), we have analysed some risks profiles. We calculated, as an instance, the probability that a woman with a slight level of anxiety, married to a man who has depression levels lower than national standards (validated by *Edinburgh Depression Scale*) and high level of satisfaction perceived into the couple, takes different modal value, when her degree of study changes (left panel of Figure 2; the opposite profile in the right panel). Other different profiles may be of course observed by varying the levels of covariates. In the selected profiles, it is

Table 3: Estimation results for the Female depression CUB model [1]

Component	Covariates	ML-estimates	Stand.errors	Wald-test
Feeling	Constant	3.760	0.738	5.098
	Female anxiety	−1.041	0.132	−7.866
	Female education	0.399	0.112	2.900
	Male depression	−0.221	0.084	−2.622
	Male dyadic satisfaction	−0.045	0.018	−2.506

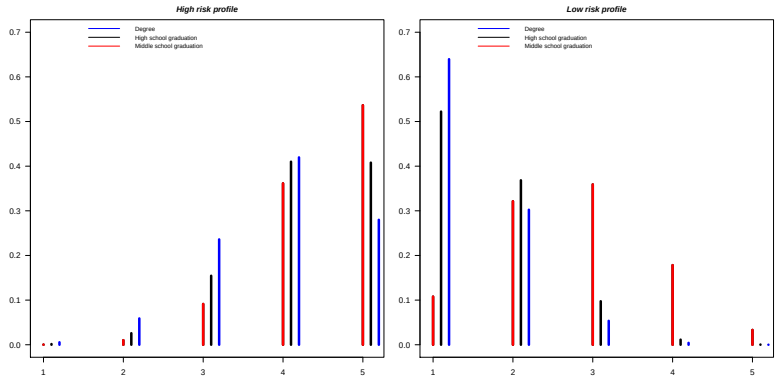


Figure 2: Risk profiles for female depression.

clear the protective role deriving from cultural training. For infertile women, *ceteris paribus*, a higher level of education corresponds lower suffering from depressive disorder. This aspect is also underlined for high risks profile (right panel of Figure 2). Finally, there are other models of dyadic relationships that correspond to other forms of dyadic non-independence. One of the aim for future researches is the comparison of the mainly used approaches and the the introduction of a new model based on the development of the presented experience.

References

Christensen, A., Arrington, A. (1987). Research issues and strategies. in T. Jacob (Ed.), *Family interaction and psychopathology: Theories, methods, and findings*, (pp. 259–296). New York: Plenum Press.

Cook, W. L. (1998). Integrating models of interdependence with treatment evaluations in marital therapy research. *Journal of Family Psychology*, 12, 529–542.

Kelley, H. H., Thibaut, J. W. (1978). *Interpersonal relations: A theory of interdependence.* New York, Wiley.

Kenny, D. A., Kashy, D. A., Cook, W. (2006). *Dyadic data analysis*. New York: Guilford..

Kenny, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships*, 2, 67–75.

Lang, J.B., Iannario, M. (2013). Improved tests of independence in singly-ordered two-way contingency tables. *Computational Statistics and Data Analysis*, 68, 339–351.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85–104.

Zurlo, M.C., Cattaneo Della Volta, M. F., Vallone F. (2017). Factor structure and psychometric properties of the Fertility Problem Inventory-Short Form. *Health psychology open*, pp. 1-11.

Zurlo, M.C., Cattaneo Della Volta, M. F., Vallone F. (2018). Predictors of quality of life and psychological health in infertile couples: the moderating role of duration of infertility. *Quality of life reaserach*, doi: 10.1007/s11136-017-1781-4

Three-way log-ratio analysis for assessing sport performance

Rosaria Lombardo [a], Ida Camminatiello [a], Antonello D'Ambra [a]
[a] Department of Economics, University of Campania 'Luigi Vanvitelli', Capua (CE), Italy

1. Introduction

Among various measures for categorical variables, the odds ratio represents a simple and very popular measure of association between two or more categorical variables. In literature, a large amount of models and methods have been developed around this measure (Andersen, 1980; Goodman, 1985; Agresti, 1996; Kateri, 2014). When modeling association structure by log-linear models, it can be found a functional relationship between the model parameters and odds ratios, as well as when using log-ratio analysis (Aitchison and Greenacre, 2002) or RC(M)-association models (Goodman, 1985). Over the last decade, some methods have been proposed (De Rooij and Heiser, 2005; De Rooij and Anderson, 2007; Sarnacchiaro et al., 2014) for describing the odds ratios in terms of point distances and inner products in a graphical representation of the variable association.

Here our focus is on studying and portraying the association among three categorical variables by using odds ratios. We propose a generalization of the log-ratio analysis presented by Aitchison and Greenacre (2002) for three-way contingency tables, based on Tucker's three-way decomposition model (Tucker, 1966; Kroonenberg, 2008). We call this generalization three-way log-ratio analysis (TLRA) which has the benefit to graphically visualize the log transformation of odds ratios computed for three-way contingency tables. As an intersting finding of TLRA, the visual interpretation of the association among the three sets of variables will be done using a distance and an inner product rule rested upon odds ratios (De Rooij and Heiser, 2005). This paper is divided into three further sections. After introducing the notation in Section 2, Section 3 provides some of the important theoretical issues concerning the three-way log-ratio analysis and its properties. Finally, Section 4 shows the main findings of the three-way log-ratio analysis when analyzing the sport perfomance of some Italian football clubs.

2. Notation

Consider a three-way contingency table that is formed from the cross-classification of n subjects according to three categorical variables, X, Y and Z, that are referred to hereafter as the row, column and tube variables. These variables consist of I rows, J columns and K tubes, respectively. Denote $\underline{\mathbf{P}}$ to be the joint relative frequency table whose $(i,\, j,\, k)$th term is p_{ijk}, such that $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk} = 1$. Define the row, column and tube marginal frequencies $p_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}$, $p_{\bullet j\bullet} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ijk}$ and $p_{\bullet\bullet k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}$, respectively. Also let $\mathbf{D}_I$, $\mathbf{D}_J$ and $\mathbf{D}_K$ be the corresponding diagonal matrices of marginal frequencies. The conditional odds ratios (OR) between X and Y for a given level k of the tube variable (Kateri, 2014, p. 67) is given by

$$ODDS_{i'i,j'j}(k) = \frac{n_{ijk}n_{i'j'k}}{n_{i'jk}n_{ij'k}} \qquad \text{for} \quad (1 \le i \le i' \le I) \quad \text{and} \quad (1 \le j \le j' \le J)$$

ISBN: 978-88-5495-135-8

Furthermore, a generalisation of Altham's index (1970) for three-way contingency tables can be seen as (De Rooij and Anderson, 2007)

$$\Theta_{A_W} = \left\{ \sum_{i'=1}^{\hat{I}} \sum_{i=1}^{\hat{I}} \sum_{j'=1}^{\hat{J}} \sum_{j=1}^{\hat{J}} \sum_{k=1}^{K} p_{i\bullet\bullet} p_{i'\bullet\bullet} p_{\bullet j\bullet} p_{\bullet j'\bullet} p_{\bullet\bullet k} \left(logODDS_{i'i,j'j}(k) \right)^2 \right\} \tag{1}$$

where $\hat{I} = I(I-1)/2$ and $\hat{J} = J(J-1)/2$. For an $I \times J \times K$ contingency table under the model of independence, all the local odds ratios are equal to 1.

3. Three-way log-ratio analysis

The three-way log-ratio analysis (TLRA) can be seen as a generalization of the log-ratio analysis for two-way contingency tables (Aitchison and Greenacre, 2002). TLRA is based on the Tucker3 decomposition method (Tucker, 1966; Kroonenberg, 2008) which is a three-way generalisation of the singular value decomposition. The Tucker3 model involves the computation of *principal components* for each of the three categorical variables and of a *core* array whose elements are generalisation of the singular values. Three-way log-ratio analysis considers the log trasformation of a three-way centred and weighted table of proportions, that is given by

$$\mathbf{S} = \left[\mathbf{D}_I^{1/2} (\mathbf{I}_I - \mathbf{D}_I) \right] log(\mathbf{P}) \left[(\mathbf{I}_J - \mathbf{D}_J) \mathbf{D}_J^{1/2} \otimes \mathbf{D}_K^{1/2} \right]$$

applying the Tucker3 decomposition, we model the dependence among variables such that

$$\text{Tucker3}\,(\mathbf{S}) = \mathbf{AG} \left(\mathbf{B}^T \otimes \mathbf{C}^T \right) + \mathbf{E}\,. \tag{2}$$

In Equation 2, $\mathbf{G}$ is the core array arranged into a two-way form of size $P \times QR$. The set of matrices $\mathbf{A}$ $(I \times P)$, $\mathbf{B}$ $(J \times Q)$ and $\mathbf{C}$ $(R \times K)$ are orthonormal with respect to the weight matrices $\mathbf{D}_I$, $\mathbf{D}_J$ and $\mathbf{D}_K$, respectively. The term $\mathbf{E}$ is the error of approximation that depends on the model dimension. It is worth noting that the inertia or global association among variables is equal to the sum of squares of the elements of the core array and is akin to the Altham (1970) index, i.e. $\Theta_{A_W} = \left(\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr}^2 \right)$. When visualizing the association, the *Interactive Biplot* (Kroonenberg, 2008) will be considered. Doing so, the column and tube variables will be interactively coded and represented by a single point in the graphical display. Similarly to weighted two-way LRA (Sarnacchiaro et al., 2014), the row and column-tube coordinates of a weighted three-way LRA have the following two important properties: 1) the sum of squared elements of the principal coordinates approximates the weighted version of Altham's index ., i.e. $\Theta_{A_w}(2) \approx trace \left(\mathbf{F}^T \mathbf{D}_I \mathbf{F} \right) = trace \left(\mathbf{H}^T \left(\mathbf{D}_J \otimes \mathbf{D}_K \right) \mathbf{H} \right)$; 2) the odds ratio can be reconstructed using the distances between the row principal coordinates $\mathbf{F}(= \mathbf{f}_i)$ and the column-tube standard coordinates $\mathbf{H}(= \mathbf{h}_{jk})$, it is verified that

$$OR_{ii',jj'} = exp \left[1/2 \left(d(\mathbf{f}_{i'}; \mathbf{h}_{jk})^2 + d(\mathbf{f}_i; \mathbf{h}_{j'k'})^2 - d(\mathbf{f}_i; \mathbf{h}_{jk})^2 - d(\mathbf{f}_{i'}; \mathbf{h}_{j'k'})^2 \right) \right] \tag{3}$$

where $d(\mathbf{f}_i, \mathbf{h}_{jk})^2$ is the squared Euclidean distance between the row variable in standard coordinates and the interactively coded column-tube variables in principal coordinates when using all the model dimensions.

4. Assessing sport performance

In this section, we briefly report some results of the three-way log-ratio analysis applied to players' performance of Italian football teams in 2017. In Table 1, we look at a $4 \times 2 \times 3$ contingency table that summarizes the football players' performance, according to three variables, *height, ball-control* and *football club* of players.

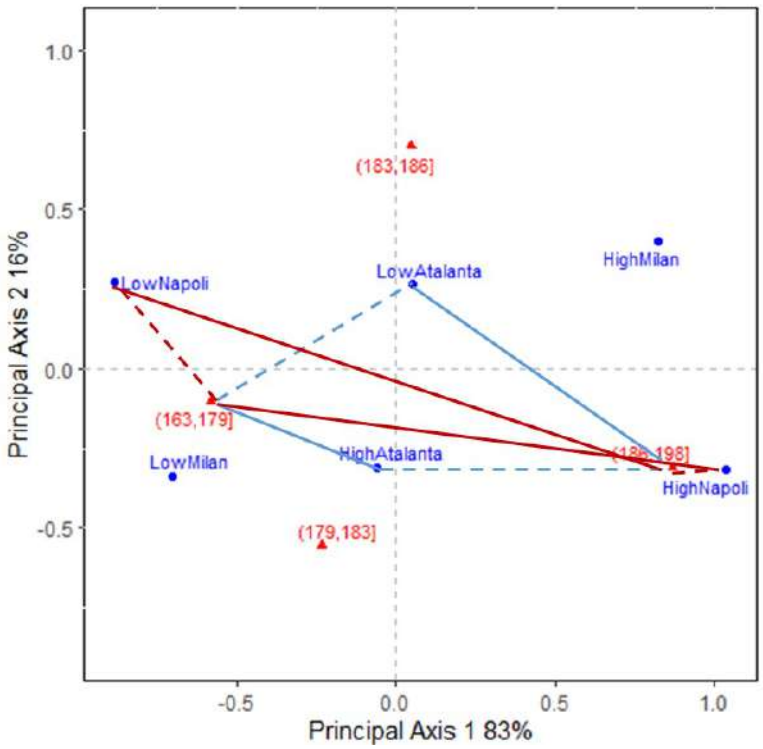


Figure 1: The biplot from the log-ratio analysis of Table 1

The row variable, *height* has four categories, that are *h1*= (163,179], *h2*= (179,183], *h3*= (183,186] and *h4*= (186,198]. The column variable is the dependent one and concerns the performace's players, in particular the *ball control*, that is coded in two categories *Low* ball control and *High* ball control. The tube variable represents the Italian football clubs, in particular we look at *Atalanta*, *Milan* and *Naples* teams.

	Atalanta		Milan		Naples	
Height	*Low*	*High*	*Low*	*High*	*Low*	*High*
$h1 = (163, 179]$	1	2	4	1	6	1
$h2 = (179, 183]$	1	3	3	1	5	2
$h3 = (183, 186]$	3	4	4	5	3	1
$h4 = (186, 198]$	2	4	1	2	1	3

Table 1: Height by Ball-control by Football Club in Italy, 2017

Figure 1 is the interactive biplot of the three variables of Table 1. Here, the players' height is in standard coordinates, while the two variables, *ball-control* and *football team* are coded interactively, forming the column-tube combined variable that is in principal coordinates. The odds ratios are represented numerically in Table 2. Their log-trasformation is also visualized in Figure 1 by means of point distances, when the solid lines are longer than the dashed lines,

	Atalanta	Milan	Naples
Height	*Low/High*	*Low/High*	*Low/High*
h1h2	1.500	1.333	2.400
h1h3	0.667	5.000	2.000
h1h4	**1.000**	8.000	**18.000**
h2h3	0.444	3.750	0.833
h2h4	0.667	6.000	7.500
h3h4	1.500	1.600	9.000

Table 2: Conditional odds ratios. In bold the two odds ratios graphically visualized in Figure 1.

then the odds ratio will be higher than one. Conversely, the odds ratio is less than one, when the dashed lines are longer than the solid lines; see Equation 3. For example for *Atalanta* team, we get an odds ratio equal 1, when comparing the performance of the smaller players against the taller players (see Table 2 *h1h4* = 1), that means that the smaller *Atalanta* players get the same *ball-control* of the taller players, indeed in Figure 1, the length of the blue solid lines is equal to the length of the dashed lines. Differently for *Naples* players, the red solid lines are longer than the red dashed ones, that means the odds is very high, i.e. the smaller *Naples* players get an high *ball-control* with respect to the taller players.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

Aitchison, J. and Greenacre, M. J. (2002). Biplots of compositional data. *Applied Statistics* **51**, pp. 375–392.

Altham, P. M. E. (1970). The measurement of association of rows and columns for an r x s contingency table. *Journal of Royal Statistical Society, Series B***32**, pp. 63–73.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland.

De Rooij, M. and Anderson, C. J. (2007). Visualizing, summarizing, and comparing odds ratio structures. *Methodology*, **3**, pp. 139–148.

De Rooij, M., Heiser, W. J. (2005). Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika*, **70**, pp. 99–123.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetric models for contingency tables with or without missing entries. *The Annals of Statistics*, **13**, pp. 10–69.

Kateri, M. (2014). Contingency Table Analysis: Methods and Implementation Using R. Birkhäuser, Basel.

Kroonenberg, P. M. (2008). Applied Multiway Data Analysis. Wiley, Hoboken, NJ.

Sarnacchiaro, P., D'Ambra, L., Camminatiello, I. (2014). Measures of Association and Visualization of Log Odds Ratio for a two-way contingency table. *Australian and New Zealand Journal of Statistics*, **57**(3), pp. 363–376.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**, 279–311.

Assessment of game actions performance in water polo: a data analytic approach

Alessandro Lubisco[a], Stefania Mignani[a], Carlo Trivisano[a]
[a] Department of Statistical Sciences, Alma Mater Studiorum Università di Bologna, Italy

1. Introduction

Water polo is an Olympic 7 vs 7 team sport. The team that scores the higher number of goals wins the match. This sport has a long tradition all over the world, especially in Europe. The best national teams come from East Europe (Hungary, Croatia, Montenegro, Serbia), but commonly other teams are also very competitive, like Spain and Italy (who won the most recent World Championship in 2019). Men's water polo was introduced to the Paris Olympic Games in 1900 as the first team sport; the first women's matches were not played until the Sydney Games in 2000 (Escalante, 2011). Each match is split into four periods of 8 real-play minutes, played in a 30×20m field with a minimum depth of 2 meters. This depth enables tall players attacking from the so-called hole set position to benefit by touching the bottom of the pool. In this way they can handle the pressure of the defender. Touching or jumping off the bottom of the pool is illegal, except for the goalkeeper. Referees will not typically call it unless it results in a clear advantage.

It makes no sense for the goalkeeper or players in other positions to try to take advantage from jumping off the bottom of the pool to block a shot or to shoot a ball.

Water polo is a highly physical and demanding sport; indeed it is famed as one of the toughest. This sport demands high levels of energy from the players and good play strategies from the coach not only in choice of tactics, but also in player changes to allow effective recovery. Consumption of energy could affect the clarity needed for rapid perception of game situations. This is a sport where tactical thinking and teamwork are fundamental. For these reasons, water polo players are subjected to very taxing training that is incomparable with what is required from most other sports.

One characteristic, which makes this sport unique, is that playing in water means that players observe the match with their eyes very close to the water surface and may move from one position to another in a completely different way compared to other sports. Nevertheless, as with all the other team sports, all schemes studied with an "aerial view" on the tactic board with the coach and practised in the water must be very clear to all the players. For this reason, match analysis could also be very useful in water polo as it would provide opportunities to understand both team and individual player performances. A second aspect, which bears significant importance on the result of a match, but which is not contemplated in this work, is the role of the referees. Playing in the water means that referees are expected to understand what is happening below the surface. Grabbing the opponent's kit or hitting him or her in any way is illegal, but they are, nonetheless, frequent occurrences. Referees must be very experienced to be able to make correct assessment of what is sometimes nothing more than a theatrical performance.

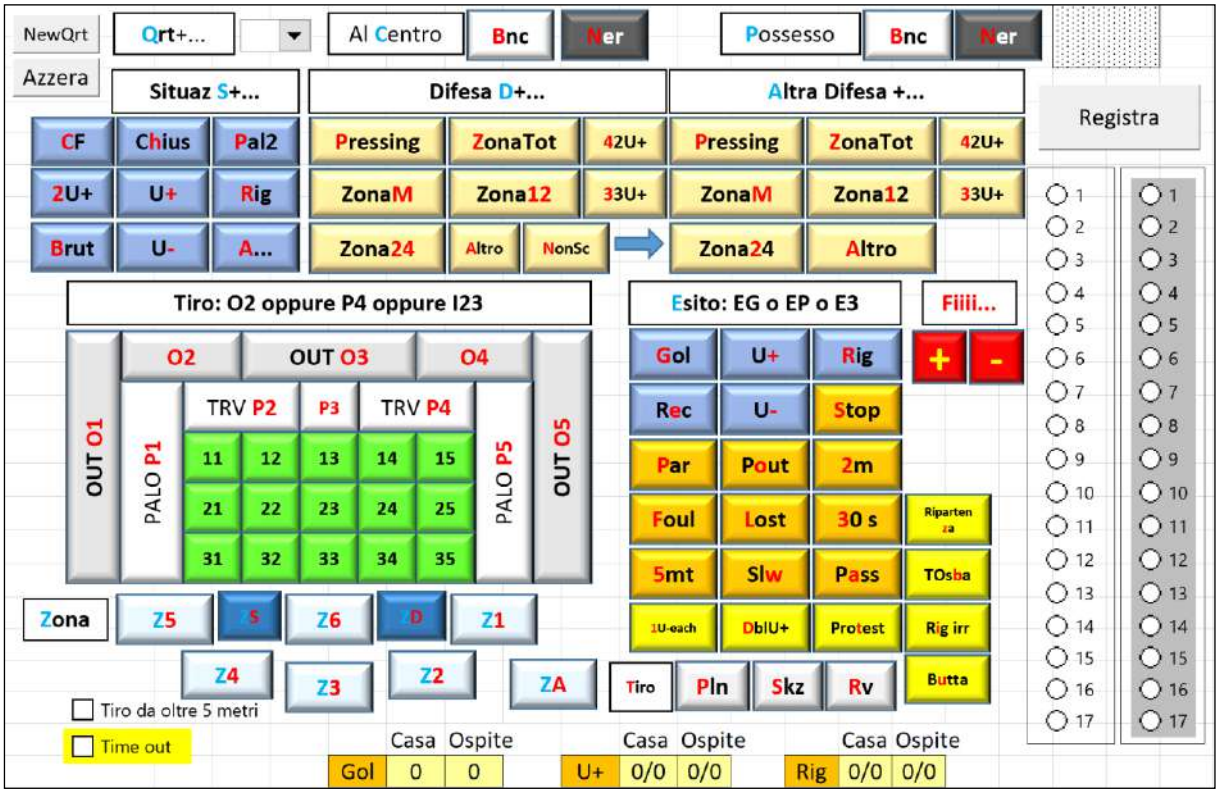
2. Data analysis

The aim of this study is to analyse the performance of President Bologna, an Italian second division water polo team, at the conclusion of the 2018 national championship. Previous studies (Graham and Mayberry, 2015; Özkol et al., 2013) are more often dedicated to matches in major tournaments (Olympics, World and European Championship). In this work, we compared the team's offense and defence performances in 17 matches (12 home and 5 away). Data were collected throughout each match by way of an original program for water polo match analysis

ISBN: 978-88-5495-135-8

specifically designed and developed in MS Excel by the first author.

Figure 1: Input computer screen for data collection



With this interface (Italian version) it is possible to record a lot of information for each play: period, situation of play (numerical equality, man-up, counter-attack…), defence used (press, zone, …), players involved (number of cap), outcome of the play (goal, exclusion, penalty, shot, turnover, …), offensive position (wings Z1 and Z5, flats Z2 and Z4, point Z3, hole set Z6, 2 meters left post ZS and right post ZD, other position ZA), number of ordinary fouls…

The speed of some actions makes it difficult to record everything in real time so subsequent video analysis is required to supplement any undetected data. In this way, information can be added for each action such as, for example, the number of passes and the duration of the action. At the same time, data can be verified and, where necessary, corrected. This is painstaking work which sometimes requires watching the same action several times and it is not always simple to recognize the players. It takes three hours to check a match that lasts one hour.

The software creates tables and graphs automatically for both teams. Team statistics: the number of shots and goals, power play trend and performance, starting and ending points of shots. Player statistics: the number of goals scored (while attacking) and conceded (while defending), exclusion (gained and suffered), balls (lost and recovered), fouls, …

The dataset consists in more than 1800 plays. Together with the variables already described, we calculated, for each play, the difference in that moment of the match between the teams in terms of goals scored and exclusions/penalties gained. GoalDifference=3 means that the team who won the match was playing that action being 3 goals ahead. ManUpPenDiff=-2 means that the team who won the match was playing that action having gained 2 exclusions or penalties less than the loser team (Table 1). An initial descriptive analysis shows that 49.3% of the goals resulted from powerplay (41.9%) or penalty (7.4%) situations, whereas, 33.8% and 16.9% respectively resulted from even and counterattack situations. These percentages are not so different from those observed in top team international tournaments.

Table 1: List of covariates

Covariate	Type	Levels
WinnerLoser	Dummy	0=Loser, 1=Winner
Period	Categorical	P1=Period 1, P2=Period 2…
GoalDifference	Discrete	
ManUpPenDiff	Discrete	
Defense	Dummy	0=Other defense, 1=Press
OffensivePosition	Categorical	9 categories (Z1, Z2, …, ZA)
HomeAway	Dummy	0=Home, 1=Away
Situation	Categorical	4 categories (Even, S+=Powerplay, SitCntAtt =Counterattack, SitRecov=Counterattack recovered)

Table 2: Team's attack performance model

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.0429	0.4633	-4.409	0.0000	***
WinnerLoser	-0.2422	0.2524	-0.959	0.3374	
P2	0.0629	0.0441	1.426	0.1538	
P3	-0.0454	0.2265	-0.200	0.8413	
P4	0.1606	0.2363	0.680	0.4966	
GoalDifference	0.2646	0.2440	1.085	0.2781	
ManUpPenDiff	0.0569	0.0237	2.399	0.0164	*
DifesaPressing	0.1076	0.2134	0.504	0.6140	
Z2	0.1953	0.4099	0.476	0.6338	
Z3	0.4694	0.4262	1.101	0.2707	
Z4	0.5438	0.3843	1.415	0.1570	
Z5	0.9675	0.4836	2.001	0.0454	*
Z6	1.5519	0.3883	3.996	0.0001	***
ZA	1.9466	0.4647	4.189	0.0000	***
ZD	1.0627	0.5457	1.947	0.0515	.
ZS	1.4666	0.4600	3.188	0.0014	**
HomeAway	0.1437	0.1839	0.781	0.4346	
S+	1.0064	0.2649	3.799	0.0001	***
SitCntAtt	1.3611	0.3415	3.985	0.0001	***
SitRecov	0.1856	0.3643	0.509	0.6105	

We estimated two logistic regression models in R using the function glm(): the first analyses the team's attack performances; the second their defending ability. The aim of the two models is to assess which are the most important covariates affecting the outcome of play, defined as "GoodOutcome" (goal, penalty or exclusion) and "BadOutcome" otherwise (lost ball, shoot out or saved …).

At the end of the championship, the analysed team showed the best defence of the 12 teams, and the fifth best attack. The model results underline that when the team is in an offensive phase (Table 2), the probability that it will complete a play with a GoodOutcome is higher when the event occurs on the two-meter line, between the two posts (hole set and nearby). This also happens during a powerplay or counterattack situation. Although, this may seem obvious, something different happens when the team is defending (Table 3): there are only two significant coefficients. The first coefficient refers to the playing situation and the positive sign meaning that, all else being equal, the GoodOutcome is more likely for counterattack if compared with other situations. The other significant coefficient is related to the covariate

ISBN: 978-88-5495-135-8

called Z2, usually occupied by a left-handed player. This coefficient has a negative sign meaning that for the opposite team the GoodOutcome is less likely from that position compared with other positions. In fact, the strategy of the analysed team throughout the year was to force the opponents to try a shot from that position. This appeared to be a good strategy, resulting in the best defence of the championship. This model, despite suffering from heterogeneity due to the presence of many factors, is a promising tool as confirmed by a percentage of correctly classified cases of around 70% estimated using the leave-one-out method.

Table 3: Team's defending ability model

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.0192	0.4130	-2.468	0.0136	*
WinnerLoser	0.0227	0.2644	0.086	0.9315	
P2	0.3213	0.2348	1.368	0.1712	
P3	0.3327	0.2396	1.388	0.1650	
P4	0.6683	0.2378	2.811	0.0049	**
GoalDifference	0.0283	0.0427	0.662	0.5080	
ManUpPenDiff	0.1101	0.0225	4.905	0.0000	***
DifesaPressing	-0.1104	0.1980	-0.557	0.5773	
Z2	-0.8456	0.4186	-2.020	0.0434	*
Z3	-0.6825	0.4270	-1.598	0.1100	
Z4	-0.1703	0.3911	-0.435	0.6633	
Z5	-0.0801	0.5053	-0.159	0.8740	
Z6	0.2818	0.3874	0.727	0.4670	
ZA	0.5247	0.4596	1.142	0.2536	
ZD	0.5624	0.5254	1.070	0.2845	
ZS	0.6541	0.4527	1.445	0.1485	
HomeAway	-0.2687	0.1821	-1.475	0.1401	
S+	-0.2108	0.2591	-0.814	0.4158	
SitCntAtt	0.8415	0.3696	2.277	0.0228	*
SitRecov	0.2796	0.3231	0.866	0.3868	

Further developments are undeniably needed. In particular, the interaction effects of some covariates should be analysed or stepwise procedures should be adopted to select the best variables

References:

Escalante Y., Saavedra J.M., Mansilla M. and Tella V., (2011). Discriminatory power of water polo game-related statistics at 2008 the olympic games. *Journal of Sports Sciences*, **29**, pp. 291-298.

Graham J., Mayberry J., (2015). Measures of Tactical Efficiency in Water Polo. *Journal of Quantitative Analysis in Sports*, **10**(1), pp. 67-79.

Lupo C., Condello G. and Tessitore A., (2012). Notational analysis of elite men's water polo related to specific margins of victory. *Journal of Sports Science and Medicine*, **11**, pp. 516-525.

Özkol M.Z., Turunç S. and Dopsaj M., (2013) Water polo shots notational analysis according to player positions. *International Journal of Performance Analysis in Sport*, **13**(3), pp. 734-749.

Selecting Features for Machine Learning in Alzheimer's Diagnostics

Luiz Sá Lucas[a], Ana Carolina Sá Lucas[a] and Rafaela Bueno[a]
[a]theopinionedge, Rio de Janeiro, Brazil

1. Introduction

Medical artificial intelligence (AI) is moving forward at considerable pace. Promising research ideas are surfacing in clinical areas (Buch, Varughese and Maruthappu, 2018; Miller and Brown, 2018)). AI is automating, for example, triage service (Burgess, 2017)); has exhibited dermatologist-level performance at identifying suspicious skin lesions , a task where experts frequently disagree (Buch, Varughese and Maruthappu, 2018); and has, among others, applications on the diagnosis of diabetes (Contreras and Vehi, 2018)), psychosis (Bedi et al., 2015) and schizophrenia (Lyle, 2019)).

Diagnosis in Medicine is the process of determining the cause of a patient's illness or condition by investigating information acquired from various sources including physical examination, patient interview, laboratory tests, patient's and the patient's family medical record, and existing medical knowledge of the cause of observed signs and symptoms. Getting a correct diagnosis is the most crucial step in treating a patient as it allows physicians and therapists to find the best treatment for the patient's condition. However, it is a complicated process and requires lots of human effort and time. Due to that complex nature, it is error-prone. Thus, misdiagnosis is very common. According to the World Health Organization, 5% percent of the outpatient encounters were misdiagnosed in 2015. This is a worrisome, especially when people lives are at stake (Gagliano et al., 2017)).

The present article explores how machine learning may be set to detect Alzheimer early stage diagnosis. The condition was first recorded by the German psychiatrist Alois Alzheimer in 1906 after he noticed changes in the brain tissue of a patient who had died from an unusual mental illness, with symptoms including memory loss, language problems, and unpredictable behavior.

In the neurological and neuro-psychological areas, clinical criteria for the diagnosis of Alzheimer's disease include insidious onset and progressive impairment of memory and other cognitive functions. There are no motor, sensory, or coordination deficits early in the disease. Tests are important primarily in identifying other possible causes of dementia that must be excluded before the diagnosis of Alzheimer's disease may be made with confidence. To help diagnosis neuropsychological tests provide confirmatory evidence of the diagnosis of dementia and help to assess the course and response to therapy. (McKhann et al., 1984). Besides medical approaches, there are also psychological tests for the disease. Psychologists have identified several promising tests (Span et al., 2005 and Duchek and Balota, 2005): Paired-associate learning tests / Perceptual identification tasks / Dichotic listening tasks / etc. The American Psychological Association mentions interesting psychological tests (see APA website).

We should also mention image studies (Shamonin et al., 2014) as very important tools for the diagnosis. A general description of tests for early diagnostics on Alzheimer can be found in the Alzheimer Association website.

2. Machine Learning and early Detection of Alzheimer Disease

Clinicopathological studies suggest that Alzheimer's disease (AD) pathology begins 10–15 years before the resulting cognitive impairment draws medical attention. Biomarkers that can detect AD pathology in its early stages and predict dementia onset would, therefore, be extremely valuable for patient care and efficient clinical trial design. In this work, based on (Craig-Shapiro et al., 2011), we utilized a machine learning (ML) predictive modeling approach to select cerebrospinal fluid (CSF) biomarkers that can augment the diagnostic and prognostic

ISBN: 978-88-5495-135-8

accuracy of current leading CSF biomarkers (besides Ab42, tau and p-tau). Our study was influenced by a ML approach to AD (Kuhn and Johnson, 2013).

In our work we used a dataset that is included in the R software package *caret* (Kuhn et al., 2016): *AlzheimerDisease*. In this set we have 333 observations, with diagnosis as a binary dependent output variable: values are *Impaired* (91 cases) and *Control* (242 cases), So *Control* has the majority of cases (around 73%). The number of *predictors* is quite big: 132. It is clearly impossible in a daily use to work with so many variables. This led us to look for a smaller set, or, in other words, to a *feature selection* procedure. We considered:

- Selecting from sets of variables / features that were highly correlated, through the CLV technique
- ICA - Independent Component Analysis
- Finding the best set through optimization techniques
- Using features quoted in the literature together with the above techniques

We implemented the CLV technique through *ClustVarLV* R package (Vigneau, Chen and Cariou, 2019). ICA is a not a so far cousin of Principal Component Analysis (PCA), with two differences. PCA is based on correlation and ICA on independence of information (Venables and Ripley, 2002 and Stone, 2004)). Besides, PCA gives an ordering of the components, which is not the case in ICA. We applied as the optimization method Simulated Annealing (Henderson, Jacobson and Johnson, 2003, and Kuhn and Johnson, 2013). We also used a set of predictors given by the importance as measured by the *randomForest* algorithm (Kuhn and Johnson, 2013).

Another interesting feature of our problem is the fact that the training dataset is highly unbalanced (73% for *Control* and 27% for *Impaired*). In order to balance this training set we used SMOTE (Torgo, 2017)). As a predictive model we used *treebag* in *caret* R package (Hauthorn and Lauthern, 2003, and Kuhn et al., 2016). With this methodology we arrived at. 10 datasets, with the following predictors:

- *rf* : based on importance as given by randomForest - Ab_42, tau, MMP10, VEGF, Cystatin_C, p_tau, age, E4, E3 and E2
- *sa*: five main predictors given by simulated annealing: tau, VEGF, p_tau, Pancreatic_polypeptide and Apoliprotein_D
- *s1*: the "classic" ones: Ab_42, tau and p_tau
- *s2*: *s1* + Genotypes E4, E3 and E2
- *s3*: *s2* + age
- *s4*: *s2* + VEGF, PYY and Apoliprotein_D
- *s5:* those given by CLV*:* Tissue_Factor, SOD, ACTH_Adrenocorticotropic_Hormon, Angiotensinogen, Apolipoprotein_A2, Fetuin_A, TIMP_1, VCAM_1, Insulin, IL_3
- *s6: s2* + those given by ICA: Lipoprotein_a, Thrombopoietin, Ferritin, Apoliprotein_CI, MIP1alpha, Pancreatic_polypeptide, Creatine_Kinase_MB, VEGF and IP_10_Inducible_Protein_10
- *s7:* first set of predictors given in Craig-Shapiro et. al. (2011) – s1 + CD5L +age
- s8: second set given by Craig-Shapiro et. al. (2011) – Cystatin_C, VEGF, TNF_RII, Osteopontin, PYY, Myoglobin, PYY.1, Myoglobin.1, MMP10, MCP_2, Fibrinogen, Fetuin_A, Eotaxin_3, ENA_78, tau and p_tau

Figures 1 and 2 show boxplots for precision measures (accuracy and kappa) for the models (2 repetitions of 10-fold Cross Validation):

ISBN: 978-88-5495-135-8

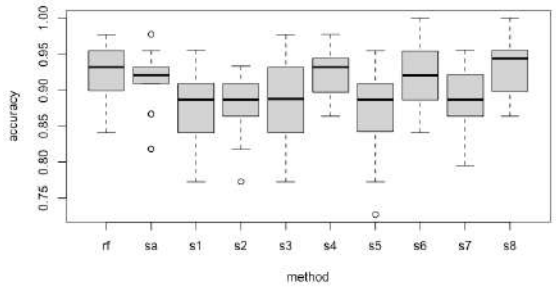


Figure 1 – accuracy for the several methods

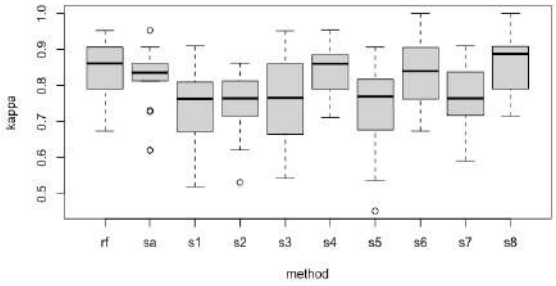


Figure 2 – kappa for the several methods

Although with some minor differences on the distributions for results, we can say that all 10 sets of predictors have similar precision. Figure 5 shows a plot for accuracy(black) and kappa(red). We see that best results were found for 1st and 2nd sets (*rf* and *sa*), 6th (*s4*), 8th (*s6*) and 10th (*s8*): around 0,93 for accuracy and 0,86 for kappa. Table 1 presents some more precision metrics:

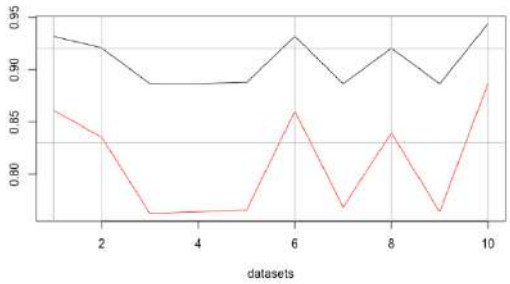


Figure 5 – Median accuracy and kappa for the several sets of variables

Table 1– Measures for the sets of variables

Variables	Accuracy	Kappa	Sensitivity	Specificity	Precision	Recall	F1	Obs
rf	***0,93***	***0,86***	***0,71***	***0,75***	***0,53***	***0,71***	***0,61***	***rf***
sa	***0,92***	***0,84***	***0,75***	***0,81***	***0,60***	***0,75***	***0,67***	***sa***
s1	0,89	0,76	0,68	0,75	0,51	0,68	0,58	s1
s2	0,89	0,76	0,68	0,75	0,51	0,68	0,58	s2=s1+Genotypes
s3	0,89	0,77	0,61	0,78	0,52	0,61	0,56	s2+age
s4	***0,93***	***0,86***	***0,75***	***0,78***	***0,57***	***0,75***	***0,65***	***s2+age***
s5	0,89	0,77	0,46	0,60	0,31	0,46	0,37	s5
s6	***0,92***	***0,84***	***0,71***	***0,79***	***0,57***	***0,71***	***0,63***	***s2+ICA***
s7	0,89	0,76	0,82	0,74	0,55	0,82	0,66	s7
s8	***0,94***	***0,89***	***0,79***	***0,71***	***0,51***	***0,79***	***0,62***	***s8***

3. Conclusion

We have seen that Supervised Machine Learning algorithms (Supervised Predictive Models - SPM) have a good enough precision in the early detection of Alzheimer Disease. But we should be aware that ML algorithms are not the final solution for diagnosis: they should always be based on clinical analysis. SPMs should be viewed as similar to blood tests. Besides in AI applications we should always be aware of the so called 'weapons of math destruction' (O'Neil, 2017): in these models, there are always false positives and false negatives, and, depending on the intensity of these false results, the results can mislead the analyst. We should never have a blind belief on the model results. Besides, clinical professionals know that a lot of other information, not given by the model, will help the diagnosis: patient's personal and family history, for example, among others.

References

Bedi, G., Slezak, D.F., Carrillo, F., Mota, N.B. (2015). Automated analysis of free speech predicts psychosis onset in high-risk, *Research Gate*.

Buch, V., Varughese, G., Maruthappu, M. (2018). Artificial intelligence in diabetes care, *Diabet. Med.*, **35**(495).

Contreras, I., Vehi, J. (2018). Artificial Intelligence for Diabetes Management and Decision Support: Literature Review, *JMIR Publications*, **20**(5)

Craig-Shapiro, R., Kuhn, M., Xiong, C., Pickering, E., Liu, J., Misko, T.P., Perrin, R.J. , Bales K.R., Soares, H., Fagan, A.M., Holtzman, D.M.(2011). Multiplexed Immunoassay Panel Identifies Novel CSF Biomarkers for Alzheimer's Disease Diagnosis and Prognosis, *PLoS ONE*, **6**(4).

Duchek, J. Balota, D. (2005). Failure to Control Prepotent Pathways in Early Stage Dementia of the Alzheimer's Type: Evidence from Dichotic Listening, *Neuropsychology*, **19**(5), pp.687–695.

Gagliano, M., Van Pham, J., Tang, B. Kashif, H., Ban, J. (2017). Applications of Machine Learning in Medical Diagnosis, *Research Gate*.

Hothorn, T., Lausen, B. (2003). Double-bagging: combining classifyers by bootstrap aggregation, *Pattern Recognition*, **36,** pp.1303 – 1309.

Kuhn, M., Johnson, K. (2013). *Applied Predictive Modelling*, Springer, New York.

Kuhn, M., Wing, J., Weston, J., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scruca, L., Tang, Y., Candan, C., Hunt, T., (2016). *caret: Classification and Regression Training*, R package version 6.0-71. https://CRAN-R-project.org/package=caret .

Lyle, A., Improved AI-based tool increases accuracy of schizophrenia diagnosis (2019). https://medicalxpress.com/pdf468148828.pdf .

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M. (1984). Clinical diagnosis of Alzheimer's disease, *Neurology*, **34**.

Miller, D. Brown, E.W. (2018). Artificial Intelligence in Medical Practice: The Question to the Answer, *The American Journal of Medicine*,**131**, pp. 129–133.

O'Neil, K. (2017)., *Weapons of Math Destruction*, Penguin Random House LLC, New York,

Staring and M., Smits, M. (2014). Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease, *Frontiers in Neuroinformatics*, **7**(50).

Spaan, P., Raaijmakers, J., Jonker, C. (2005). Early Assessment of Dementia: The Contribution of Different Memory Components, *Neuropsychology*, **19** (5), pp. 629 – 640.

Torgo L. (2017)., *Data Mining with R: Learning with Case Studies*, CRC Press, Boca Raton.

Vigneau, L, Chen, M., Cariou, V. (2019). ClustVarLV: Clustering of Variables Around Latent Variables. R package version 2.0.0, https://CRAN.R-project.org/package=ClustVarLV

Missing values in Social Media: an application on Twitter data

Paolo Mariani [a], Andrea Marletta [a], Nicholas Missineo [a]
[a] Department of Economics, Management and Statistics, University of Milano-Bicocca, Milano, Italy

1. Introduction

The 2018 annual report of Excelsior ed Anpal provides a lack of alignment regarding the 26% of the labour contracts predicted by the productive system. This is principally due to changes in the professional structure of the needs of the firms that requiring more specified profiles in a digital and sustainable field. It is difficult to find personnels both in the North of Italy where the labour market is more competitive and efficient and in the South of Italy where the unemployment rates are doubled with respect to the North. The difficulty of recruiting is more evident for young people: $1.267.000$ contracts have been predicted for under 30 years employees and the 28% is believed as not easy to find. This value reaches a peak of 62% for physics, chemicals and computer sciences specialists, 45% for the engineering sector and 43% for metalworkers (Unioncamere, 2018).

During the last three-month period employed increased in Italy: 60.000 workers more, the $0,4\%$of the total, 44.000 with a long-term contracts. But having a look on the long period statistics, the situation of the Italian labour market is still complicated. Italy did not recovered after the economic crisis and Italian workers are divided in two groups. The former is composed by those who were able to keep the safeguard of the past, this are on average old and close to the retirement and they are addressed to decrease. The latter is composed by younger and precarious workers and their number is increasing. Under occupation and involuntary part-time are more spread, while the wages are locked or in decrease.

A very important index for the labour market is the number of worked hours of the total amount of the workers in a year. In 2008 in Italy the worked hours were $45,8$ billions, in 2018 they reduced to $43,6$. This means that if the number of workers is similar to pre-crisis levels, the intensity of working is lower with respect to ten years ago. The comparison with the world situation is clear following the OECD data. The total unemployment rate decreased to $10,2\%$, but it is still twice the OECD average $5,2\%$, and over the average value of EU countries $(6,5\%)$. With respect to the young unemployment, the rate comprise 33%, three times to the rate of the average value of OECD and EU countries (OECD, 2019). In conclusion, the average income of the families decreased during last years and the increasing of the employed is only due to the increase of precarious workers.

One of the measured approved by the Italian government to improve the situation of the economy is the introduction of guaranteed minimum income for specified categories of citizens. The guaranteed minimum income (Reddito di cittadinanza in Italian language) is a form of sustain for families in difficult condition composed by two parts: an economic part, providing a minimum income and a contribution for the house location; a project for the insertion in the labour market of the person receiving the minimum income. This measure lasts 18 months and it could be renewed for other 18 months. The Italian governement approved this measure on 17th January 2019. It is a selective measure, only oriented to those who present a determined profile of difficulty. It is not universal and it requires a precise commitment.

In this work the aim is to understand the perception of this economic measure for Italian citizens and stakeholders before the introduction using social media data. In particular, the work

involves Twitter users about the percepetion of the Italian guaranteed minimum income on the basis of different categories of users. The main distinction about users is made between verified Twitter users and not verified users. The first category is related to politician, institutional authorities and other official stakeholders. The second one is represented by citizens and other subjects not directly involved in the process of realization of this measure. A classification method based on tweets, retweets and quotes posted by users with hashtag #redditodicittadinanza will be able to discern between verified and not verified users. Moreover, an analysis of the KPI (Key Performance Indicators) will be conducted using their presence and absence through the use of the complementary values. This tool is very useful to give a meaning to an absence of behaviour distinguishing between no interest and a negative opinion.

2. An explorative approach to treat social media data

Social networks are identified as an online informative system allow the realization of virtual social interactions. They are websites or technologies permitting to share textual contents, images, videos and interactions among users (Finger, 2013). Social media data are data collected from social network. Among social media, Twitter is one of the most spread and well-known. Differently from Facebook or Instagram, Twitter has been used to share news, official contents about economics and political issues. This is why for this study, Twitter data have been provided and statistical units are represented by Twitter record.

In particular, here each record represent a tweet, retweet or quote made by a Twitter user. A tweet is a written post on Twitter with a maximum of 280 characters. A retweet is a reproduction of a written post by another user, with, eventually, a comment of maximum 280 characters. A quote is a comment of maximum 280 characters on a tweet or a retweet.

To evaluate the effective communicative capability of a Twitter record, a comparison has been implemented among the above-listed KPIs and the respective complementary frequencies. The effective difference between the distribution of KPIs and the respective complementary frequencies has been carried out using a chi-squared test (Pearson, 1900) and a factor analysis (Cattell, 2012). A difference between these two distribution could lead to a presence of bots or a dislike effect (Mariani et al., 2019).

Here two indicators are taken into account: the number of likes and the number of a retweet obtained by each record. Social networks are communication media based on the interaction between users, in which the participants tend to express in a clear way their subjectivity. This behaviour is simplified by the possibility of forwarding to the followers posts written by other users or tagging them with a like. The difference between a like and a retweet is that while a like expresses an appreciation for a post, the retweet does not imply an approval but it equals to show the attention for that post to the followers.

Starting from the fact that a tweet is more effective if it receives more visualizations, like and retweets make bigger the attention on it. The hypothesis is that the behaviour of the complementary observations was similar to those of real observations.

The procedure to apply the proposed approach is the following: creating 3 tables (tweet, retweet and quote) with number of likes and retweets; creating 3 complementary tables; computing the chi-squared on the two distributions (real and complementary); plotting the first two principal components of the factor analysis underlining the positioning of variables and observations.

3. Results

Data are collected in April 2019 using the official API Twitter for the entire Italian territory. Using this scraping method, 4797 records and covariates have been obtained. These records have been divided into three categories aforementioned, 945 tweets, 3774 retweets and 78 quotes. The covariates available for each records are 63 and they could classified in three class: variables of the element, variables of the user, variables of the follower. Some examples of covariates in the first class are the entire text of the element, date of publication, type of element. In the second class, the characteristics of the user who published the element. Finally in the third class, the characteristics of the user who interact the element.

Results of chi-squared test to verify the equivalence of the real and complentary distribution for tweet, retweet and quote are presented in Table 1. It is possible to note that in all cases the H_0 hypothesis of equal distribution is rejected.

Table 1: Chi-squared test values for testing hypothesis of equal distribution

Element	$\chi^2_{oss,\alpha}$	$\chi^2_{(p-1)(n-1),\alpha}$	Result
Tweet	14869.01	594.12	Rejected H_0
Retweet	304134.4	6258.45	Rejected H_0
Quote	532.84	117.43	Rejected H_0

Once the hypothesis of equal distribution between real and complementary frequencies for all the considered elements, it is possible to extract latent dimensions using a factor analysis. In particular after the choice of first two components for all elements (explaining 80% of the variance), here in Figure 1, the correlation circle has been shown only for tweet element. For space reasons, here the same plots for retweet and quote have been omitted, but results are similar.

As it is possible to note from the graph, variables for real frequencies and complementary observations are not diametrically opposed. This supports the initial hypothesis of a behaviour between users, probably due to a potential presence of a dislike effect or bots among observations.

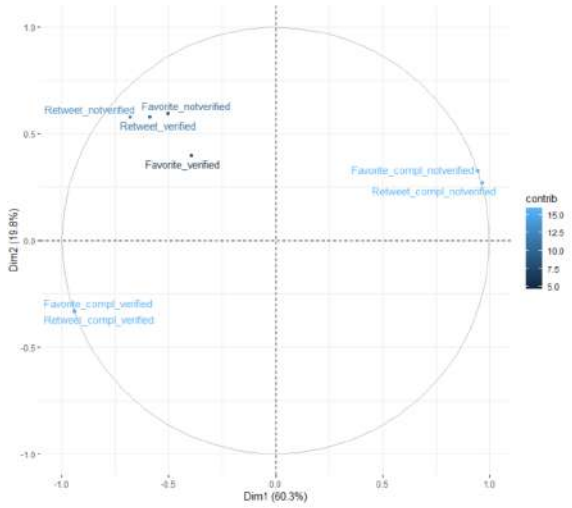


Figure 1: Correlation Circle from Factor Analysis on element Tweet

4. Conclusions

The analysis of social media data could be very useful when the aim is to detect the effect of an innovation on citizens. For this reason, the aim of this study is to verify how citizens responded to the introduction of a new economic measure to face the crisis of Italian labour market. The introduction of guaranteed minimum income has been very discussed from a political point of view, but now since it was approved by Italian government, an exploratory analysis has been carried out after the approval and before the application. Data source are the Twitter elements containing the hashtag "#reddito di cittadinanza" and among the covariates the attention has been focused on the number of likes and retweets for those posts.

From a methodological point of view, the use of the frequency distributions of the complementary elements allowed to verfify the presence of a behaviour in the lack of expression for the observations. Chi-squared tests and factor analysis reinforced the hypothesis of a different distribution in tweet, retweet and quote for real and complementary elements both for verified and not verified users. This could be due to presence of a dislike effect or to presence of bots, fake users just created to spread (in positive or negative way) Twitter elements about this topic. Future research could be addressed to a procedure based on predictive models for the identification of these fake profiles signalling these abuse to Twitter.

References

Cattell, R. (Ed.). (2012). The scientific use of factor analysis in behavioral and life sciences. Springer Science & Business Media.

Finger, L (2013), Ask, Measure, Learn: Using Social Media Analysis to Understand and Influence Costumer Behaviour.

Mariani, P; Marletta, A; Grammatica, E (2019). A missing value approach on Facebook Big Data: Like, Dislike or Nothing? *In Data Science & Social Research 2019 - Book of Abstracts*.

Organisation for Economic Co-operation and Development (OECD) Staff. Rapporto Economico OCSE Nota di Sintesi.(2019).

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302), 157-175.

Unioncamere. Sistema informativo excelsior. In Progetto Excelsior. Sintesi dei principali risultati. URL http://excelsior.unioncamere.net (2018).

Application of multivariate statistics in sports: Exploration of recall and recognition of UEFA Champions League sponsors

Milica Maricic[a]
[a] Department of Operational Research and Statistics, University of Belgrade, Belgrade, Serbia

1. Introduction

Sponsorship can be, according to Meenaghan (1983), defined as "the provision of assistance either financial or in-kind to an activity by a commercial organisation for the purpose of achieving commercial objectives". Accordingly, sponsorship has become an important mean of communication in the company's marketing strategy. The opportunity of displaying the company's logo or advertisement to massive audiences around the globe, sponsorship comes with an extreme price tag (Kwak and Pradhan 2019). The most common type of sponsorship is sports sponsorship which envelops sponsoring of teams, athletes, competitions, and leagues. Having the above mentioned in mind, the interest of marketing experts and academics in decision-making and effectiveness measurement of sponsorship activities increased (Olson and Mathias Thjømøe 2009, Lee and Ross 2012, Walraven et al. 2016).

The specific mean of sponsorship and advertising, which is attracting the attention of sponsors is embedded advertising. Cain (2011) provides a definition of embedded advertising as "advertising embedded in a TV show, print magazine or sports match, whereas it is competing for attention with other stimuli". There are several benefits of this kind of approach to sending sponsorship messages: there is no persuasive tone of the message, the message is easier to repeat, and there is place for repeated exposure (Pitts and Slattery 2004, Cowley and Barron 2008, Schmidt and Eisend 2015).

Studies on the application of multivariate statistics in the field of sport are mostly related to the application on data gathered from the training and matches. However, this paper attempts to apply statistical analysis in the field of sport sponsorship and sponsorship decision making. Namely, marketing managers are given not an easy task to choose which events to sponsor, how to sponsor them, and how to increase the effectiveness of such activities. This paper addresses two issues the managers are faced with: whether to use embedded advertisements (Nebenzhal and Jaffe 1998) and how to measure the effectiveness of embedded sponsorship (Walraven et al. 2016).

To achieve the aim of the paper, a survey was designed and conducted in Serbia on the attitudes towards sponsors of the specific sports competition – UEFA Champions League (UEFA CL). Afterwards, a conceptual model was created and tested on the gathered data. Specific attention was put on the difference of impact taking account the respondents' previous involvement with the UEFA CL. The statistical analysis chosen to measure the impacts among constructs is structural equation modelling (SEM). There is hope this paper will initiate further research on the topic of sport sponsorship, sponsorship efectiveness and embedded advertisement.

2. Research model

The research outlined here presents a continuation of the work by Maricic et al. (2019). Namely, in their paper, the authors created a novel conceptual model for measuring the mutual impact of *Involvement* (measured though involvement into football as a sport), *Exposure* (measured as the exposure to a certain sport competition), *Awareness* (defined as the recall and recognition of sponsors), *Attitude* (defined as the attitude towards the sponsors), and *Purchase intention & experience* (observed as the purchase intention of sponsors'

ISBN: 978-88-5495-135-8

product/service). Their observed hypotheses were (Maricic et al. 2019):

H1: Consumer involvement in a particular sport has a direct positive effect on the consumer's exposure to a particular sport competition.

H2: Consumer involvement in a particular sport has a direct positive effect on sponsorship awareness.

H3: Consumer involvement in a particular sport has a direct positive effect on the consumers' attitude towards sponsorship and the sponsor.

H4: Consumer exposure to a particular sports competition has a direct positive effect on sponsorship awareness.

H5: Consumer exposure to a particular sports competition has a direct positive effect on the consumers' attitude towards sponsorship and the sponsor.

H6: Consumer awareness has a direct positive effect on the consumers' attitude towards sponsorship and the sponsor

H7: Consumer awareness has a direct positive effect on purchase intentions and experience.

H8: Sponsorship attitude has a direct positive effect on purchase intentions and experience.

To accept or reject the listed hypothesis, they conducted an online survey in 2017 among the viewers of the UEFA CL in Serbia. Their conclusions were that the hypothesis H1, H2, H4, H7 and H8 could be accepted. Nevertheless, they conducted the SEM analysis on all respondents. The question raised here is whether there is a difference in the SEM models and rejection of hypothesis if more experienced viewers, "old", and first-season viewers, "new", are observed separately?

3. Results

The data set on which the study was conducted is the same as in Maricic et al. (2019). They have received 444 responses on the complex seven-section survey. They have mostly covered young, male students from Belgrade, Serbia, who regularly watch UEFA CL. For more details on the questionnaire and descriptive statistics, please consult the original paper.

Herein, two SEM models were created: one on 351 respondents who stated they watched the previous season of the UEFA CL, and second, on 93 respondents who stated the opposite. The aim is to explore are the mutual impacts of constructs the same or not if previous exposure to UEFA CL is taken into account.

The initial model for multi-year viewers had relatively solid fit to the data (Chi-square=1172.784, df=295, $p<0.000$, RMSEA=0.092, CFI=0.825, TLI=0.807). To additionally modify the model, modification indices were used. The final model had good fit to the data (Chi-square=757.444, df=279, $p<0.000$, RMSEA=0.070, CFI=0.904, TLI=0.889). The obtained coefficients are presented in Figure 1. As presented, four out of eight hypotheses have been accepted. Involvement proved to have an impact on exposure and awareness, as expected. It should be pointed out that awareness had an impact on purchase intention. Similarly, attitude plays a role when it comes to purchase intention. The best-explained construct is *Exposure*, with 61.2% variance explained, followed by *Purchase intention & experience*, with 20.7% variance explained.

ISBN: 978-88-5495-135-8

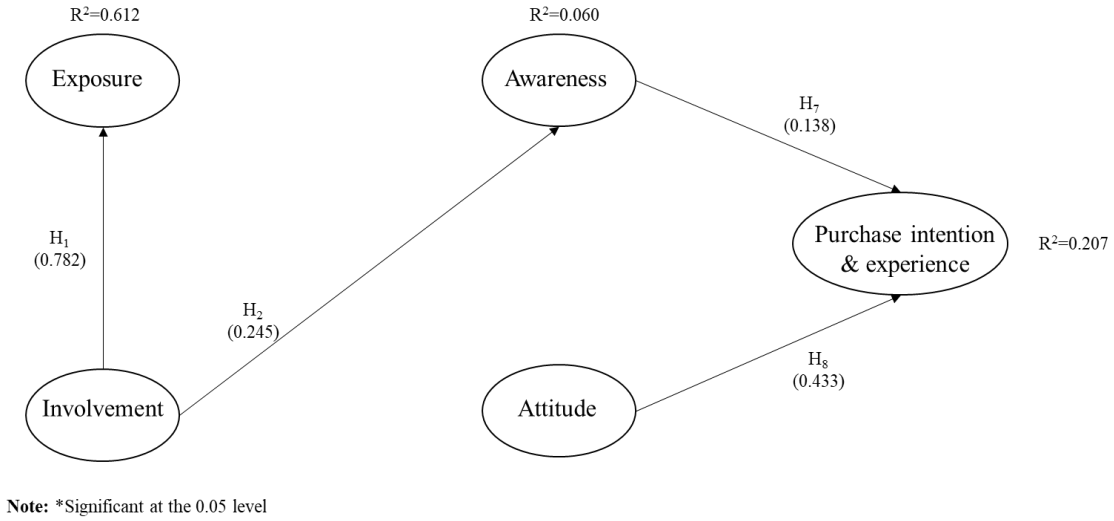


Figure 1: Structural model for multi-year viewers

The initial model for first-year viewers had relatively solid fit to the data (Chi-square=668.081, df=295, p<0.000, RMSEA=0.117, CFI=0.855, TLI=0.840). To additionally modify the model, modification indices were used. The final model had good fit to the data (Chi-square=463.464, df=280, p<0.000, RMSEA=0.084, CFI=0.929, TLI=0.917). The obtained coefficients are presented in Figure 2. Again, four out of eight hypotheses have been accepted. Involvement proved to have an impact on exposure and awareness, while interestingly, exposure has an impact on attitude. The obtained result differs from the result of Maricic et al. (2019). Attitude towards sponsors, as expected, had an impact on purchase intention and experience. Again, the best-explained construct is *Exposure*, with 59.2% variance explained, followed by *Purchase intention & experience*, with 29.3% variance explained.

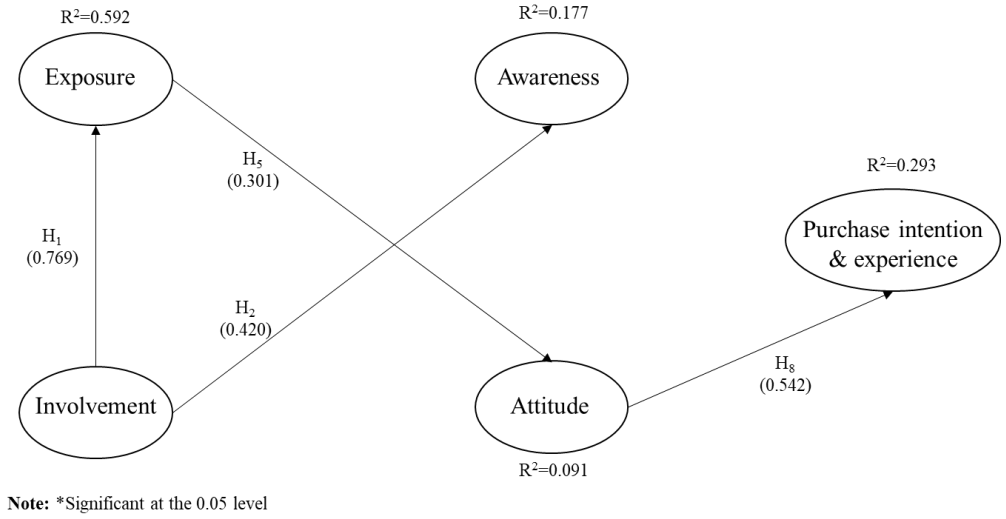


Figure 2: Structural model for first-year viewers

The two models should be more closely compared. When it comes to H1 and H2, they were bot accepted in the two models. This indicates that no matter how the respondent often watches the UEFA CL, his or her involvement in football has a role in the determination of the exposure to the league and awareness of the sponsors. Hypothesis H3 and H4 have been rejected in both models, indicating that consumer involvement in a particular sport does not

have an effect on the consumers' attitude no matter whether he has been watching the league previously. Also, pure exposure, for multiple seasons or just one, does not have an impact on recall and recognition of sponsors. The hypothesis H5 is of interest as in the model of first-year viewers; exposure had an impact on attitude towards sponsors. It can be presumed that the first-year viewers were amazed with the on-screen organisation, stadiums, field surroundings and transferred the content to the sponsors. H6 was rejected in both models meaning that the awareness of the sponsors does not have any effect on attitude towards sponsors. However, H7 was accepted in the model for older viewers, which might indicate that multi-year exposure and awareness leads to increased purchase intention. Finally, H8 was accepted in both cases.

4. Conclusion

The proposed conceptual model tries to explore the relationship between several constructs which are believed to have an effect on sponsorship awareness, taking into account the previous exposure to the sponsored event. The results might have several impacts. First, they show that the attitudes of first-year viewers and multiple-year viewers differ and that different constructs have a mutual impact. Second, that pure exposure to the sponsored event has an impact on attitude towards sponsors for first-time viewers. And third, that affection to the particular sport plays a detrimental impact on the exposure to a particular event or league, no matter whether it is the first or later exposure. Future directions of the study could include other constructs such as team loyalty or fan identification. The study could also be recreated in a country which has more football clubs participating in the UEFA CL.

References

Cain, R. M., (2011). Embedded Advertising on Television: Disclosure, Deception, and Free Speech Rights. *Journal of Public Policy & Marketing*, **30**(2), pp. 226–238.

Cowley, E. and Barron, C., (2008). When Product Placement Goes Wrong: The Effects of Program Liking and Placement Prominence. *Journal of Advertising*, **37**(1), pp. 89–98.

Kwak, D. H. and Pradhan, S., (2019). Fans' responses to the National Basketball Association's (NBA) pilot jersey sponsorship program: An experimental approach. *Journal of Sports Analytics*, **5**(2), pp. 121–136.

Lee, S. and Ross, S. D., (2012). Sport sponsorship decision making in a global market. *Sport, Business and Management: An International Journal*, **2**(2), pp. 156–168.

Maricic, M., Kostic-Stankovic, M., Bulajic, M., and Jeremic, V., (2019). See it and believe it? Conceptual model for exploring the recall and recognition of embedded advertisements of sponsors. *International Journal of Sports Marketing and Sponsorship*, **20**(2), pp. 333–352.

Meenaghan, J. A., (1983). Commercial Sponsorship. *European Journal of Marketing*, **17**(7), pp. 5–73.

Nebenzhal, I. D. and Jaffe, E. D., (1998). Ethical dimensioins of advertising executions. *Journal of Business Ethics*, **17**(7), pp. 805–815.

Olson, E. L. and Mathias Thjømøe, H., (2009). Sponsorship effect metric: assessing the financial value of sponsoring by comparisons to television advertising. *Journal of the Academy of Marketing Science*, **37**(4), pp. 504–515.

Pitts, B. and Slattery, J., (2004). An Examination of the Effects of Time on Sponsorship Awareness Levels. *Sport Marketing Quarterly*, **13**(1), pp. 43–54.

Schmidt, S. and Eisend, M., (2015). Advertising Repetition: A Meta-Analysis on Effective Frequency in Advertising. *Journal of Advertising*, **44**(4), pp. 415–428.

Walraven, M., Koning, R. H., Bijmolt, T. H. A., and Los, B., (2016). Benchmarking Sports Sponsorship Performance: Efficiency Assessment With Data Envelopment Analysis. *Journal of Sport Management*, **30**(4), pp. 411–426.

ISBN: 978-88-5495-135-8

Short-run and long-run persistence of bad health among elderly

Daria Mendola[a], Paolo Li Donni[a]
[a] Department of Economics, Business and Statistics (SEAS), University of Palermo, Palermo, Italy.

1. Introduction

The assessment of elderly health and quality of life is receiving increasing importance in both social and economic health policy planning for which longevity is not the only primary goal, while well-being is nowadays assuming a central role (Angelini et al., 2012). In 2016 21.8 per cent of non-institutionalized persons aged 65 and over are in fair or poor health in US, and 6.4% need help with personal care from other persons (National Center for Health Statistics, 2017). Persistence in bad health (not limited to chronic disease) is one of the main assailant to the elderly chance to a successful aging. A prolonged period of bad health as well as repeated and close episodes of bad health significantly affect the quality of life of people, threatening elderly ability to fulfil occupational, social and family roles.

Relationships between elderly health and their socioeconomic and demographic characteristics have been clearly and almost consistently assessed in several studies, both in cross-sectional and in longitudinal settings (e.g. Hernàndez-Quevedo et al., 2008; Contoyannis et al., 2004). Among others, Buckley et al. (2004) study how wealthier and better educated older Canadians achieve better health outcomes; and Cheung's (2000) empirical findings suggest that lack of social support is negatively correlated with health persistence while marital status has an unclear effect. Moreover many studies found that health-risk behaviours such as smoking status, alcohol consumption, having a sedentary lifestyle, and obesity are associated with a variety of health-risks and poorer health status (Lantz et al., 2001).

The importance of studying more in depth the dynamics of health to set policy intervention has been widely supported by many empirical studies using panel data models and measuring the persistence effect in terms of state dependence from the previous level of health stock. In particular Contoyannis et al. (2004) identify a state dependence effect, related to the fact that some illnesses are inherently chronic and long-lasting, and a "purely" heterogeneous dynamic" effect. Therefore this state dependence simply translate into a short-run effect, which can be related to several contingent factors such as education, material deprivation, lifestyles, socio-economic status and environment that may also have a long-lasting influence on individuals' health. This long-run effect may not necessarily be the same of the one-period lagged effect captured by state dependence, and it could be the results of a cumulative (non linearly additive) repeated sequence of bad health events. Thus there are many factors affecting health and quality of life whose effect cannot be disentangled by using cross-sectional data and that are relevant to set policy interventions.

The present paper aims at proposing two different approaches to measure health dynamics among the elders by distinguishing between long- and short-run effects. Interesting similarities were found between the two approaches.

2. Data and methods

The Health and Retirement Study (HRS) is a biennial survey targeting elderly Americans over the age of 50, sponsored by the National Institute on Aging of US. Since 1994 it has been providing longitudinal data for a rich array of information, consistently administrated, on several different fields such as health and health care utilization, lifestyles, socioeconomic

ISBN: 978-88-5495-135-8

conditions. Particularly we drawn a fully balanced panel of 3,365 individuals reporting consistent information for health over ten waves from 1996 to 2014. We focused on the self-assessed health measure, largely recognized as a trustworthy indicator of health status (Dardanoni and Li Donni, 2012).[1]

In this paper we propose two possible strategies to model subjective health trajectories over time. The first strategy relies on a dynamic econometric approach to model contribution of state dependence (the short-run effect) and unobserved heterogeneity on individual health. The second strategy, which has not yet been used in the health context, relies on the "spells approach" exploited to synthesize sequences of ill/good health episodes and then it models the long-run effect. The first approach assumes the reported health as generated by a short-memory latent process, modelled as a latent Markovian (LM) process. Marginal distribution of individual health can be modelled as follows:

$$\log \frac{p(h_{it}=1|U,\boldsymbol{x}_{it-1},\boldsymbol{z}_i,h_{i,t-1})}{p(h_{it}=0|U,\boldsymbol{x}_{it-1},\boldsymbol{z}_i,h_{i,t-1})} = \alpha_{it}(U) + \boldsymbol{x}'_{it-1}\boldsymbol{\beta} + \boldsymbol{z}'_i\boldsymbol{\gamma} + h_{i,t-1}\delta \qquad (1)$$

where h_{it} is the health status of individual i at time t (bad=1 or good=0); $\boldsymbol{x}_{it}$ is the vector of individual explanatory variables at time *t-1* and $\boldsymbol{z}_i$ are time invariant personal characteristics. A crucial issue also in this setting is the initial condition problem. To explicitly take into account this problem we adopt a multinomial logit parameterization of the initial probabilities and allows π_{iu} (the probability of the *i-th* individual to belong to the latent state u) to depend on the initial observation of h, $\boldsymbol{x}$ and $\boldsymbol{z}$ (Vermunt et al., 1999). This empirical strategy deals with unobserved time-varying individual heterogeneity affecting health and it accounts for state dependence among consecutive health statuses over time. The LM model takes into account the structure of the underlying individual heterogeneity, by identifying unobserved groups with different propensity over time to report bad health. We set this model by using some recent developments on marginal modelling and multivariate LM model (Bartolucci and Farcomeni, 2009). In particular we estimate a LM model such that unobservable heterogeneity is captured by including a vector of subject-specific parameters, which are time-varying and follow a first order homogeneous Markov chain. Explanatory variables are demographic (age, gender, marital status, ethnic group), socioeconomic (education, income, and occupational status), and health and lifestyle characteristics (being a smoker, sedentary, obese; having a chronic disease).

On the other side, the proposed spell approach requires very few assumptions and is able to exploit all the longitudinal information available on health at individual level (going beyond lag 1) in a very flexible way. We adapted an index drawn from the literature on the measurement of persistence in poverty (Mendola et al., 2011), in order to propose a fully not parametric measure of persistence in bad health based on the observed sequences of self-assessed health statuses. The hypotheses at the basis of this proposed index (that here we rename as Health Persistence Index, HPI) are a) *Cumulative damages*: close and repeated bad health spells have a more detrimental impact on health related quality of life of individuals than occasionally and distant experienced bad health spells; b) *Time monotonicity*: other things being equal, persistence increases with the increase of the total number of waves spent in bad health; c) *Path-dependence*: other things being equal, persistence increases if the number of consecutive years in bad health rises; and d) *Volatility*: given the total number of years spent in bad health, persistence decreases with the increase of the number of transitions

[1] Noteworthy, in this paper both approaches are based on a dichotomous assessment of the health status (e.g. good/bad health) over time hence we proceeded in splitting self-assessed health in excellent+very good+good vs fair+poor health.

ISBN: 978-88-5495-135-8

out of bad health status (volatility) over time, and, in contrast, persistence is higher if there is a low number of transitions in and out of bad health status (stability).

HPI=0 if the individual enjoys a good health along the whole observation period, while HPI=1 when individual suffers a bad health along all the waves, with no status's change. All values in (0,1) can be observed on any individual and express different degrees of persistence in bad health. Moreover, in computing the HPI, we decided not to account for the intensity of health status given the lack, in our dataset, of a measure sufficiently differentiating individuals. For details on the analytical formulation of the index please refer to Mendola et al. (2011) and Mendola and Busetta (2012).

3. Results and discussion

Estimates from the LM model in (1) revealed the existence of three unobserved groups differing in their propensity to report bad health. The first group (latent state) is characterized by a lower probability of reporting bad health over time as compared to the other states; the second one has a lower probability to report poorer health over time as compared to the third; and the third one refers to individuals frequently in bad health, hence with the highest propensity of being sick. Presumably, individuals in the first group are those with unobserved factors (e.g. genetic endowment, high health-related risk aversion, etc.), which persistently positively affect health over time. Similarly individuals in the third group there are those with unobserved factors (e.g. genetic endowment, unhealthy habits, etc.), which persistently negatively affect health over time.

State dependence was confirmed: health stock in the previous period affects negatively current health (b=0.373***), that is being in bad health at time *t-1* increases the persistence of bad health in *t*.

Regarding the impact of the personal characteristics on the persistence in bad health, we observed a decreasing relationship with education level indicating that better educated elders are able to live a healthier life (b=-0.113***). Persistence in bad health decreases with income (b=-0.095**), with physical activity (b= -0.272***), and when being never married (vs married, b= -0.781***); while it increases with depression intensity (measured via the CESD scores, b= 0.115***), number of declared diseases (b=0.386***), and being part of an ethnic minority (vs white people, b=0.454** for Black people, and b=0.672*** for other ethnic groups). No significant effect are detected for occupation, body mass index, separated people, current smokers, other things being equal.

We then proceeded by computing the HPI. Our results showed that the great number of elders in the sample is never in bad health along ten waves. The higher values of the index are consistently observed with the main socio-demo-economic risk factors, indicating that individuals with lower education, unhealthier lifestyles and limited participation in the labour market experience higher persistence in bad health.[2]

Interestingly the factors affecting bad health persistence do not differ from a short-run to a long-run perspective. This suggests that whether a factor is associated with bad health in the short-run, it is also likely to be persistently associated with subsequent health statuses in the long-run period.

Furthermore, both approaches provide a measure of the degree of persistence. While this measure is directly provided by the HPI values, it can be recovered in the LM model by computing the posterior probabilities of experiencing persistently bad health over time. In our case this can be obtained by averaging over time the individual probability of being in the third latent state in every time occasion. To make this comparison more extensive, and with

[2] Although results from the HPI and the LM model provide a very similar picture, as explained in the following, their structure is substantially different, which renders their analytical comparison unfeasible.

ISBN: 978-88-5495-135-8

the aim of providing a simple sensitivity analysis, we also estimated the LM model by excluding $\boldsymbol{x}_{it-1}$ and $\boldsymbol{z}_i$ from equation (1), and computed again the above mentioned posterior probabilities of persistence in bad health. It stands out that the two strategies are highly correlated indicating that they are identifying the same degree of persistence at individual level and that both approaches (and not only HPI) are substantially able to catch a longer view over health dynamics, going beyond lag 1. Thus individuals with higher persistence in bad health, namely higher HPI values, are also those with higher propensity of experiencing bad health over time. It is noteworthy that including time-varying covariates (i.e. $\boldsymbol{x}_{it-1}$) lowers the correlations between HPI and estimated posterior probabilities indicating that HPI takes into account the persistence in poor health (implicitly) inclusive of the explanatory time-varying factors.

References

Angelini, V., Cavapozzi, D., Corazzini, L., Paccagnella, O. (2012). Age, health and life satisfaction among older Europeans. *Social indicators research*, **105**(2), pp. 293-308.

Bartolucci, F., Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, **104**(486), pp. 816-831.

Buckley, N.J., Denton, F.T., Leslie Robb, A., Spencer, B.G. (2004). The transition from good to poor health: an econometric study of the older population. *Journal of Health Economics*, **23**(5), pp. 1013-1034.

Cheung, Y.B. (2000). Marital status and mortality in British women: a longitudinal study. *International Journal of Epidemiology*, **29**(1), pp. 93-99.

Contoyannis, P., Jones, A.M., Rice, N. (2004). The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics*, **19**(4), pp. 473-503.

Dardanoni, V., Li Donni, P. (2012). Reporting heterogeneity in health: an extended latent class approach. *Applied Economics Letters*, **19**(12), pp. 1129-1133.

Hernàndez-Quevedo, C., Jones, A.M., Rice, N. (2008). Persistence in health limitations: A European comparative analysis. *Journal of Health Economics*, **27**(6), pp.1472-1488.

Lantz, P.M., Lynch, J.W., House, J.S., Lepkowski, J.M., Mero, R.P., Musick, M.A., Williams, D.R. (2001). Socioeconomic disparities in health change in a longitudinal study of US adults: the role of health-risk behaviors. *Social Science & Medicine*, **53**(1), pp. 29-40.

Mendola, D., Busetta, A. (2012). The importance of consecutive spells of poverty: A path-dependent index of longitudinal poverty. *The Review of Income and Wealth*, **58**(2), pp. 355-374.

Mendola, D., Busetta, A., Milito, A.M. (2011). Combining the intensity and sequencing of the poverty experience: a class of longitudinal poverty indices. *Journal of Royal Statistical Society- Series A*, **174**(4), pp. 953-973.

National Center for Health Statistics (2017). *Health, United States, 2016: With Chartbook on Long-term Trends in Health*. Hyattsville, MD.

Vermunt, J.K., Langeheine, R., Bockenholt, U. (1999). Discrete-time discrete state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, **24**(2), pp. 179-207.

ISBN: 978-88-5495-135-8

Harmonised Administrative Databases: a new approach in the era of Big Data

Vittorio Nicolardi [a], Caterina Marini [a]
[a] Department of Economics and Finance, University of Bari Aldo Moro, Bari, Italy

1. Introduction

The real challenge that in the nowadays society needs to be scientifically faced is to accurately handle the enormous flow of information that in an IT world can be tremendously powerful to analyse the social and economic changes. The official statistics that the National Institutes of Statistics yield are facing their limits in all the research areas and the problem is already recognised and discussed in the international scientific context. Undoubtedly, the Public Administration (PA, hereafter) datasets can be a very useful source of additional and detailed data to complete the statistical information about phenomena that are still partially depicted by means of the official data. And, in some cases, limitations in dealing with the real size of many socio-economic territorial developments are significant. This is the case, for instance, of phenomena such as health, labour forces, well-being, real estate. In this context, however, it is important to underline that the administrative databases present two main problems that need to be considered, both difficult to deal with. The first is related to the primary purpose by which the administrative databases are fulfilled, which is obviously not statistical. The second is related to the huge amount of data that, in the era of Big Data, is stored in the administrative databases, which experience all the consequences related to this aspect. An additional problem, which is sometime encountered, is related to the entirety of the administrative data on the phenomenon analysed when its complete information requires the merging of two or more databases that often belong to two or more PA offices. In this paper, we focus the analysis on the Italian real estate phenomenon and how the administrative data are powerful in adding new information on the phenomenon in terms of both volume and value. In particular, the analysis is circumscribed to the territory of the city of Bari, in the South Italy, because it is part of a national research project but the outcomes we generated can be perfectly replicated in any dimensional territorial area, both metropolitan and regional and national. We built a unique administrative database starting from 4 independent administrative databases, normally managed by independent PA offices (the Italian Real Estate Registry and the Italian Revenue Agency), that provide autonomous information. The record linkage has required the basic practise of the big data methods to deal with both missing data, duplication and erroneous information, and the identification of the useful variables to merge the 4 data sources. Although we have restricted the analysis to one city, the amount of data has also required the application of GIS processes to guarantee the exact matching of data and depict the real estate framework in detail. In fact, the results of our work allow researchers and policy makers to deeply analyse the territory, even the single real estate unit. And the differentials between the real estate market monetary values reported by the Italian Revenue Agency (IRA, hereafter) and the real estate values reported by the Italian Real Estate Registry have shown significant results in terms of potential revaluations of city neighbourhoods or areas. Therefore, the potential effects on the local economies in terms of tax revenues can be highly innovative. The importance of the analysis we yielded in this paper is unique and original in its attempt to describe an economic phenomenon that, not only in Italy, still suffers the consequences of the dearth of a complete and harmonised data warehouse. Our work is the first challenge in this sense.

2. The data, the method and the outcomes

The main problems that are normally encountered when it is necessary to work with administrative databases are the typology of data and the corresponding quality that they contain. The typology is fundamental to plan the type of the analysis. The quality is likewise important to guarantee the reliability of the outcomes. Both issues require a great attention when the size of the databases is remarkable because the opportunities relying on the major availability of information risk to be a weakness for the purposes of the studies. And in this sense, one of the most delicate phases in the study of huge databases regards the cleaning and the management of the same. Therefore, in our work, we decided to independently work on each database to pre-process data and select the key features of each database to finally proceed with the merging action. Three of the 4 datasets belong to the Real Estate Registry (RER, hereafter) of the city of Bari and they contain all information related to the real estates. Although the PA office is the same, the 3 databases are independent and autonomous in providing the corresponding information. Therefore, the Italian RER has the complete information on real estates though utilises that in a roundabout way that complicates its same use. The main database is that named Real Estate Units (REU, hereafter) and it is a list of records in which all the technical and economic cadastral information of each unit is recorded. Information we use in this study is referred to the real estate cadastral category to identify the various typologies of units such as, for instance, dwelling or shop or office, the cadastral income and the size of each unit. The size of the database is 283,217 records, without duplication, referring to each unit but 20,240 records lack cadastral income because they belong to units of a particular cadastral category without income (i.e. the F category). Therefore, without statistically affecting the analysis, a first cleaning has been necessary to delete the uncomplete records to guarantee a homogeneous dataset in terms of information. The other 2 RER databases are functional to build the final database. The first is named Cadastral Identifiers (CI, hereafter) and includes all the cadastral information, mainly the Urban Section and the Cadastral Sheet, Subordinate and Parcel, to merge the REU dataset and the Census Section dataset by the Italian Institute of National Statistics (ISTAT, hereafter). The latter is important to support the geo-localization in the GIS process. The CI dataset includes 421,324 records, a number much higher than REU records because of both duplication, caused by some administrative change, and the presence of some real estate unit whose record has not been deleted though the building was demolished and is not really anymore existing. Therefore, a cleaning action has been necessary to homogenise information between the CI and REU datasets, and the final size of the CI database coincides with that of the REU database. The last RER database is named Cadastral Addresses (CA, hereafter) and comprises the toponyms of the real estate units. Toponyms are important to identify the exact localisation of each unit on the urban territory. The size of the CA dataset is 668,302 records and, likewise the previous databases, a cleaning action has been necessary because of the same reasons of the CI dataset previously described although the CA duplication is caused by the modifications of some toponym and/or building number. Therefore, the final size of the CA database equalises the other 2 databases. The cleaning of the CI and CA databases has been yielded by means of, respectively, the *Protocol Number* field and the *Sequential* field that report the several modifications that involved the real estates over time. Once the numeric homogeneity of the database size is obtained, the successive step is to merge them. We have identified as merging field the Cadastral Office Real Estate Identification Code (COREIC, hereafter) because that is the unique field in common between the 3 datasets. We obtained, therefore, the Unique Real Estate Registry (URER, hereafter) database by means of COREIC. The fourth database we used belongs to the Real Estate Italian Observatory (REIO, hereafter) of IRA. In Italy, this source of data is the main and one of the most reliable to analyse the real estate monetary value dynamics. The REIO real estate value data are calculated based on the trade price per square

meter of the properties. They are open data on biannual basis referred to the minimum and maximum price for all the different types of real estates at the level of the council territory. In order to use a univocal REIO value in our analysis, we calculated the midrange value for each record. The council territory is split in homogeneous areas that experience the same economic and socio-environmental characteristics, i.e. the REIO areas. All the areas of the same city are then grouped in 5 territorial districts that delineate precise geographical portions of the urban space: Centre, Near-centre, Outskirts, Suburbs and Extra-Urban. In this work, we use the REIO dataset of the city of Bari for the years 2015 and 2018. Biannual data are referred to the 14 typologies of the real estates that exist in Bari, for a total of 7,814 records. In the case of REIO database, finally, the cleaning action is not necessary because data are already statistical values. Table 1 shows the size of all the databases we used in this work in terms of fields and records, original size and final size after cleaning when occurred. To attain the purpose of our work, that

Table 1: Dataset contents.

Dataset	Original Size		Final Size	
	Fields	Records	Fields	Records
Real Estate Units	29	283,217	9	262,977
Cadastral Identifiers	13	421,324	6	262,977
Cadastral Addresses	5	668,302	5	262,977
Real Estate Italian Observatory	24	7,814	6	7,814
Istat Census Sections	4	82,576		

is comparing the real estate cadastral income with the corresponding market value, it is necessary that the URER and REIO databases are aligned. In fact, it is important to highlight that the 2 databases previously described are unlinked and not directly connectable through any field although they are referred to same object, and in literature there is not any attempt in this sense. To align the 2 databases, we need to solve two technical problems that involve the procedure. The first and most important issue is related to the specific territorial context that is differently defined in each database. In particular, the territorial context is the single Cadastral Parcel in URER, while in REIO the geo-context is the REIO area. Therefore, to surmount the obstacle we yielded a GIS procedure by means of two additional databases: the first belongs to the Italian RER and includes all the geo-localization data that allow to link each real estate unit to the ISTAT Census Section database; the second belongs to IRA and includes the geo-localization data of the REIO areas. Afterwards, we overlapped the two GIS maps and yielded our own database (BRIDGEDB, hereafter) that connects ISTAT census section and REIO areas. Therefore, the procedure allows us to link the ISTAT census sections, and indirectly the cadastral units, with each REIO area. The second problem in the alignment of URER and REIO is related to the real estate typologies, because in the 2 datasets they are differently classified. To surmount this obstacle, we built a Transformation Matrix to relate the 2 different classifications. Finally, the use of BRIDGEDB and the Transformation Matrix allows to assign the REIO real estate midrange value to each real estate unit for each cadastral category and compute the market value through the cadastral size of each real estate unit. The final database includes, therefore, all the harmonised cadastral and market data for each real estate unit. The extraordinary potentialities of this outcome are very large and can involve many aspects of the PA activities on one hand, and the household/private economy on the other. In this work, we used our Harmonised Real Estate Database to analyse the real estate allocation on the urban territory of the city of Bari relating to the differentials between the cadastral income and the REIO market value. The outcomes of the study are economically and statistically noteworthy in depicting a value discrepancy that is de facto considered as known but never numerically quantified. In this paper, we describe only the outcomes of 2 typologies of real estates, the Economic Dwellings and Villas and De-

tached Houses, based on the percentage average differentials per ISTAT census section and geo-referred to only 2 urban neighbourhoods because the latter experience the most impressive results. The average differentials are statistically necessary to guarantee the robustness of the estimates against the single differentials. Figure 1 shows the percentage average differentials of the Economic Dwellings in the Murat neighbourhood and compares 2015 data with 2018 data. Figure 2 describes the percentage average differentials of Villas and Detached Houses in the neighbourhood of Carbonara and compares 2015 data with 2018 data. As we can see in Figure 1, in the Murat neighbourhood the percentage average differentials highlight that the real estate cadastral income is always much lower than the market value, up to 80% in some cases. Furthermore, time comparison shows an increase of the differentials between 2015 and 2018 underlining the effects of requalification actions that involved the Murat historical area. In the neighbourhood of Carbonara, contrariwise, the opposite is the case. In fact, the percentage average differentials show that the real estate market values are lower than the cadastral income in the great part of the area (Figure 2). Time comparison of 2015 and 2018 data highlights that the distance between the two data increased underlining that it is less appealing to live in Villa or Detached Houses because they are located in an Outskirt or Suburb area.

Figure 1: Average differentials of Economic Dwellings in Murat. Percentage values.

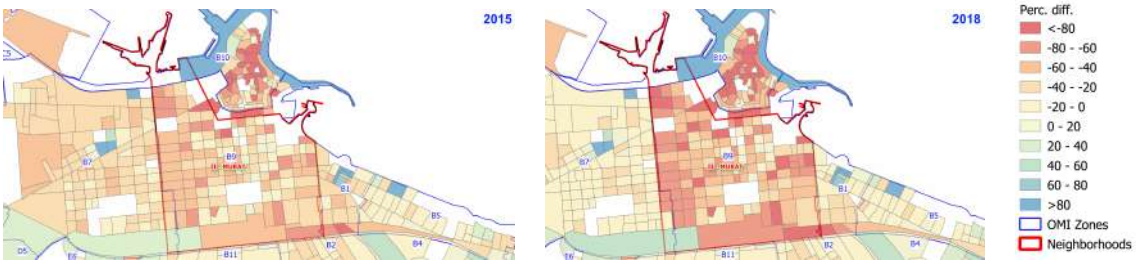
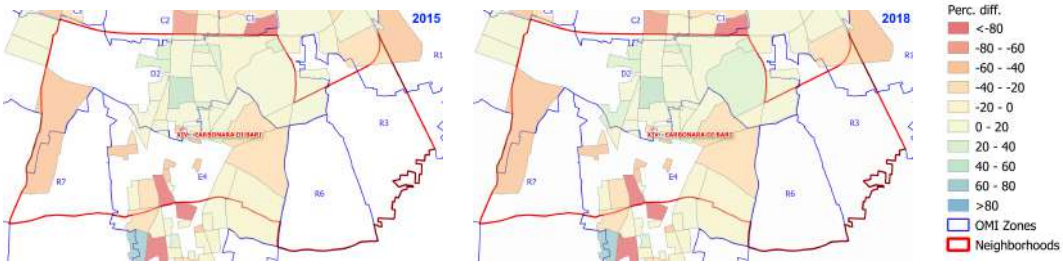


Figure 2: Average differentials of Villas and Detached Houses in Carbonara. Percentage values.



References

Calzaroni, M., (2008). Le fonti amministrative nei processi e nei prodotti della statistica ufficiale. *Atti della Nona Conferenza Nazionale di Statistica*.

Kitchin, R., (2015). The opportunities, challenges and risks of bigdata for official statistics. *Statistical Journal of the IAOS*, **31**(3), pp. 471–481.

Nordbotten, S. (2010). The Use of Administrative Data in Official Statistics - Past, Present, and Future - With Special Reference to the Nordic Countries. *Journal of official statistics*, pp. 205–223.

Thomsen, I., Holmoy, A.M.K., (1998). Combining Data from Surveys and Administrative Record Systems. The Norwegian Experience. *Journal of official statistics*, **66**(2), pp. 201–221.

The blockchain for the certification of the dairy supply chain, the "Lucanum" basket and the bakery products for well-being.

Antonio Notarnicola[b], Vito Santarcangelo[a], Nicola Martulli[b], Francesco Abbondanza[c]

[a] iInformatica Srls (PMI innovativa), Trapani, Italy;
[b] Capurso Azienda Casearia S.r.l., Gioia del Colle (BA), Italy;
[c] L'Abbondanza Lucana, Matera, Italy.

1. Introduction

Dairy products have always been the flagship of the southern Italian peninsula, and around this reality an important supply of raw materials suppliers stands.

However, preserving traditions and flavors in compliance with quality and compliance standards it is a primary objective to protect the image and values of dairy products, in fact, more and more often producers enter in the market with some product whose name remembers high quality products without guaranteeing the same standards. In this regard is really important the protected designation of origin, better known by the acronym PDO, denomination recognized to the mozzarella of Gioia Del Colle. However, keeping track of the quality of the entire supply chain process in order to remain in compliance with the PDO designation presents some problems.

In the procedural guideline of the mozzarella of Gioia Del Colle (Italian government gazette n. 18A01843) we can deduce all the characteristics to be respected to maintain the quality and the standard of the product. From some articles we can immediately notice that there is the need of technologies that allow us to keep track of the entire supply chain process to be in compliance with the PDO designation. In fact, the ART. 3 established that the area of production, processing of milk and packaging of the mozzarella of Gioia del Colle includes the administrative territory of some municipalities in the province of Bari, Taranto and part of the municipality of Matera, in Basilicata. The ART. 5 established that the milk used to produce the mozzarella of Gioia del Colle comes from dairy cows who must be raised in farms with the possibility of grazing for at least 150 (one hundred and fifty) days per year, in grazing with autumn monophytes or autumn - spring polyphites, composed by leguminous essences (clover, vetch, field bean and pea protein) and cereals (oats, barley, durum wheat, soft wheat and rye), or from natural pastures of wild herbs. The food of livestock, whose milk is used to produce the mozzarella of Gioia del Colle, is made up of grass and / or hay from polyphite in at least 60% of the total dry substance.

The aim of this work is to present the potential of the blockchain for the purpose of certifying compliance with the PDO processes of the dairy production chain (Mozzarella of Gioia del Colle), bakery products and territorial menus (Lucanum menu).

2. Blockchain for the dairy and bakery industry

The most common example of blockchain is the Bitcoin network. Bitcoin is an example of a public blockchain. When a user makes a transaction on the Bitcoin network, this is sent in broadcast to all the nodes that are part of it, in this way it is possible to certify that the transaction took place in a specific time (timestamp). Each node, better known as "miner", collects all new transactions forming a block. When a block is complete, it can be inserted into the chain as long as a cryptographic test is passed. The prof-of-work at the base of Bitcoin consists in the increase of a numerical value, called "nonce", which causes the block's

ISBN: 978-88-5495-135-8

hash to start with a certain number of zeroes. Once the cryptographic problem has been solved, the winning "miner" node communicates the solution found to all other nodes. The nodes accept the block only if all the transactions inside are valid and express their acceptance by creating the next block that contains the hash of the block just accepted. The block is therefore written in a database of blocks called "ledger", which is present in every node of the blockchain. Thanks to its distributed structure and the large number of nodes, the blockchain guarantees data immutability and business continuity. In permissioned blockchain, unlike public blockchain, access is restricted to some users only. In addition, the central authority defines the role of each user within the network, and the information it can access. The application cases presented in this paper can be implemented using both types of blockhain (permissioned and public). As already mentioned, there are several problems in tracing the whole chain process of the PDO mozzarella of Gioia Del Colle, in this regard the Gioiella app, developed by iInformatica Srls, using internet of things technology make it possible to keep track of the compliance of the supply chain process.

Each individual interaction is then certified by blockchain (public on NEM framework based or private within a MySql oriented database).

All interactions are appropriately recorded in the ledger in an anti-elusive perspective, providing validity in a legal perspective and transparency towards the final consumer.

The same applies to bakery products, following the blockchain-oriented approach of the Barile Bakery, historic producer of Altamura Bread. In this sector, even flour producers are interested to preserve the quality of their product, so to guarantee the origin of wheat from growers who use a certified organic cultivation method. In this way, flour producers defend the image and the quality of their product from unfair competition preserving consumers health and rights.

3. Semantic Fuzzy Blockchain for the Lucanum menu and LTV analysis

The same concepts presented in the above paragraph can be applied, in the same way, for the certification of territorial menus. This is the case of the restaurant “L'Abbondanza Lucana” in Matera. At present, the certification of the origin of a product is carried out with paperwork and declarations of the single producer which do not provide guarantees of temporal certification.

The use of the semantic blockchain is a valid possibility for the temporal certification of the localization and of the declarations made by the single producers. In this context the problem of mathematical computation of blockchain can be declined considering the origin of the ingredients of the individual dishes, that is to represent as hash a codification of some characters in relation to the place surveyed at the time of the insertion of a new block. In this way it is possible to correlate the ingredients by territorial origin, represented as the initial hash of the mined block. For example, if a product is registered at the GPS level as "STIGLIANO", a hash could be calculated with an initial string "STIG". This semantics is defined in a suitable knowledge base (KB) which allows to map the chain of territorial origin, according to the location required by each single ingredient. The block mining of a dish is therefore connected to the hash of the individual ingredients certified in the previous blocks, also according to the quantities requested by the dish and available at the level of the blocks. This semantics is defined in a suitable knowledge base (KB) which allows the individual ingredients to be mapped, also taking into account the quantity required by the individual dish. Considering the Lucanian dish of "chicory with bean puree", the individual chicory producers will have to certify the localization of the collected chicory through the HMI system (also with the help of multimedia support files), the same will be true for the producers of legumes and extra virgin olive oil producers.

The information about the block relative to the restaurant plate, therefore, will be correlated to the territorial semantics of the blocks of the single raw materials (beans, chicory,

ISBN: 978-88-5495-135-8

oil) that compose it and that are appropriately registered in the knowledge base. The lack of even one of these products in terms of certified blocks in the ledger will not allow to carry out the mining of the block of the realized plate. The more the dish is articulated, the more the level of complexity increases. The knowledge base (appropriately called Lucanum basket, to recall the enogastronomic elements of the Game of Basilicata) also allows to record the quantities used by each individual dish, which determine its possibility of undermining or not a certified dish based on quantities present in blocks of raw materials not yet used. In the case of mining, it will therefore be possible to certify the entire dish and also the individual ingredients that make it up.

Semantic hash increases a lot the computation, so to balance the numeric complexity, our approach considers a fuzzy hash, a revolutionary approach, that consent to obtain mining with a low complexity and better speed compared to standard blockchain. The most diffuse algorithm about hash similarity are ssdeep, sdhash, mrsh-v2, SimHash (Harichandran, V. (2016)).

The algorithm considered is ssdeep, that considering testing on dataset of literature (Breitinger,F.(2014)) is characterized by performance (accuracy) over 90%, with the best TP rate and TN rate. For this reason, the accuracy of fuzzy blockchain can be consider suitable to this semantic blockchain providing a better speed, allowing miners to mine the hash with the nonce in a lean way.

To evaluate the quality of blockchain approach applied to menu, we consider a LTV (life time value) parameter for an estimation of the investment return of the initiative over time considering the blockchain structure and user/producers value perceived.

Consequently, the LTV (t) parameter (life time value at "t" time) can be correlated to the level of complexity of the dish (CP) which is a function of the number of ingredients of the territory and the correspondence also to qualifications such as PGI and PDO.

The LTV level over time is linked to the number of blocks correctly mined (βm) considering the total blocks to mine (β), to the contribution of all the evaluation score weighted (σ) by individual user (μ), to the weighted (π) contribution of all value perceived by producers of the supply chain (ρ) related to the incremental value of the number of producers adhering to the system over time ($\Delta(\partial(\Delta(t))$).

$$LVT(t) = \frac{\beta m(t)}{\beta(t)} * (\sigma * \sum \mu) * (\pi * (\sum \rho) * (\Delta(\partial(\Delta(t)))$$

The parameter of the complexity (CP) of the dish can therefore be defined considering the presence of PGI and PDO related to the totality of ingredients (T) and a parameter called CR that is the retrieval complexity, depending on the seasonality of the product or of the offer present in the territory.

$$CP = \frac{PGI + PDO}{T} * CR$$

Moreover, thanks to the activity of implementing information in the blocks and thanks to the mining activity, the system also lends itself to being a meeting point between the enogastronomic demand and offer of the territory, signaling in time the dishes at risk, due to the complexity of finding the ingredients, in order to also generate objective reports for regional policies for the protection of typical products. In terms of HMI, the end user will then be able to view the certified menu and check the chain of individual products.

4. Conclusions

This paper showed innovative blockchain-oriented approaches aimed at complying with the product specification and improving product quality within the dairy supply chain applied to DOP processes of the mozzarella of Gioia del Colle. Alongside this application there is a

ISBN: 978-88-5495-135-8

new blockchain-oriented semantic-fuzzy initiative applied to the "Lucanum" basket of typical Basilicata products (in order to respect the territorial enogastronomic peculiarities) and to the traceability of bakery products, focused on a well-being target. We hope that these examples will be useful for improving the quality of production processes and developed products, in a context that requires always more ethics and social responsibility.

References

Capurso, F., Brandonisio, A. (2018). *Sistema basato su blockchain per la certificazione di filiera di prodotti caseari*, UIBM, 102018000021313.

Santarcangelo, V. , Massa, E. (2019). *Metodo avanzato per il miglioramento continuo basato su blockchain semantica e analisi della serendipità*, UIBM, 102019000001931.

Nakamoto, S. (2009). *Bitcoin: A Peer-to-Peer Electronic Cash System* https://bitcoin.org/bitcoin.pdf.

Calabrese, P., Stella, G., p.102018000021445, *Sistema basato su blockchain per la distruzione certificata dei documenti*.

Coretti S., Lamagna, F., Quarto, S., n. 102018000020785, *Sistema intelligente per la preventivazione rapida ed il controllo blockchain oriented*.

Fanari, F., Santarcangelo, V., Sinitò, D.C. (2018). *Esperienze di Ricerca e Sviluppo applicate alle brillanti realtà del nostro sud*, RCE Multimedia.

Barile, A. (2018). *Metodo e sistema innovativo di certificazione di ingredienti e metodi di produzione di prodotti alimentari da forno*.

Breitinger, F., Roussev, V. (2014). Automated evaluation of approximate matching algorithms on real data, *Digital Investigation*, **11**(Supplement 2), pp. S10-S17.

Harichandran, V., Breitinger, F.,Baggili, I. (2016), Bytewise Approximate Matching: The Good, The Bad, and The Unknown. *Journal of Digital Forensics, Security and Law*, **11**(2), Article 4.

Another look at the relationship between perceived well-being and income satisfaction

Omar Paccagnella[a], Ilaria Zanin[a]
[a] Department of Statistical Sciences, University of Padova, Padua, Italy.

1. Introduction

Wealth and well-being are highly investigated topics in the literature. However, they are difficult to measure because they are multidimensional concepts and there are no (or too few) objective features to take into account in their measurement. Therefore, the most adopted solutions involve subjective measures. Life satisfaction is often used to evaluate individual well-being, but it may be seen as the composition of several (subjective) domains (van Praag et al., 2003). Differential Item Functioning – DIF (Holland and Wainer, 1993) might affect individual self-assessments, because respondents may interpret, understand or use in different ways the response categories expressed on Likert (or rating) scales for the same question. If this is the case, self-reported answers become incomparable across respondents and the analysis of the variable of intertest is misleading. *Anchoring Vignettes* were introduced by King et al. (2004) as a tool to identify and correct such heterogeneity in the response scales. This allows comparability across countries or socio-economic groups of the individual subjective assessments.

Exploiting the features of anchoring vignette data to remove individual unobserved heterogeneity in the use of the response scales, this work aims at enriching the literature showing that analysing subjective well-being measures without correcting for such heterogeneity in the response scales may lead to some misleading conclusions. We do not focus on the causal links between two spheres of well-being, that is life satisfaction and income satisfaction, rather, before studying these links, we judge important to establish the strength of the relationship between the two types of satisfaction forming the object of investigation (especially when they are measured bearing in mind individuals' heterogeneous use of rating scales in this type of self-assessment).

2. Methods

Vignettes are "short descriptions of a person or a social situation which contain precise references to what are thought to be the most important factors in the decision-making or judgement-making process of respondents" (Alexander and Becker, 1978). Therefore, anchoring vignettes are additional questions, where a scenario in the same domain of the concept of interest of the self-assessment is described, to be answered by respondents after o before the self-evaluations. The self-reported questions, in the domains of this paper, are:

- How satisfied are you with your life in general?
- How satisfied are you with the total income of your household?

The life satisfaction anchoring vignettes are:

- *[John]* is 63 years old. His wife died 2 years ago and he still spends a lot of time thinking about her. He has 4 children and 10 grandchildren who visit him regularly. John can make ends meet but has no money for extras such as expensive gifts to his grandchildren. He has had to stop working recently due to heart problems. He gets tired easily. Otherwise, he has no serious health conditions. How satisfied with his life do you think *[John]* is?
- *[Carry]* is 72 years old and a widow. Her total after tax income is about € 1,100 per month. She owns the house she lives in and has a large circle of friends. She plays bridge twice a week and goes on vacation regularly with some friends. Lately she has

ISBN: 978-88-5495-135-8

been suffering from arthritis, which makes working in the house and garden painful conditions. How satisfied with her life do you think *[Carry]* is?

While the income satisfaction vignettes are:

- *[Jim]* is married and has two children; the total after tax household income of his family is €1,700 per month. How satisfied do you think *[Jim]* is with the total income of his household?
- *[Anne]* is married and has two children; the total after tax household income of her family is €3,400 per month. How satisfied do you think *[Anne]* is with the total income of her household?

For all of these questions to be assessed, the answer categories are: 1. Very Dissatisfied; 2. Dissatisfied; 3. Neither Satisfied, Nor Dissatisfied; 4. Satisfied; 5. Very Satisfied.

King et al. (2004) introduced a parametric and a non-parametric solution to analyse anchoring vignette data. The second one is the strategy adopted in this paper: it allows to recode self-ratings according to the connected set of anchoring vignettes. Self-assessments may be compared mapping them according to the scale fixed by the anchoring vignette's evaluations in each country or socio-economic group. However, each respondent has to answer to all vignette questions and evaluates the anchoring vignettes according to their *natural ranking*. Based on the content of all anchoring vignettes, that is the severity of the problem depicted in each scenario, the natural ranking may be defined as the ranking of the anchoring vignette evaluations used by the majority of the individuals belonging to the analysed country or group. Let Y_i be the self-evaluation and Z_{ij} the evaluation of the anchoring vignette j (j =1,...,J) for respondent i (i =1,...,n) and let $Z_{ij-1} < Z_{ij}$ – for all i, j – be the natural ranking. The adjusted (DIF-corrected) C_i variable is defined as:

$$C_i = \begin{cases} 1 & if \quad Y_i < Z_{i1} \\ 2 & if \quad Y_i = Z_{i1} \\ 3 & if \quad Z_{i1} < Y_i < Z_{i2} \\ \vdots & \\ 2J & if \quad Y_i = Z_{iJ} \\ 2J+1 & if \quad Y_i > Z_{iJ} \end{cases}$$

This provides a new ordinal variable, that can be analysed by standard ordered probit models, contingency tables, and so on. The main problem of this approach is that a respondent might evaluate the anchoring vignettes in a way different from the natural ranking or provide the same rating to more scenarios. This leads to some inconsistencies, defined as ties. If this happens, all answers provided by the respondent cannot be used to construct the *C*-scale.

The relationship between life and income satisfaction is investigated by means of different bivariate analyses. In a descriptive framework, several measures of association among ordinal variables are proposed, such as Kendall's *tau-b*, Somers' D and polychoric correlations. Then, bivariate ordered probit models are estimated (Greene and Hensher, 2010). This solution allows to take into account the ordering of the answer options, as well as the correlation (ρ) we expect to exist between the two self-reported measures. In the end, Heckman sample selection solutions (Heckman, 1979) are used to check whether previous results are affected by selectivity effects on the construction of the *C*-scale (individuals who do not report the natural ranking might be different to those who report it), even if we do not have reasons to support this hypothesis.

3. Data

This paper uses data from SHARE Wave 2 (DOI:10.6103/SHARE.w2.700), collected in 2006/2007 (Börsch-Supan, 2019). See Börsch-Supan et al. (2013) for methodological details. SHARE is a panel survey that collects detailed cross-national information on health, socio-

ISBN: 978-88-5495-135-8

economic status and social and family networks of citizens aged 50 and over from a large set of European countries, ranging from Scandinavia to Mediterranean nations.

More specifically, we keep all individuals who answered both to the self-reported questions on life and income satisfaction and to all their connected anchoring vignettes. Therefore, the total sample size is equal to 7353 respondents, living in 11 countries (Belgium, Czech Republic, Denmark, France, Germany, Greece, the Netherlands, Italy, Poland, Spain, Sweden). The sample is mainly composed by females (55%) and 64.3 years old on average (median: 63 years): about half of them have middle education (27% low education). About 50% of the respondents are retired, while 30% are workers.

The analysis of the collected self-assessments shows a large cross-country heterogeneity, more pronounced for income than life satisfaction. In all countries, the majority of respondents are satisfied or very satisfied with their life in general, ranging from more than 90% in Denmark and the Netherlands to less than 60% in Greece. On the other hand, studying the income domain, in six countries out of 11 the majority of respondents are not satisfied or very satisfied with their income: while Danes and Dutch people are still the most satisfied, the least satisfied country is Poland (more than 40% of not satisfied with their income).

All countries show the same natural rankings for both sets of anchoring vignettes: Carry is rated more satisfied with her life in general than John and Anne is rated more satisfied with her income than Jim. However, in both domains this natural ranking is not respected for all respondents: in each country, there are people who either identify the opposite ranking or report the same answer category for both anchoring vignettes of the domain under investigation. Such disagreement is low in the income satisfaction vignettes (about 15%, that is 1120 respondents), but larger (about 37% - 2774 individuals) for life satisfaction anchoring vignettes. The joint condition, that is the correct identification of the natural ranking in both life and income satisfaction domains, is not respected by about 45% of respondents.

4. Main results

The association between life satisfaction and income satisfaction is investigated, comparing both the collected and the rescaled self-ratings. Results are reported in Table 1. All measures show a positive association between the original life and income satisfaction variables. Analysing DIF-corrected variables, the association between these two variables are still positive, but halved in magnitude.

We then perform some multivariate analyses to support our findings. First, we estimate a bivariate ordered probit model, where life and income satisfaction are the dependent variables, both in the original dimension and in the rescaled form, controlling for a large set of explanatory variables. Many individual variables are statistically significant in each model (not always the same in both equations) and the signs have the expected directions. Results in terms of correlation are reported in Table 2. Analysing the original collected self-ratings on the whole sample, the estimated correlation between life and income satisfaction is equal to 0.390 and statistically significant at 1% of level, *ceteris paribus*. This value is very similar to the descriptive findings reported in Table 1. The same estimate is then obtained analysing only the subset of respondents who satisfied the natural rankings of the anchoring vignettes. However, the estimated correlation between the two self-evaluations strongly decreases (more than half) when the bivariate ordered probit model is estimated on the rescaled variables.

DIF-corrected measures are obtained according to the subsample of respondents who respected the natural ranking of the anchoring vignettes. We do not have reasons to support the idea that unobserved factors may affect the judgement of the anchoring vignettes in the group of respondents who do not respect the natural ranking; however, this hypothesis may be tested through the estimation of some Heckman sample selection models. Life and income satisfaction domains are investigated separately and results are reported in Table 3: analysing the original collected self-ratings and looking at the correlation coefficient between the two

ISBN: 978-88-5495-135-8

equations, selectivity effects are strongly supported, while they are no statistically significant effects studying the DIF-corrected evaluations.

Table 1: Estimation of some measures of association between income and life satisfaction

Variable	Kendall coefficient	Somers' D on life	Somers' D on income	Polychoric
Original	0.409 (0.009)	0.447 (0.010)	0.373 (0.009)	0.524 (0.011)
DIF-corrected	0.208 (0.013)	0.225 (0.014)	0.191 (0.018)	0.284 (0.017)

Table 2: Estimation of ρ in the bivariate ordered probit model analysis

Income & life satisfaction variables	ρ estimate
Original – all respondents (N = 6920)	0.390 (0.013)
Original – respondents with respected natural rankings (N = 3769)	0.379 (0.018)
DIF-corrected (N = 3769)	0.166 (0.019)

Table 3: Estimation of the correlation coefficient in the Heckman model analysis

Variable	Original	DIF-corrected
Income satisfaction	0.739 (0.125)	-0.098 (0.208)
Life satisfaction	-0.527 (0.141)	-0.040 (0.337)

5. Conclusions

On the whole, life satisfaction can be seen as an aggregate concept that can be explained by considering its several domains, which include satisfaction with one's job, health, financial situation, income, social life, and so on. The literature on life satisfaction tends to emphasise the role of income. However, researchers have to be very careful studying self-reported (subjective) data: indeed, correcting for DIF, the correlation between life and income satisfaction is *lower* than the one observing without correcting for DIF. Differences are not due to sample sizes of the analysed samples. At the same time, investigating uncorrected variables, DIF may introduce some forms of sample selection that have not reasons to exist.

References

Alexander, C.S., Becker, H.J. (1978). The use of vignettes in survey research. *Public Opinion Quarterly* **42**, pp. 93-104.

Börsch-Supan, A. (2019). *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2*. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w2.700.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology,* **42**, pp. 992-1001.

Greene, W.H., Hensher, D.A. (2010). *Modelling Ordered Choices: A primer*. Cambridge University Press. Cambridge.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, pp. 153-162

Holland, P.W., Wainer, H. (1993). *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale.

King, G., Murray, C., Salomon, J., Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, **98**(1), pp. 191-207.

van Praag, B.MS, Frijters P., Ferrer-i Carbonell, A. (2003). The anatomy of subjective well-being. *Journal of Economic Behavior & Organization*, **51**, pp. 29-49.

Profile pattern of italians NEET by nonlinear PCA

Anna Parola[a], Francesco Palumbo[b]
[a] Department of Humanities, University of Naples Federico II, Naples, Italy.
[b] Department of Political Sciences, University of Naples Federico II, Naples, Italy.

1. Introduction

This work aims to investigate psychological factors associated with the status of NEET: those young people who are not in education, employment or training. The recently increased complexity in the labor market has had an impact on the school-to-work transition. In particular, in the Italian context, the existence of groups who during the critical period of the late teens spend a substantial amount of time outside any form of education, employment, or training (NEET) takes on ever-increasing aspects of risk of predicting and planning the working future (Savickas, 2012), and of psychological well-being (Parola and Donsì, 2018; Paul and Moser, 2009).

According to Bynner and Parsons (2002), the social and economic context of youth transitions is critically important in determining their shape and outcomes. Due to transformations in the economy, society, and technology, the world of work has seen dramatic changes resulting in employment insecurity, uncertainty and fragmented career paths (Savickas, 2012). Among individuals choosing jobs and constructing careers, the current work world provokes feelings of anxiety and insecurity (Savickas, 2012). Moreover, the most recent literature shows that unemployed young people are a high-risk mental health problem group. Several reviews and meta-analyses link unemployment and psychological health (McKee-Ryan et al., 2005; van der Noordt et al., 2014). A recent systematic review (Bartelink et al., 2019) reports that unemployment is associated with increased mental health problems, depression, and anxiety disorders.

To study the symptomatology of young NEET we referred to the Achenbach and Rescorla (2001) model. According to this model, problem behaviors have dichotomized into two empirically established syndromes reflecting internalizing problems (Anxious/Depressed, Withdrawn, Somatic Complaints) and externalizing problems (Aggressive Behavior, Rule-Breaking Behavior, Intrusive).

This article proposes a study on a sample of 150's Italian NEET young people aiming to identify homogenous groups concerning the NEET condition symptom-profiles. A psychometric questionnaire was administered to the subjects in the sample, and then a hierarchical clustering algorithm was used to identify homogeneous groups in a reduced subspace obtained through the nonlinear principal component analysis (NL-PCA) according to Gifi (1990). Finally, the diverse emerged NEET symptom-profiles were considered concerning their socio-demographic variables.

The following section illustrates in brief the statistical approach; section 3 presents and comments the main findings and it empirically shows that the classical PCA can lead to unsatisfactory biased results when the variable distributions are skewed and correlation among variables is non-linear. Last section presents some concluding remarks.

2. Statistical analysis

The proposed typological approach of identifying groups based on the presence/absence of problem behaviors is closely related to clinical practice. Such a typological approach allows for investigating the degree to which different problem behaviors co-occur in young adults with an unemployment condition. Therefore, it allows to discover different problem groups and how many young can be assigned to these groups. To support that, we propose a cluster analysis (CA) on the nonlinear principal component analysis coordinates. CA allows identifying homogeneous groups with respect to the symptom-profiles related to the NEET condition. Generally, to get

ISBN: 978-88-5495-135-8

more stable results, CA is performed on a reduced subspace obtained by a principal component analysis (PCA) on the original variable set (Fabrigar et al., 1999; Linting et al., 2007). However, in the present work, we empirically show that in the analysis of our NEET data a nonlinear PCA offers better results, as it accounts non-linear relationships between variables. Such a non-linear association largely depends on the highly skewed variable distributions. Therefore, nonlinear PCA better highlights the different symptom profiles of the youngs NEET.

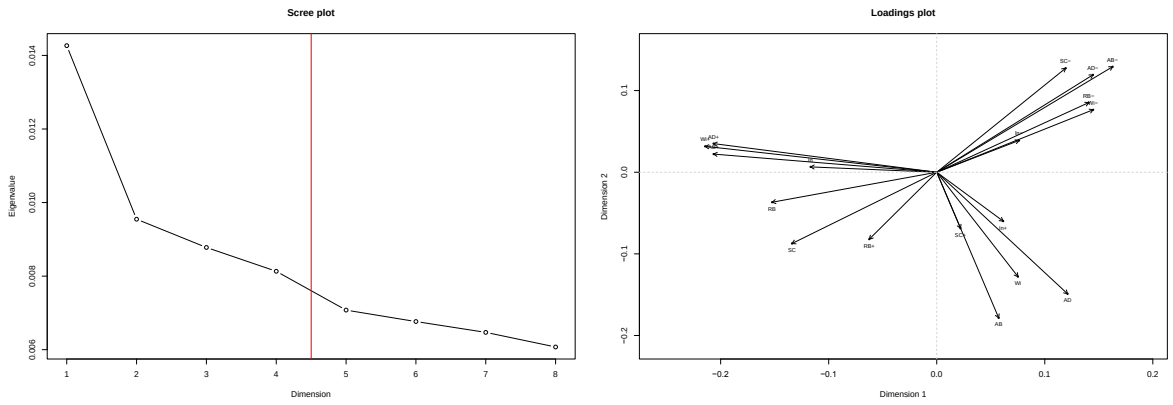
In particular, we have compared the results of the two-step analyses, PCA and Non-linear PCA, followed by a hierarchical CA, performed on the data collected through the administration of the Adult Self Report 18-59 questionnaire (ASR; Achenbach and Rescorla, 2001) to a sample of 150 Italians NEET. The instrument consists of 77 items that define six psychometric scales corresponding to the following six-syndrome structure: Anxious/Depressed (AD; 18-items), Withdrawn (W; 9-items), Somatic Complaints (SC; 12-items), Aggressive Behavior (AB; 15-items), Rule-Breaking Behavior (RB; 14-items), and Intrusive (I; 6-items). Items were recorded using a three-point ordinal scale.

Performing Non-Linear PCA on the six scales, a basis expansion was considered to recode the original variables; via a three-knots, order two b-splines function, where the median corresponds to the central node, scores were fuzzy recorded in three ordinal levels: low, medium and high. Then, HOMALS analysis was performed on the recoded variables (HOMogeneity by means of Alternating Least Squares, Gifi 1990), finally a hierarchical clustering algorithm based on the Ward criterion was applied on the first principal coordinates in the factorial subspace. Ward criterion aims to minimize the variance within clusters at each step of grouping.

Five respondents were removed from the analysis as were identified as being straightliners, or flatliners respondents, as they selected the top level for all questions.

3. Results and discussions

The plots in Figure 1 show the results of the Non-linear PCA. As PCA, Non-linear PCA renders two representations that provide useful information: (1a) PCA scree-plot and (1b) factor loadings. The scree-plot helps to identify the intrinsic dimensionality in four factors. The loading plot interpretation is like the classical PCA, albeit in this approach each variable has been recorded in three levels: (+) high, medium and low (-). Looking at the loadings, it is worth noting that modalities associated with the lower values are in the first quadrant (clockwise). Central and high categories of the considered variables are represented in the remaining quadrants. Specifically, the highest values of the scales AD, W and AB are collocated in the fourth quadrant, whereas the highest value of the scale RB is in the third quadrant. Finally, the highest values of the variables SC and I are collocated in the second quadrant. Plots 2(a) and 2(b) in Figure 2 show the representation of the statistical units with respect to the first two factors (left-hand

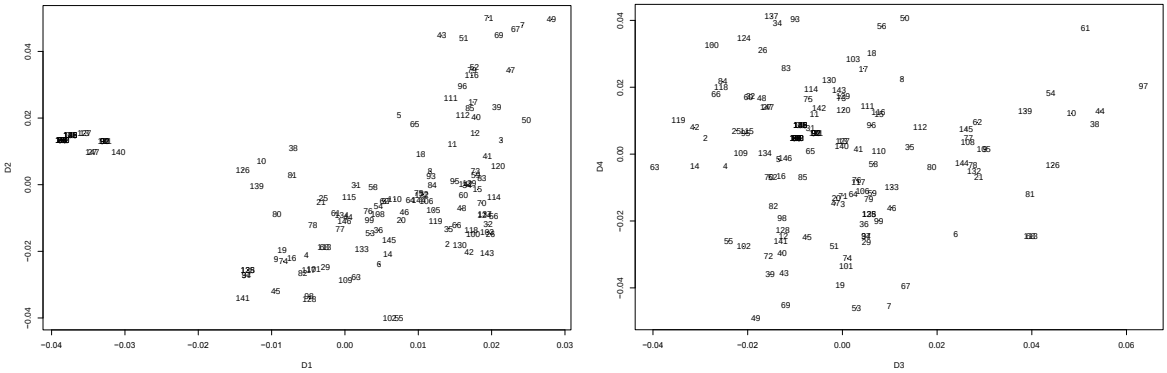


(a) Scree-plot (D1-D8)

(b) Loadings (D1-D2)

Figure 1: First results of Homogeneity analysis

ISBN: 978-88-5495-135-8



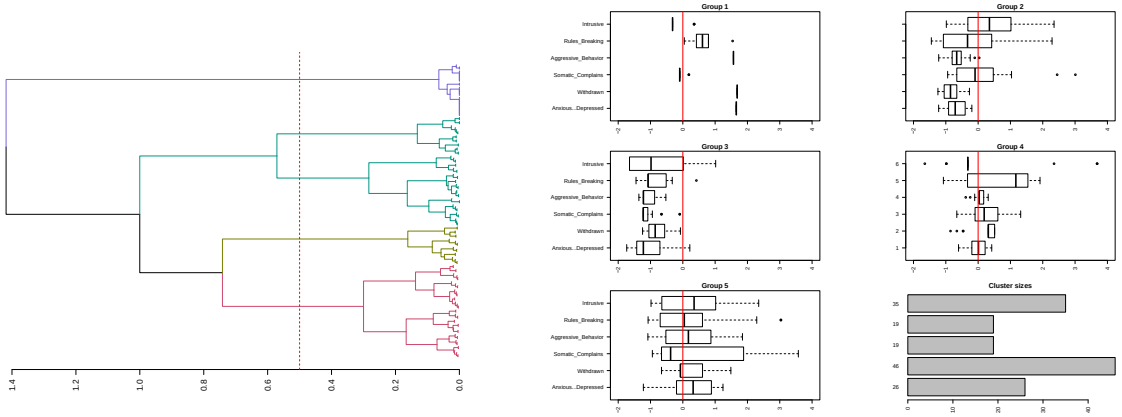
(a) First factorial plan (D1-D2) (b) Second factorial plan (D3-D4)

Figure 3: Representation of individuals on the first two factorial plans

side) and third and fourth factors on the (right-hand side), respectively.

Along with Non-linear PCA, the cluster dendrogram illustrates the hierarchical structure of the clusters defined by the analysis (Figure 3(a)). Clusters have been defined by cutting the dendrogram (cut=5). Five boxplots of the scale *z scores*, one for each cluster, and a bar chart of the cluster sizes are shown in Figure 3b. Looking at Figure 2(b), the first cluster (n=26) includes individuals having the highest value for the scales AD, W and AB; the second one (n=46) presents values above the average for the I scale. In the third cluster (n=19), all syndromic scale values are below the average. The fourth cluster (n=19) presents the highest value above the average for the RB scale. In the last cluster (n=35), the scales I, AB and AD have median values above the average. Among the possible configurations, in the first profile anxiety and depression are linked to a socially withdrawn behavior but also to aggressive behaviors, having difficulties in in managing internal dynamics with consequent acting out. In the second profile, the intrusive problem stands out. This health evidence could indicate a less adaptive attempt to regulate unpleasant mental states through the ambivalent search for the relationship with the other. The third profile presents a less malaise NEET condition. In the fourth profile, a highly externalizing problematic related to the transgression of social norms emerges. As in the first profile, in the fifth profile anxiety and depression are linked to an intrusive and aggressive behavior. In this case, in addition to acting out, the individuals manifest ambivalent search for the relationship with the other.

For sake of space the paper does not show the results obtained with the classical PCA. However, to highlight the difference in the cluster analysis, Table 1 cross-classifies the units with respect to the two different factorial approaches. The clustering algorithm is based on the



(a) Clustering Tree (Ward Criterion) (b) Boxplots and cluster size (5 Groups)

Figure 2: Results of the hierarchical cluster analysis

Ward criterion in both cases. First, it is important to remark that PCA reports four groups unlike the well-separated five groups obtained on the NL-PCA coordinates. Moreover, it is worth noting that the first group has been equally identified by both approaches, whereas it is difficult to find any correspondence between the remaining groups.

		Five groups on NL-PCA				
		1	**2**	**3**	**4**	**5**
Four groups on the PCA	**1**	26	0	0	0	1
	2	0	19	18	8	11
	3	0	9	1	10	11
	4	0	18	0	1	12

Table 1: Confusion matrix between PCA and NL-PCA groups

4. Conclusion

The exploratory statistical analysis of the six psychometric scales considered for the study of the NEET psychological profiles revealed that the distribution were skewed or highly skewed with non linear correlations. Such a dependence structure in the data can mask the presence of the true clusters and may lead to trivial results. To avoid this drawback, this paper work a non-linear analysis for the identification, classification and discrimination of different symptom-profiles related to the NEET condition. This analysis has demonstrated to be a valid alternative to the use of PCA. The identification of different profiles is useful for clinical intervention and planning the career counseling.

References

Achenbach, T. M., Rescorla, L. A. (2001). *The manual for the ASEBA school-age forms and profiles.* University of Vermont: Research Center for Children, Youth, and Families, Burlington, (VT).

Bartelink, V. H., Zay Ya, K., Guldbrandsson, K., Bremberg, S. (2019). Unemployment among young people and mental health: A systematic review. *Scandinavian journal of public health*, pp. 1-15.

Bynner, J., Parsons, S. (2002). Social exclusion and the transition from school to work: the case of young people not in education, employment or training. *Journal of Vocational Behavior*, **60**, pp. 289-309.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, **4**(3), pp. 272-299.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Wiley, Chichester, England.

Linting, M., Meulman, J. J., Groenen, P. J., van der Koojj, A. J. (2007). Nonlinear principal components analysis: introduction and application. *Psychological methods*, **12(**3), pp. 359-379.

McKee-Ryan, F., Song, Z., Wanberg, C. R., Kinicki, A. J. (2005). Psychological and physical well-being during unemployment: a meta-analytic study. *Journal of applied psychology,* **90**(1), 53-76.

Parola, A, Donsì, L. (2018). Sospesi nel tempo. Inattività e malessere percepito in giovani adulti NEET. *Psicologia della Salute*, **3**, pp. 44-73.

Savickas, M. L. (2012). Life design: A paradigm for career intervention in the 21st century. *Journal of Counseling and Development*, **90**, pp. 13-19.

van der Noordt, M., Ijzelenberg, H., Droomers, M., Proper, K.I. (2014). Health Effects of Employment: A Systematic Review of Prospective Studies. *Occupational and Environmental Medicine*, **71**, pp. 730-736.

ISBN: 978-88-5495-135-8

Assessing mental health therapeutic communities functioning

Anna Maria Parroco[a], Vincenzo Giuseppe Genova[b], Laura Mancuso[a], Francesca Giannone[a]

[a] Department of Psychology, Educational Science and Human Movement, University of Palermo, Italy.

[b] Department of Economics, Business and Statistics, University of Palermo, Italy.

1. Introduction

This work is part of a research that aims at introducing empirical investigation methods for understanding and evaluating treatments in Mental Health. A commitment that is attributable to a more general clinical and social requirements: the ethical and scientific interest of the professional community of Mental Health for the promotion of effective care interventions as well as the entrepreneurization processes of the health sector which is designed to identify and consolidate proven effective care practices.

The aim of this paper is the evaluation of the functioning of residential therapeutic communities for severe patients (Angelini *et al*. 2017). Therapeutic communities are complex settings where numerous organizational and relational variables act (structures, activities, care characteristics, relationship between members, group dynamics). The empirical assessment of their functioning is a complex challenge, with clear implications for the promotion of effective interventions and for the improvement of the quality of care. The results of this study constitute a first step forward to understand, through a quantitative approach, a context which is complex both for the multiplicity of the involved stakeholders and the treatment variables to be analyzed.

2. Materials and methods

In order to analyse functioning, survey data have been collected. Two instruments have been used to gather the data: a self-report questionnaire at an individual level and an assessment table at the community one. The first, the VIVACOM Questionnaire (VIsiting for VAluation of COMmunities), (Biaggini *et al*. 2012) consists of 77 items, clustered in ten dimensions representing the major areas of communities functioning (general organisation, personalisation and rights, therapeutic climate and setting comfort, general treatment features: individual and group, family-focused activities, resident and caregiver safety, staff management and training, organisational supplements and collaboration, clinical documentation and reporting system, quality assessment and research). Respondents are asked to give a score y_{ij} on these set of items, with a scale bonded at both ends (1 to 5). For each dimension an index of functioning ($FI_{i\ i=1,2,..10}$) has been calculated as the average of the items score assigned by the respondent. A general Functioning Index (FI) has also been obtained to get a general score for each community.

The assessment table, the GAS-SET (Grid for the Analysis of Set(ting)-Set) (Giannone and Lo Verso, 2011; Bruschetta, 2014) reports data on each community that mainly concern the range of services that aid and support the patient during the care program. The table is organized into different sections: structure data, collaborations, treatment data, intervention area, psychotropic drugs management, community regulation, activities area, staff member and patients data. Based on these information, several structural and organizational variables have been identified. In this study, variables have been aggregated to obtain dichotomous indicators, which assume value equal to 1 when the community offers a number of service equal or grater than the average, and 0 otherwise.

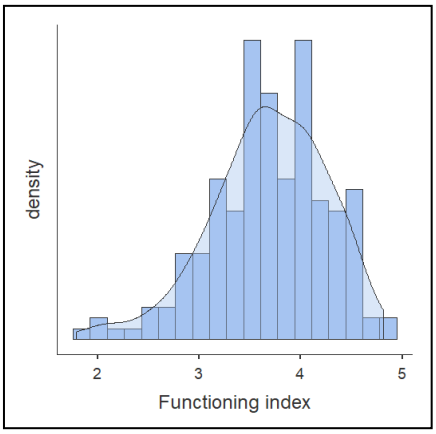
The sample under investigation includes 18 adult therapeutic communities, located in Italy,

ISBN: 978-88-5495-135-8

and 191 units. All the participants have been involved in the Visiting DTC project, a national Italian - training, evaluation, research and accreditation of "Community Group-Quality" project, promoted by LegaCoop-sociale.

As shown in figure 1, FI has an asymmetric distribution (*M = 3.68, S = 0.570*). Although it provides a synthetic measure of functioning, a tendency to assess functioning with medium and high scores (*Q1= 3.32; Q2=3,7)* has been observed. We think that this function could be considered as a *proxy* of the recognition, in staff members' perception, of a good work that has very often reached high levels of efficacy, so corresponding to the programmed targets.

Figure 1: FI Plot



The main descriptive statistics of the Functioning Index (FI) according to the levels of the factors with which it could be potentially associated, are reported in Table 1.

Table 1: Summary statistics of the Functioning Index by the levels of the potentially explanatory variables.

Variable	Categories	Functioning Index (FI)			
		N	Mean	Median	St.dev
Gender	F	129	3.72	3.73	0.561
	M	57	3.61	3.58	0.584
Age	<=30	19	3.76	3.71	0.559
	31-40	41	3.77	3.85	0.535
	41-50	45	3.70	3.73	0.525
	>=50	69	3.67	3.69	0.576
Professional Role	Expert	13	3.85	3.94	0.489
	Health	105	3.62	3.65	0.595
	Socio-Pedagogical	71	3.75	3.73	0.538
Collaborations	0	86	3.51	3.61	0.575
	1	105	3.82	3.82	0.530
Treatment data	0	64	3.51	3.53	0.657
	1	127	3.77	3.74	0.503
Intervention area	0	46	3.91	4.14	0.741
	1	145	3.61	3.63	0.486
Psycotropic drugs management	0	106	3.48	3.51	0.508
	1	85	3.94	4.06	0.543
Community regolation	0	73	3.50	3.53	0.546
	1	118	3.80	3.82	0.558
Activities area	0	56	3.65	3.67	0.555
	1	135	3.70	3.72	0.578

This first results reveal a high presence of women workers (69%) compared to men (31%). Nearly two out of three (65%) were more than 40 years of age while only 10% were younger than 30 years of age. Regarding their professional role, 60% of respondents carried out its role in

health, 30% in socio-pedagogical area, 10% were experts (psychologist, psychiatrist and pedagogist). The average values of FI conditioned to the levels of potentially explanatory variables (Col.4, Table 1) show that respondents working in communities whose number of activities is higher than the mean, award higher scores than the others, with the exception of intervention fields.

We are interested in the effects of the variables described in table 1 on FI. This presentation offers only the analysis of the general functioning index, as an example of the adopted method both for the sake of brevity and because data collection is still in progress. Due to the characteristics of Y, a beta regression model with random intercept, has been adapted and thus the dependent variable has been converted in a (0,1) interval (Verkuilen and Smithson, 2012). After estimating the null model without any explanatory variables, which highlights that the between-community variance is non-zero, a beta regression model, with a link logit and a random intercept was estimated.

3. Results

In order to consider the effects of the set of the considered independent variables on FI, the results relating to the estimated random intercept model are reported in Table 2. This model has been compared with the one including the fixed effects of all the independent variables above listed. The Chi-squared test used to evaluate the best fit of data does not show a significant difference between the two models, so only the estimates relating to the more parsimonious model have been reported.

Albeit subject to the limitations pointed out by the reduced sample size, which may explain a large part of overdispersion, these findings could make stakeholders reflect on some aspects of communities work. By analysing the estimates reported in Table 2, we observe that a positive effect on the dependent variable is higher when the indicators "Care data" and "Psychiatric drugs" take the '1' value, compared to the reference level, that is they count a greater number of interventions than the average. This is more evident for the "Care data-Section" for which the estimated coefficient is three times larger then the size of "Psychiatric drugs" estimated coefficient. In the opposite direction "Intervention Area" and "Activities Area" move: a positive effect is obtained as the number of activities related to these dimensions is below the average. Similarly looking at professional roles: for both levels of this variable, the results show a negative effect compared to the reference category, which is composed by experts.

Table 2: Model results

Random effect	Variance	Std. Dev		
Community (Intercept)	0,04561	0,2136		
Conditional model:	Estimate	Std. Error	Z value	Signif. level
(Intercept)	1,8897	0,2157	8,761	***
Treatment data	1,3383	0,3049	4,389	***
Intervention area	-0,6169	0,1482	-4,161	***
Psycotropic drug management	0,3314	0,1234	2,686	**
Activities area	-1,2856	0,3146	-4,086	***
Health role	-0,5389	0,1438	-3,755	***
Socio-pedagogical Role	-0,5084	0,1455	-3,47	***
Overdispersion parameter for beta family (): 32.9				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

So, among the variables examined by the Gas-Set, the activities indicated in the "Care data" section are those that seem to have the greatest impact on a positive evaluation of the communities functioning. These activities are related to the work of construction, sharing and transparency of therapeutic planning for individual users and for the regulation of community life. Conversely, a

greater commitment in the "intervention area" and "activities area" sections shows worse assessments. These results are more difficult to interpret: do the operators believe that the commitment to more areas of intervention and activities does not allow to sufficiently focus on their work? Is this true also on the relationship with the user? Could the awareness of the complexity and commitment for an adequate care intervention reduce the evaluation of the quality of one's work? The issue will require greater reflection and in-depth study.

4. Conclusion

As we have seen, the final results of this work appears to provide useful knowledge and insights into the community functioning as well as the model seems to have the potential for wider replication, even if some challenges remain. Based on empirical data, we have observed that advices received from health care professionals are better, primarily, when the community commitment is focusing on the implementation of customized therapeutic residential projects, shared therapeutic interventions, as well as when a democratic construction of the general regulation is adopted. It is also important the psychiatric drugs management system: promoting the user self-management and the taking in charge directly by the community and the involved services. The relationship with several auxiliary activities needs to be addressed further. Finally, to sum up, the use of a quantitative approach allows to identify specific, recognizable and in some way measurable actions, implicated in the functioning of these contexts, trying to connect "what is actually done" (Guarnaccia et al. 2019), with the evaluation of people living the experience and this could have a direct operational usefulness for the community stakeholders. Equally, conditions concerning the statistical model performance draw our attention to the need for a large sample, in order to put forward practical proposal for the improvement of the involved organizations under a framework based on significant variables to the understanding of the community therapeutic treatments.

References

Angelini A., Bruschetta S., De Crescente M., Gaburri L., Giannone F., Mingarelli L., Pismataro C.P., Vigorelli M. (2017). The “Visiting in Italy” project: origins, organisation and prospects. *Funzione Gamma*, **39**, pp. 1-18.

Bruschetta, S. (2014). GAS-Set Dispositivi terapeutici di sostegno all’abitare. In R. Barone, S. Bruschetta, and A. Frasca. *Gruppoanalisi e sostegno all’abitare*. Franco Angeli, Milano, pp.1-10.

Biaggini M., Bisanti R., De Crescente M., Gaburri L., Ghisotti N., Martini S., Moschetti S., Pismataro C., Vigorelli M. (2012). Visiting per la valutazione delle Comunità Terapeutiche. Manuale VIVACOM. In *Le Comunità Terapeutiche. Psicotici, borderline, adolescenti e minori*, eds A. Ferruta, G. Foresti, M. Vigorelli. Raffaello Cortina, Milano, pp. 551-564.

Giannone, F., Lo Verso, G. (2011). Epistemologia, Psicologia Clinica e Complessità. In *Gruppoanalisi Soggettuale*, eds G. Lo Verso, M. Di Blasi. Raffaello Cortina, Milano, pp. 17-57.

Guarnaccia, C., Ferraro, A.M., Lo Cascio, M., Bruschetta, S., Giannone, F. (2019). The SCIA Questionnaire: standards for communities for children and adolescents. A tool for the evaluation of good practices. *Therapeutic Communities: The International Journal of Therapeutic Communities*, (40), pp.1-15.

Verkulein, J., Smithson, M. (2012). Mixed and Misture regression Models for Continuous Bounded Responses Using the Beta Distribution. *Journal of Educational and Behavioural Statistics*, **37**(1), pp.82-113.

ISBN: 978-88-5495-135-8

A pre-post sensory experiment on the effect of a seminar on olive oil preferences of Italian consumers

Eugenio Pomarici[a], Alfonso Piscitelli[b], Luigi Fabbris[c], Raffaele Sacchi[d]
[a] Department of Land, Environment, Agriculture and Forestry, University of Padua, Italy.
[b] Department of Political Sciences, Federico II University of Naples, Naples, Italy.
[c] Department of Statistical Sciences, University of Padua, Padua, Italy.
[d] Department of Agricultural Sciences, Federico II University of Naples, Naples, Italy.

1. Introduction

The relative importance of different attributes of olive oil helps consumers to determine their preference for a hypothetical product (Van der Lans *et al.*, 2001; Krystallis and Ness, 2005; Mtimet *et al.*, 2013). Physical characteristics such as colour, taste, and flavour play an important role in consumers' perception (Grunert, 1997).

The purpose of this study was to measure the effect of information on taste and other sensory aspects of olive oils on a sample of Italian consumers. This paper describes the design and the results of a pre-post experiment aimed at evaluating the possible effects of a brief seminar on sensory perceptions of olive oil. The experiment was designed in such a way as to control the effects of both the selection of the sample of consumers treated with the seminar (namely, 'treated' vs 'control' samples) and the possible spontaneous maturation of treated consumers due to a pre-seminar measurement of their preferences for the same evaluated olive oils (namely, 'post' vs 'pre' measurement). The experimental design and the types of assessed olive oils are described in detail in Section 2.

The seminar, lasting about one hour, was held after a first tasting session and aimed to inform participants about the commercial classification of olive oils, their olfactory and flavour characteristics and problems, their aftertaste, and how to identify qualified and/or fresh oil. Olfactory sensations referred to fruity olive as well as tomato, artichoke, apple, eucalyptus, and other flavours. Tasting properties referred to a significant but not extreme sour and spicy sensation. The participants were involved in the topical issue of the seminar as if they were expert judges able to sensorily evaluate the chemical and physical qualities of oils. The participants showed interest in the seminar issues, asking questions and generally interacting with the experts giving the seminar.

Three categories of olive oils were included in the experiment: virgin, extra virgin, and crystal clear. Of these, the extra virgin one is considered the best. Some extra virgin productions can be protected either by a designation of origin (PDO) or by a "100% Italian" label. Crystal clear oils, though belonging to the general category of virgin oils, possess chemical and physical as well as organoleptic characteristics that make them unsuitable for human consumption. For this reason, they are refined and possibly mixed with other virgin oils and then commercialised as "olive oils".

In this paper, we present a synthesis of the results of the experiment (Section 3) and the conclusions drawn thereupon (Section 4).

2. Data and methods

A pre-post sensory evaluation experiment was conducted on five olive oils, four extra-virgin olive oils (EVOOs) and one olive oil. For the first tasting session, 246 assessors were randomly assigned either to the treated (n = 117) or to the control group (n = 129). After one week, the second tasting session involved 99 assessors who had in the meantime participated in the seminar and 106 who had not participated. The data analysis included only people who participated in both tasting sessions.

For each oil tasted, the assessors were asked to express a preferential judgment by stating a

ISBN: 978-88-5495-135-8

value on a nine-point scale, where one means *awful* and nine means *excellent*. All tasting experiments were conducted under a double-blind control procedure. After the first tasting session, the treated group attended the seminar.

Five oils were tasted in each session. Oils differed in quality, price, and sensory profile. All oils can be found in retail shops. Sixty per cent of the sampled olive oils were produced in Italy and the remaining in other EU countries. In particular, the experiment included the following samples:

(i) a so-called first prime-price (PP) oil which has a light olive fruitiness and is strongly characterised by a note of eucalyptus;

(ii) a lower-price Italian (ITA) oil which presents a fruity flavour of clean olive with subtle notes of apple;

(iii) a medium-price Italian oil (HQ) characterised by a medium-intensity fruity olive flavour with notes of leaf and almond and a slightly prevalent pungency of bitter;

(iv) a high-price Protected Designation of Origin (PDO) olive oil with an intense fruitiness, in whose aromatic profile it is possible to distinguish the aromas of green tomato, fresh leaf, and green almond; and

(v) a mixture of rectified crystal-clear oil and virgin oil (OLV) almost completely devoid of the olfactory-gustatory attributes of virgin oils.

The olive oil samples tasted, were served at room temperature in plastic cups. Each cup contained 3 ml of oil and had a transparent plastic lid. White bread and water were provided as a sample carrier. A commercial brand was selected in order to prevent taste or quality variability in the bread (Porretta, 2000). All participants tasted the full set of five oils and the administration order was randomised so as not to condition the judgements on the tasting sequence.

To facilitate comparability between the treated and control samples, as far as possible, we equalize the two samples through propensity score approach. We computed propensity score that allowed matching each treated unit using a near-neighbour strategy (Rosenbaum and Rubin, 1983). Thereby, 188 cases were matched, equally partitioned between the treated and control samples. The unmatched samples were excluded from further statistical analysis.

The effect of the seminar on the ability of participants to perceive oils' qualities required a joint comparison between the pre- and post-experiment tastings and between the treated and control groups, which is called difference-in-difference estimation. To estimate the conditional difference-in-difference effect of the seminar on the olive oil evaluation skills we adopt an ordinary least squares regression, given also the tendency to normality of scores. The results of the before (0) and after (1) seminar measurements and of the effect estimation are presented in Table 1.

Predicted propensity score and matching between assessors were performed using packages MatchIt (Ho *et al.*, 2011), and matching (Sekhon, 2011) implemented in the statistical computer programming environment R (R Core Team, 2019). Also, the ordinary least squares regression was performed using the R package.

3. Results

The post-hoc judgements of the experimental group tended to be lowest for the rectified olive oil, which is considered of the lowest quality, intermediate for extra virgin oils, and highest for the oil considered of the highest quality among the five administered. However, the trend was similar, though less evident, both during the first tasting and among the control group. This shows that people participating in the experiment were generally aware of oil quality and that they took their role in the experiment seriously. Delving deeper into the data, one can observe that the control group's scores were substantially similar in both the first and the second tasting session. However, in the second trial, the scores were higher for the medium-quality oils and lower for the rectified

ISBN: 978-88-5495-135-8

oil. This phenomenon was also observed among the experimental sample. This could mean that second judgements are conditioned by instinctive psychological factors. One factor could be complaisance, that is, a desire to please the experimenter by enhancing their judgement in a way that they guess the experimenter expects. Another factor could be maturation, that is, the ability to learn from the experience, which means that the more the assessors tasted, the more their judgemental capability grew. As we do not have enough evidence to answer this question, we leave this topic to future research.

A first consequence of the simultaneous changes between the scores from the treated and control samples is that, for four out of five oils, the post-pre differences and, even more, the difference-in-differences were not significant. This is attributed to the wide variability of scores among participants and, in any case, it shows that the seminar did not affect the participants' ability to identify the low- and medium-quality oils.

The only statistically significant difference in time and through the samples concerns the best olive oil, the one qualified by a protected designation of origin. The difference in time between the treated units is 1.3, and that between the treated and control samples during the second tasting is of 1 point out of a maximum of 9. This eye-catching difference concerns the interaction between time and group, which is 1.4 points and is statistically significant. Likely, the consumer sample recognised in this type of oil not only an intense fruity flavour and an equilibrium between bitter and spicy tones but also the note of tomato which characterises this type of oil and was an issue during the seminar.

Table 1: Results of the experiment on the effects of an educational seminar on olive oil preferences, by type of oil and experimental category (significance within parentheses).

Type of oil	*Time*	Average scores		OLS coefficients		
		Control	*Treated*	*Time*	*Treatment*	*Time*Treat*
OLV	0	3.096	3.564	-0.181	0.468	-0.319
	1	2.915	3.064	*(0.537)*	*(0.110)*	*(0.441)*
PP	0	4.851	5.362	0.266	0.511	0.219
	1	5.117	5.840	*(0.408)*	*(0.112)*	*(0.639)*
ITA	0	5.479	5.489	0.319	0.011	0.383
	1	5.798	6.191	*(0.233)*	*(0.968)*	*(0.312)*
HQ	0	5.585	5.409	0.468	-0.177	0.219
	1	6.053	6.096	*(0.121)*	*(0.559)*	*(0.608)*
PDO	0	5.872	5.511	-0.064	-0.362	1.383
	1	5.809	6.830	*(0.833)*	*(0.233)*	*(0.001)*

4. Discussion

The results indicate that the information provided may influence the capacity of consumers to recognise the essential sensory characteristics of olive oils. The seminar focused on the ability of a consumer to recognise certain general indicators of oil quality which correlate with the intrinsic presence of virtues and the absence of flaws. That is why people participating in the experiment were able to correctly order the oils by quality.

It is worth noting that previous research, both at the national and international level, focused on the effect on consumer preferences that information about extrinsic characteristics had, such as geographical origin, price, trademark, and the like (see, among others, Grolleau and Caswell, 2005; Espejel *et al.*, 2007; Dekhili *et al.*, 2011; Salazar-Ordóñez *et al.*, 2018). Of course, choices based on extrinsic characteristics differ widely from those based on extrinsic ones. However, it is important to stress that the consumers involved in the experiment judged oils according to easy-to-perceive (that is, non-technical), general attributes, which this puts them closer to experts.

Another conclusion is that practicing tasting may make consumers aware of the quality of

ISBN: 978-88-5495-135-8

many oils but the very good ones. All treated and control groups identified the low quality of the rectified oil. Indeed, the seminar significantly affected the capacity of consumers to recognise the intrinsic qualities of the best oil. This means that if different oils are systematically tasted, the comparison capacity develops instinctually, at least up to a certain point. Beyond that point, only specific training can improve the capacity to recognise quality.

References

Dekhili, S., Sirieix, L. and Cohen, E. (2011). How consumers choose olive oil: The importance of origin cues. *Food quality and preference*, **22**(8), pp. 757-762.

Espejel, J., Fandos, C. and Flavián, C. (2007). The role of intrinsic and extrinsic quality attributes on consumer behaviour for traditional food products. *Managing Service Quality: An International Journal*, **17**(6), pp. 681-701.

Grolleau, G. and Caswell, J.A. (2005). Interaction between food attributes in markets: the case of environmental labeling, Working Paper No. 7, Centre d'Economie et Sociologie Appliquées à l'Agriculture et aux Espaces Ruraux.

Grunert, K.G. (1997). What's in a steak. A cross-cultural study on the quality perception of beef. *Food Quality and Preference*, **8**(3), pp. 157-174.

Ho, D.E., Imai, K., King, G. and Stuart, E.A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, **42**(8), pp. 1-28.

Krystallis A. and Ness M. (2005). Consumer preferences for quality foods from a South European perspective: a conjoint analysis implementation on Greek olive oil. *International Food Agribusiness Manage Review*, **8**(2), pp. 62–91.

Mtimet N., Zaibet L, Zairi C. and Hzami H. (2013). Marketing olive oil products in the Tunisian local market: the importance of quality attributes and consumers' behavior. *Journal of International Food Agribusiness Marketing*, **25**(2), pp. 134–145.

Porretta, S. (2000). *Analisi sensoriale & consumer science*. Chiriotti Editori, Pinerolo (TO).

Rosenbaum, P.R., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), pp. 41–55.

R Core Team (2019). R: A language and environment for statistical computing. Wien: R foundation for statistical computing. Available online at: *http. www.R-project.org*.

Salazar-Ordóñez, M., Rodríguez-Entrena, M., Cabrera, E. R. and Henseler, J. (2018). Understanding product differentiation failures: The role of product knowledge and brand credence in olive oil markets. *Food quality and preference*, **68**, pp. 146-155.

Sekhon, J.S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, **42**(7), pp. 1- 52.

Van der Lans I.A., Van Ittersum K., De Cicco A. and Loseby M. (2001). The role of the region of origin and EU certificates of origin in consumer evaluation of food products. *European Review of Agricultural Economics*, **28**(4), pp. 451–477.

Understanding local administrations policies effects on well-being in Italian inner areas

Luca Romagnoli[a], Luigi Mastronardi[a]
[a]Department of Economics, University of Molise, Campobasso, Italy

1. Introduction

The paper proposes an analysis of the relationships between Italian local administrations policies and well-being at municipality level (LAU, Local Administrative Units of territorial units nomenclature), basing on the Classification of the municipalities defined in the National Inner Areas Strategy (Strategia Nazionale per le Aree Interne, SNAI).

Well-being is a broad concept, including both economic status and quality of life, e.g. life expectancy at birth or schooling years (UNDP, 2018). The lack of updated indicators at municipal level, however, allows to address the topic only by means of income data. The latter, in turn, evoke the concepts of poverty and inequality, which are strong determinants of well-being (Stiglitz, 2012).

Inner areas are territorial contexts characterized by a significant distance from the main supply poles of essential services (health, education, mobility) (Barca et al., 2014). They have recently been the subject of specific policies, which find their main reference in the SNAI (DPS, 2014), founded on Place-Based approach (Barca, 2009).

Territorial dimension plays a central role in explaining quality of life, which depends on the presence/absence of environmental, social and economic resources and public and private services (Stiglitz, 2012).

Municipal administrations contribute, together with Regions and Ministries, to implement optimal levels of "essential services of citizenship" in inner areas (DPS, 2014); moreover, their expenditures have a multisectoral valence (i.e. social assistance, environmental protection, and so on), while Regions spend their budget mostly for health care. In this scenario, the paper aims at answering to the following research questions (RQs):

a) Is a convergence taking place at municipality level, with respect to the per capita incomes?
b) Are there any spatial spillover effects regarding the distribution of income variations?
c) What are, in the end, the effects of local administrations policies on income level in Italian inner areas?

2. Methodology

Our study takes into account: a) the per capita incomes at municipality level, and b) the budgets of Municipalities, in particular for what concerns the expenditures part, as subdivided in the so-called "Missions". We have considered the following expenditure categories: 1) Education (EDU); 2) Cultural heritage (HER); 3) Youth, Sports and Leisure (YSL); 4) Tourism (TOU); 5) Planning and Housing (PAH); 6) Environment protection (ENV); 7) Social policies (SOC); 8) Economic development (ECO); 9) Agriculture (AGR). Individual income data come from Revenue Agency, while per capita budget data have been gathered on AIDA database of Bureau van Dijk. Reference years are 2008 and 2016; in this latter case, all of the values have been converted to 2008 constant price values. We have been compelled to remove from our analyses the two biggest Italian islands, Sardinia and Sicily, owing to problems in the linking of the municipality codes both in the two databases, and in the two different years considered. The total number of continental Italy municipalities amounts to 7129, 3448 of which classified as "Inner areas", and 3681 as "Centres".

ISBN: 978-88-5495-135-8

Following the RQs presented at the end of Section 1, we have at first analysed the per capita incomes convergence, on the basis of a simple conditional model. Then, we studied the spatial distribution of income variations between 2008 and 2016; finally, we have built, for municipalities belonging to inner areas, a regression model linking the mean per capita incomes with the per capita expenditures. In particular, the convergence model is the following:

$$\ln\left(\frac{Y_{2016}}{Y_{2008}}\right) = SNAI + NUTS1 + \ln Y_{2008} + \ln EXP_{2008} + \varepsilon \quad (1)$$

where Y_t is the per capita income at year t, $SNAI$ is two-level factor classifying Italian municipalities into Inner Areas (IA) and Centres (CC), $NUTS1$ is another factor identifying the various macro-regions: North-West (NW), North-East (NE), Centre (CE), South (SO), and EXP_{2008} are the total public expenditures at year 2008.

Exploratory Spatial Data Analyis (ESDA) has been performed in order to highlight the presence of local clusters (hot spots) suggesting the presence of spatial spillovers. The variable taken into account has been the dependent one of model (1). Spatial contiguity has been introduced by means of the classic row-normalized matrix $\boldsymbol{W} = [w_{ij}]$, with $w_{ij} = 1/n_i$ if two municipalities i and j are contiguous, and $w_{ij} = 0$ otherwise, where n_i is the number of neighbours of zone i. We calculated the well-known Moran's index of spatial autocorrelation, I, at a global scale, and the Local Moran (LM) index (Anselin, 1995) at municipal level, in order to highlight the municipalities contributing in a significant way to global spatial autocorrelation.

As a final step we have estimated, with respect to IA municipalities, a random effects panel regression model (Wooldridge, 2010) relating per capita incomes with the public expenditure categories. Three dummy variables have also been added in matrix $\boldsymbol{D}$, for the territorial partitions (with reference macro-region being NW), giving:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{D\gamma} + \boldsymbol{\varepsilon} \quad (2)$$

where: $\boldsymbol{y} = (y_{11}\ y_{12} \cdots y_{1T}\ y_{21} \cdots y_{2T}\ \cdots\ y_{n1} \cdots y_{nT})'$; $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_k]$ is the independent variables matrix, with $\boldsymbol{x}_h, h = 1, \cdots, k$ having the same structure as $\boldsymbol{y}$; $\boldsymbol{D} = [\boldsymbol{d}_1\ \boldsymbol{d}_2\ \cdots\ \boldsymbol{d}_m]$ is the dummy variables matrix, with $\boldsymbol{d}_l, l = 1, \cdots, m$ having the same structure as $\boldsymbol{y}$; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are parameter vectors of dimension, respectively, $(k \times 1)$ and $(m \times 1)$; and the $(nT \times 1)$ error vector $\boldsymbol{\varepsilon}$ is modelled as $\varepsilon_{it} = \alpha_i + u_{it}$, with α_i being the random individual component (constant across time periods) and u_{it} is the residual noise term.

3. Results and discussion

The mean and median per capita incomes are lower in IA with respect to the CC (Table 1). It is important to stress the fact that in the period considered IA have recorded a 0.95% increase in mean per capita incomes, while CC have experimented a decrease in real incomes (-0.77%).

Table 1: Univariate statistics for the per capita incomes in years 2008 and 2016

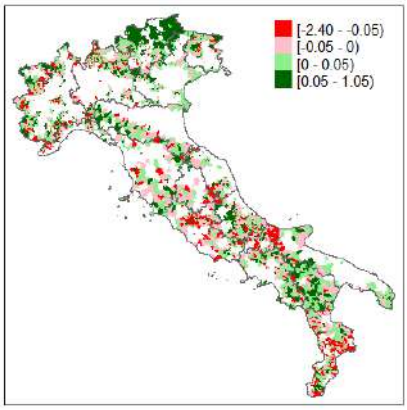
Year	SNAI	Min	Q_1	Q_2	Mean	Q_3	Max	SD	CV
2008	**IA**	2366.82	8441.20	10772.86	10627.99	12518.37	21869.17	2610.23	24.56
	CC	4699.28	11590.22	13400.92	13067.92	14865.83	29471.96	2892.58	22.13
2016	**IA**	2402.18	8427.60	10763.42	10728.85	12665.25	32315.17	2787.46	25.98
	CC	1237.21	11472.06	13297.55	12917.53	14618.56	28507.43	2823.23	21.86

Figures 1a. and 1b. show the logarithms of the ratio between per capita incomes at years 2016 and 2008, respectively for inner areas and centres. In particular, 1840 IA municipalities show an increasing income (53.36% of the total), while only 1378 CC municipalities achieve the same record (37.44%). This suggests a convergence process between IA and CC.

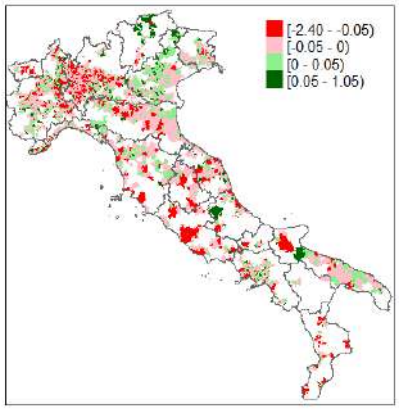
The results for conditional β-convergence model (1) are in Table 2. All the variables are significant. β coefficient presents a negative sign, again highlighting convergence. The negative sign of dummy variable CC indicates a higher increase in incomes in IA with respect to CC.

ISBN: 978-88-5495-135-8

Figure 1: Cartograms reporting the values of $\ln(Y_{2016}/Y_{2008})$ in Italian municipalities.



1a. Inner areas



1b. Centres

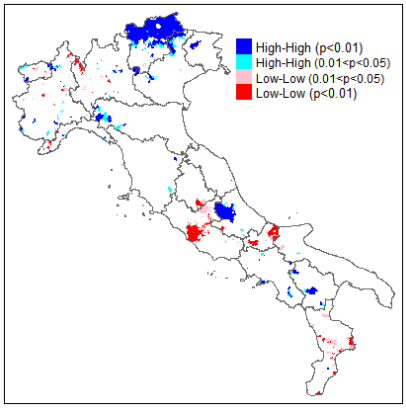
Regarding $NUTS1$, only NE has positive sign, and this implies a differentiated convergence process among macro-regions, being faster in Northern than in Central and Southern Italy. The last, important observation is that total public expenditures at year 2008 has had a positive effect on per capita income growth.

Table 2: Parametric estimation results for model (1).

Parameter	Estimate	Std. Err.	t value	Pr(>|t|)	Parameter	Estimate	Std. Err.	t value	Pr(>|t|)
Intercept	0.4563	0.0522	10.48	< 0.001	**CC**	-0.0078	0.0020	-3.84	< 0.001
NE	0.0280	0.0024	11.73	< 0.001	**$\ln Y_{2008}$**	-0.0541	0.0054	-10.09	< 0.001
CE	-0.0170	0.0029	-5.60	< 0.001	**$\ln EXP_{2008}$**	0.0077	0.0014	5.33	< 0.001
SO	-0.0183	0.0033	-5.49	< 0.001					

As to the spatial distribution of income (log-)variations between 2008 and 2016, Moran's index is positive ($I = 0.151$), and highly significant ($p < 0.001$): this result suggests the presence of spatial spillovers. LM index has then been calculated for all of the considered municipalities; in the following significance map (Figure 2) we report only the High-High and Low-Low combinations, for $p < 0.01$ (dark colours) and $0.01 < p < 0.05$ (light colours).

Figure 2: LM significance map (H-H and L-L combinations) for $\ln(Y_{2016}/Y_{2008})$.



As an example, a blue zone indicates that both the considered municipality and its neighbours average have grown in the period 2008-2016, and that this raise is significant at 0.01 level. Local clusters show how economies with different development degrees converge only in presence of not too dissimilar structural conditions, e.g. technological level or propensity to save (Barro and Sala-i-Martin 1991).

The final step has been parametric estimation of panel random model (2); before that, we calculated the univariate statistics reported in Table 3.

It is easily recognizable the difference both in mean and in relative variability amounts.

Table 3: Statistics for the per capita public expenditures in Italian IA (average 2008-2016).

	EDU_a	HER_a	YSL_a	TOU_a	PAH_a	ENV_a	SOC_a	DEV_a	AGR_a
Mean	268.9	40.02	44.57	39.24	325.07	255.09	108.08	20.35	5.97
CV	169.23	340.65	300.63	472.91	143.32	101.33	131.13	388.60	609.88

Parametric estimation results for panel regression model (2) are shown in Table 4.

Table 4: Parametric estimation results for model (2)

Effects:	**Var**	**Std.Dev**	**Share**	
idiosyncratic	0.0031	0.0556	0.122	
individual	0.0223	0.1495	0.878	
theta:	0.7439			
Coefficients:	**Estimate**	**Std. Error**	**z-value**	**Pr(>|z|)**
Intercept	9.4995	0.0143	663.43	< 0.001
NE	0.0604	0.0083	7.27	< 0.001
CE	-0.1542	0.0085	-18.22	< 0.001
SO	-0.4162	0.0069	-60.54	< 0.001
EDU	-0.0211	0.0020	-10.33	< 0.001
HER	0.0026	0.0008	3.20	0.001
YSL	0.0027	0.0008	3.48	0.001
TOU	0.0071	0.0008	9.17	< 0.001
PAH	-0.0075	0.0014	-5.50	< 0.001
ENV	-0.0014	0.0013	-1.08	0.279
SOC	0.0051	0.0014	3.71	< 0.001
DEV	0.0030	0.0009	3.32	0.001
AGR	-0.0011	0.0012	-0.90	0.366
R-Squared:	0.8476	**Adj. R-Squared:**	0.8474	

Adjusted R-squared accounts for about 85% of the observed variability; only two variables are not significant, i.e. Environment protection and Agriculture enhancement expenditures. The territorial effect has been caught by means of the dummy variables relating to macro-regions: in this regard, it is possible to see how North-Eastern IA municipalities have experimented a much higher per capita income growth with respect to North-Western ones, while Central and Southern IA municipalities have had a negative growth when compared with NO ones.

Other two variables show negative sign (*EDU* and *PAH*): this means that, across times, an increase in one of these categories has led, in mean, to a downward effect on per capita incomes variation. Conversely, Tourism and Social policies play the most important positive role.

4. Conclusions

A process of convergence, conditioned to economic policies and to territorial features, has emerged from the analysis. Spatial differences are important, with IA having better performances than CC. In detail, with reference to IA, per capita incomes have raised in Northern macro-regions, and not all of the local administrations policies have had a positive effect. It is needed a change in the distribution of expenditures between categories, in order to achieve a bigger impact on per capita incomes.

References

Anselin, L. (1995). Local Indicators of Spatial Association – LISA. *Geographical Analysis*, **27**(2), pp. 93-115.

Barca, F., (2009). *An agenda for a reformed Cohesion Policy. A place-based approach to meeting European Union challenges and expectations.* Bruxelles (BE).

Barca F., Casavola P., Lucatelli S. (2014), *Strategia nazionale per le Aree interne: definizione, obiettivi, strumenti e governance*. Materiali Uval, Roma (IT).

Barro R. J. and Sala-i-Martin X. (1991), Convergence across States and Regions, *Brookings Papers on Economic Activity*, **22**(1), pp. 107-182.

DPS (2014). *Strategia Nazionale per le Aree Interne. Accordo di Partenariato 2014-2020.*

Stiglitz J.E. (2012). *The Price of Inequality: How Today's Divided Society Endangers Our Future*. Penguin, London (UK).

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge (MA).

UNDP (2018). *Human Development Indices and Indicators*. New York (USA).

Intelligent systems to support patients

Vito Santarcangelo[a], Emilio Massa[a], Diego Carmine Sinitò[a], Giuseppe Scavone[b]
[a] iInformatica Srls (PMI innovativa), Trapani, Italy;
[b] Centro Rham S.r.l., Matera, Italy;

1. Introduction

The research focused on innovative software development and the creation of new working flows can bring great benefits in qualitative and economic terms to companies operating in the medical sector. In this work we present several innovative tools implemented by the Rham medical center to improve the quality of the service offered to its patients and its personnel. The research has several objectives: the improvement of novel diagnostic techniques related to orthoptics, the development of new therapeutic tools related to gaming, the administrative and management aspects of the medical center.

2. Innovative tools

The research focused on two fronts: the need to innovate diagnostic tools that had some limitations highlighted by the doctors of the medical center and the need to create new therapeutic tools modelled on the needs of the patients of the medical center.

Regarding the enhancement of diagnostic tools, our studies led us to the development of a system for improving awareness of visual coupling and monocular response. After an analysis of the classic tools of the orthoptics area, such as the test of the four images of Purkinje [1]: which assesses the luminous reflection and therefore if there are or not deviations of the visual axis through the use of a central light stimulus, we started to rethink the diagnostic tools in order to make them more autonomous and accurate.

These diagnostic systems are characterized with a strong manual interaction by the operator and a considerable limitation in systematically and digitally recording and monitoring. The objective of our work is essentially to provide a highly innovative system for improving awareness of visual engagement and visual well-being. The novel system provides an HMI interface for the user, usable from a common smartphone or an input terminal, a web server with a database and a semantic system, one or more cameras and a computer vision module.

Thanks to the computer vision module is possible to carry out the analysis of the awareness of the visual coupling considering the engagement times (calculated considering the gaze response, through eyetracking, from a stimulation on the terminal) and the distance to which the subject is placed with respect to the terminal. From this analysis it is also possible to detect the possible delta of the anomalous ocular focus with respect to the expected one.

It is possible to define the VCA parameter (visual coupling awareness) according to the coupling and focusing times, appropriately weighted for the distance. The VCA parameter used for determination of the delta of the focus obtained with an appropriate comparison with an ideal reference (calculated by an expert system from the semantic parameters related to the physiology of the subject). The delta is fundamental to highlight the presence of some related diseases and can be examined as a trend over time.

Further research has instead led to the development of new therapeutic tools in the form of games, with very specific requests from the doctors of the medical center, so to satisfy the needs of patients. Gaming represents today a new frontier for training and learning. In this perspective, the medical center has devised and designed innovative tools (traditional in form but not in method and content) to support clients, especially those affected by Alzheimer and Learning Disability. For the Alzheimer, a souvenir box was created, with the use of cards that represent appropriate situations to implement the space-time sequences of common actions. For Learning

ISBN: 978-88-5495-135-8

Disability, a Carhambola game was created, set in a scholastic context and based on the lexical learning of adjectives according to the paradigm of the AIN Thesaurus, to learn lexical polarity and semantics (in line with the knowledge bases of artificial intelligence systems). The two products represent a real revolution for traditional learning tools through the use of innovative gaming logic.

3. Patient management system

Research activities was focused on improving internal company processes related to the management of patients' treatment plans. An algorithm and related information system was developed to manage and monitor the treatment plans over time, considering all requests from users and their workforce, with the aim of pursuing greater service quality, user satisfaction and staff motivation. This method will be expressed in full in the following paragraph.

Expert systems (ESs) are technological applications that belong to the branch of artificial intelligence and can be identified as software programs that attempt to reproduce the performance of one or more experts in a given domain. An expert system is able to autonomously implement inference procedures (with an inductive or deductive process, a conclusion is reached following the analysis of a series of facts or circumstances) suitable for solving a problem.

The ES are composed of three basic elements:

- The knowledge base: the database in which the production rules are stored, i.e. the deductive rules that allow the system to follow a logical reasoning on a particular branch of knowledge. It is often referred to as Knowledge Base (KB);
- The inferential engine: the software component that processes the knowledge contained in the knowledge base, interprets the user's need and provides a solution to the problem. The reasoning is based on a set of deductive rules of the type IF condition THAN action;
- The user interface: the intermediate element between the user and the inference engine. In most cases it is an input/output interface in which the user enters the data of the problem to be solved and displays the results of the processing.

In this specific case, the use of an ES was used to address the problem of managing the calendar of patient appointments. Appointment management is a major task for the center, a correct schedule of doctors makes it possible to make the best use of the workforce and to maximize the number of services provided by the center, which implies a significant impact both on the quality of the center from the point of view of doctors and patients but also allows you to maximize revenue. This task was always carried out by the administrative staff of the structure, highlighting two incomplete aspects: the first is the large amount of data that the staff must examine to schedule the calendar of appointments and the second is the impartial choice on the assignment of shifts. For these reasons the aim of this work was the design of an automatic method for the optimization and maximization of the therapies administered by the Rham medical center to its patients.

Each patient in the center is associated with a therapeutic plan which consists of a document describing the therapies administered to the patient, the doctors who follow him and the frequency of the therapies. Therapeutic plans are therefore the starting point for planning the work carried out by the doctors of the center. Often the patient can also express preferences about the hours or days in which the different therapies will be administered, this happens very often with school-age patients who for example cannot receive therapies in the morning hours as they are busy with the school.

Initially we proceeded with the formalization of the KB where all the information regarding the treatment plans, the patients and their availability, the doctors and their work shifts are stored. The main entity is the therapeutic plan (TP), which includes: a patient, a start and an end date, a set of therapies and the number of total benefits (TB). Each therapy within the plan is coded

ISBN: 978-88-5495-135-8

by means of a product code, the duration expressed in minutes, the weekly frequency. The doctors in the field who are in charge of the patient are associated with each therapy. Two supporting entities are those that store information regarding the availability of patients and doctors.

In summary it is possible to express a therapeutic plan P through the following formulation:

$$P(p, PT, d, OP_1(f,m) \cup \ldots \cup OP_n(f,m))$$

where:

- p: is the patient;
- PT: the number of total benefits to be paid;
- d: the cycle end date;
- $OP_1(f,m) \cup \ldots \cup OP_n(f,m)$: the union of the n individual therapies with relative frequency f and duration m.

The inference engine is built based on a priority assignment system. Therapies with higher priorities are always scheduled, thereafter we proceed with the scheduling of therapies as long as doctors are available. The priority value of an $OP_n(f,m)$ therapy, where n is the index of the therapy we are taking in exam, will be very low if the expected number of therapies to be delivered is close to that of the therapies provided, it will be 0 if the two values coincide. The formula changes getting close to the end of the cycle.

The algorithm involves the following main steps:

a) Selection of all active therapeutic plans (date of examination < date end of cycle);
b) Removal of treatment plans with number of services >= PT;
c) Priority calculation for each therapy $OP_n(f,m)$;
d) Selection of the activity $OP_n(f,m)$ with the highest priority;
e) Start a cyclic search in the configured time slots (center opening times). Looking for a doctor available from those assigned to the patient. If there are no doctors available, we go to the next time slot, increasing the start time of the appointment by 15 minutes;
f) If a doctor is available we check if the user has some time preferences and if we can match one of them.
g) Once the scanning of all the activities is completed, we move on to the next day of the week and we run the algorithm again;

The run of the algorithm usually takes place within the starting date that is a Monday and the ending date that is a Friday in order to schedule a weekly calendar with all the appointments of medical center. After each run of the algorithm we are able to calculate a performance index. Using the information stored in the KB we can define: *Medical center's load of work, Employees' Load of work, Patient satisfaction*.

$$\text{Medical center's load of work} = \frac{\text{theoretical maximum number of appointments of a week}}{\text{number of appointments scheduled from the run of the algorithm}}$$

$$\text{Employees load of work} = \sum_{i=1}^{n} \frac{\text{number of contracted working hours}}{\textit{number of the worked hours in the proposed scheduling}} \text{ for each employee}$$

$$\text{Patient satisfaction} = \frac{\text{number of users preference expressed}}{\text{number of users preference matched}}$$

$$\text{Algorithm performance index} = \text{Medical center's load of work} * \text{Employees load of work} * \text{Patient satisfaction}$$

ISBN: 978-88-5495-135-8

Since our objective is reach a high quality standard we can run the algorithm more times in order to reach a fixed performance index until we get the perfect scheduling of the appointments.

4. Conclusions

The use of innovative technologies has allowed the medical center to provide a service modelled on the different needs of each patient while maintaining high standards of quality and professionalism. Thanks to the novel algorithm the medical center is able to schedule a calendar for the therapies that takes into account the patients needs in an easy way, whit a low human support. Moreover, the research aimed at defining new therapeutic protocols was positively perceived by patients who got benefits by the less invasive nature of some diagnostic tools or the dynamism of new therapeutic tools, such as the use of some games. This new kind of therapy introduce to new future investigations, extending what has been done to other disorders and therapies.

References

Hewitt D. Crane, Carroll M. Steele, (1985). Generation-V dual-Purkinje-image eyetracker, *Applied Optics*, **24**(4), pp. 527-537.

Scavone, G. (2018). *Metodo e sistema innovativo per la gestione del piano terapeutico degli assistiti*, UIBM, 102018000021196.

Scavone, G., Giugliano, F., Santarcangelo, V. (2019). *Sistema ad alta innovazione e metodo per il miglioramento della consapevolezza dell'aggancio visivo e della risposta monoculare*, UIBM, 102019000009996.

Santarcangelo, V. (2018). *Sistema e metodo intelligente per la somministrazione di un gioco da tavolo*, UIBM, 102018000005400.

Notarnicola, A., Scavone, G. (2019). La blockchain come strumento a supporto della CSR. *Quality & Engineering*, **3**(1).

Food quality perception in children: a comparison between Bayesian Network and Structural Equation Modelling

Anna Simonetto[a], Silvia Golia [b], Buirma Malo[a], Gianni Gilioli[a],
[a] Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy
[b] Department of Economics and Management, University of Brescia, Brescia, Italy

1. Introduction

Recent surveys indicate that poor eating habits and unhealthy lifestyles are growing in young people (Al-Nakeeb *et al.*, 2015). The spread of overweight and obesity among young people is particularly worrying if we consider the future socio-health consequences related to the expected increase in chronic-degenerative diseases. School age is a critical phase for the acquisition and consolidation of correct eating habits, therefore widespread and effective educational interventions are among the most promising actions preventing the effects of an unhealthy eating habits, as widely demonstrated in scientific literature (Holsten *et al.*, 2012). In a more global perspective, food education should take into account all the several components that define the food quality, not only the nutritional value. Aspects such as food safety, sustainability, social responsibility, food security and organoleptic properties are important drives of food consumption choices. It is appropriate that children are aware of the complexity that characterizes the systems of production and transformation of food and what are the implications of their food choices, both for themselves and for the environment and the society in which they live. For this purpose, the education to multidimensional food quality is of particular interest for political institutions, educational agencies and health professionals to promote the health and quality of life of today's and tomorrow's citizens.

The objective of this study is to compare two different approaches, based on Structural Equation Models (SEM) and Bayesian Networks (BN) for the assessment of the multidimensional concept of food quality in students of first grade secondary school (11-14 years old).

2. Methodologies

The research team developed a questionnaire to investigate the concept of food quality in children. This tool is composed by four main parts: i) socio-demographic data, ii) qualitative description of the idea of food quality, iii) a match-question in which students are asked to match the six dimensions of food quality to their correct definitions (listed randomly), iv) a set of true/false questions investigating the knowledge on several aspects of food quality. The fourth part of the questionnaire consists in 60 statements regarding the six dimension of food quality (10 items for each dimension) considered in this analysis: Food Safety (FSaf), Food Security (FSec), Nutritional Value (Nutr), Organoleptic Quality (OrgQ), Environmental Sustainability (Sust) and Corporate Social Responsibility (SocR). For each statement students should indicate if it is ‘True’, ‘False’, or he/she doesn’t know the answer (Don’t know). The questionnaire was preliminary validated through its administration to 20 children randomly chosen among the participants of a presentation event of the university's research activities. The validated version of the questionnaire was administered to 696 students from six schools in the Province of Brescia (Italy) from November 2018 to January 2019. The ‘Don't know’ response is codified as wrong answer since our aim was to investigate the presence of an adequate knowledge with respect to the question. To investigate the six dimensions of food quality we analyse the results of the fourth part of the questionnaire following two different approaches.

The first approach, based on the SEM (Bollen, 1989), provides for the simultaneous estimation of latent constructs (in this case the six food quality dimensions) and the relationship between them. The second approach the BN is applied to investigate the relationships between the six constructs, estimated through Rasch Model (RM).

Structural Equation Model Structural equation models constitute a family of statistical methods representing one of the most widespread methodologies in the analysis of behavioural data and/or perceptions since they allow studying existing interrelationships between variables that cannot be directly measured, called latent variables or factors. Information contained in the interrelationships between many variables (indicators or observed variables) can be traced back to a smaller set of variables (latent variables), making it easier to identify a structure underlying the data. SEMs allow testing complex models involving both direct and indirect effects. Applying a SEM, it is possible to examine whether a conceptual model, assuming relationships between a set of latent variables and a set of indicators, is consistent with the empirical data. SEMs include the structural model, a set of equations defining links between the latent variables, and the measurement model, a set of equations that specify the relationships between the latent variables and the indicators. A useful tool for the interpretation of SEMs is the path diagram, the graphical representation of a system of simultaneous equations, which shows the relationship between all the variables, including disturbance factors and errors. All SEM approaches involve the same basic sequence through which the analysis is carried out on the explanatory models of the hypothetical causal links at the base of the observed data. This iterative sequence consists of: i) specification of the conceptual model; ii) estimation (of the structural parameters) of the model; iii) evaluation of the model; iv) modification of the model.

Rasch model and Bayesian Networks The RM is a measurement model which converts raw scores into linear and reproducible measurements; if the data fit the model, the obtained measures are objective and expressed in logit. Its distinguishing characteristics are separable person and item parameters, sufficient statistics for the parameters and conjoint additivity, whereas the prerequisites are unidimensionality and local independence (Bond and Fox, 2015). RM assumes that the probability of response to an item given by a person is only governed by the difficulty of the item and the ability of the person, which in this study represents the degree of knowledge of the food dimension.

The BNs belong to the class of probabilistic networks, which explicit, through a graph, the interactions among a set of variables represented as nodes of the graph (Kjærulff and Madsen, 2013). A BN is composed by the pair (G, P), where G is the Directed Acyclic Graph (DAG), and P is a probability distribution which factorizes according to G. G is composed by a set of nodes V, which correspond to a set of random variables X_V indexed by V, and a set E of directed links between pairs of nodes in V. A BN can be used as a descriptive tool of the dependence/independence relationships among the variables, as well as a predictive model.

3. Results

The analyses were performed using Mplus 7 to implement the SEM approach, Winsteps 3.75 to create the Rasch measures and the R package `bnlearn` to identify the BN.

The full conceptual model tested trough the SEM approach is represented in Figure 1-a. Each latent variable is described by a set of 10 items (S01-S60) and it can be correlated to the other dimensions. We performed several SEM exploratory factor analyses modifying the set of indicators to test the hypothesized relationships of the full conceptual model, but the models never converged. The errors were mainly associated with the first dimension of analysis: Food Safety. This is the dimension least understood by scholars, with the lowest percentage of correct answers (13.9%). This led to the exclusion of this dimension and the related indicators from the

ISBN: 978-88-5495-135-8

analysis. The final version of the estimated SEM is shown in Figure 1-b. Fit indexes of the SEM are quite good, the parsimony-adjusted index RMSEA is very low (0.02), and the Comparative Fit Index (CFI), comparing the fit of a target model to the fit of an independent, or null, model, is high (0.923).

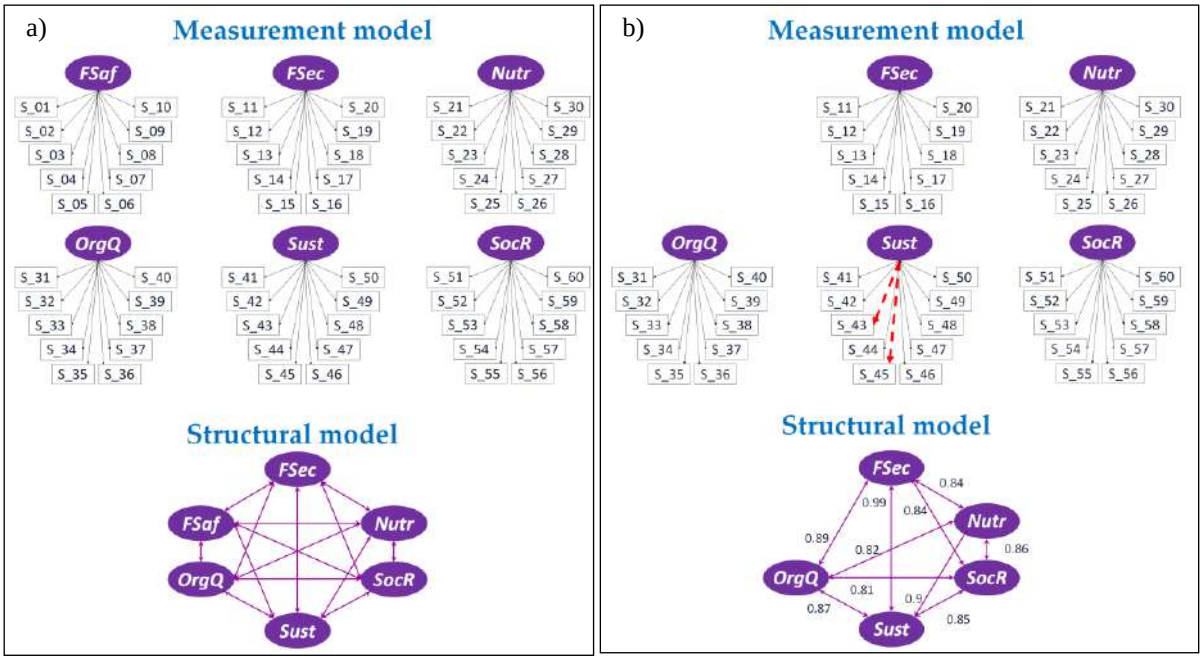


Figure 1: a) path diagram of the full conceptual model of SEM, b) path diagram of the final version of the estimated SEM (red dashed arrows represents not significative estimates)

The results of the SEM show a high linear correlation between all the dimensions investigated, always greater than 0.8 (Table 1). In particular, Food Security and Sustainability are characterised by an almost perfect positive correlation ($corr(FSec, Sust) = 0.99$). Social Responsibility and Organoleptic quality are the two latent variable with a less intense correlation link, though still high ($corr(SocR,\ OrgQ) = 0.81$).

Table 1: Variance-covariance matrices of the SEM (standardized latent variables)

	FSec	Nutr	SocR	Sust	OrgQ
FSec	1				
Nutr	0.836	1			
SocR	0.841	0.861	1		
Sust	0.985	0.900	0.845	1	
OrgQ	0.889	0.816	0.806	0.869	1

Regarding the second approach, we have obtained reliable measures of all the six dimensions through the RM, as shown in Table 2, which reports the Item Reliability Index (IRI), the Greatest Lower Bound (GLB) and the mean degree of knowledge of each dimension, with standard deviations in brackets. The IRI is used to verify the hierarchy of the items, whereas the GLB is a reliability index alternative to the Cronbach's Alpha, which in many situations underestimates the reliability of a test. Analysing the average measures of the six dimensions, the ones of Nutritional Value, Organoleptic Quality, Food Security and Social Responsibility are higher than zero, so we can conclude that, overall, the students involved in the survey have a high degree of knowledge of these food dimensions.

Table 2: Reliability indices and mean degree of knowledge of each dimension

Dimension	IRI	GLB	Mean (sd)
FSaf	0.99	0.71	-0.21 (1.01)
FSec	0.99	0.78	0.45 (1.22)
Nutr	0.99	0.76	0.82 (1.22)
OrgQ	0.98	0.77	0.62 (1.10)
Sust	0.99	0.69	-0.12 (1.04)
SocR	0.98	0.91	1.04 (1.42)

In order to compare the structure of relations between the variables obtained applying SEM and BN, we have discarded the dimension of the Food Safety. Figure 2 reports the preliminary DAG underlined the BN obtained applying the score based algorithm Hill Climbing with BIC as the score.

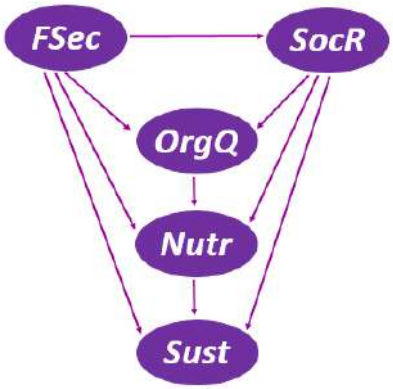


Figure 2: Preliminary DAG for the considered measures of food quality

The DAG is not fully connected, as the path diagram of the SEM; only Food Security, Nutritional Value and Social Responsibility are connected with all the other four dimensions. Moreover, from the DAG in Figure 2 we can read the conditional independence relationships between the variables; for example Environmental Sustainability is independent from Organoleptic Quality given Food Security, Social Responsibility and Nutritional Value.

References

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. John Wiley and Sons, Inc., New York.

Bond, T.G., Fox, C.M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Holsten, J.E., Deatrick, J.A, Kumanyika, S., Pinto-Martin, J., Compher, C.W. (2012). Children's food choice process in the home environment. A qualitative descriptive study. *Appetite*, **58**(1), pp. 64–73.

Kjærulff, U.B., Madsen, A.L. (2013). *Bayesian networks and influence diagrams: a guide to construction and analysis*. Springer, New York.

Al-Nakeeb Y., Lyons M., Dodd L.J., Al-Nuaim A. (2015) An investigation into the lifestyle, health habits and risk factors of young adults. *International Journal of Environmental Research and Public Health*, **12**(4), pp. 4380–4394.

The studyholism comprehensive model: towards a bayesian reanalysis

Federico M. Stefanini[a], Yura Loscalzo[b]
[a]Department of Statistics, Computer Science, Applications, University of Florence, Italy;
[b]Department of Health Sciences, School of Psychology, University of Florence, Italy;

1. Introduction

In the psychological literature, scholars recently argued the existence of a potential new clinical condition associated to problematic overstudying, namely Study Addiction (Atroszko et al., 2015) or Studyholism, i.e. obsession toward study (Loscalzo and Giannini, 2017). Even if the constructs of Studyholism and Study Addiction refer to the same problem behavior, there is disagreement concerning their theory and operationalization (Atroszko et al., 2015; Griffiths et al., 2018; Loscalzo and Giannini, 2018a,b). Atroszko et al. (2015) defined problematic overstudying as a behavioral addiction characterized by the seven core components of substance addictions (i.e., salience, tolerance, mood modification, relapse, withdrawal, conflict, and problems).

Loscalzo and Giannini (2017) instead went beyond the addiction model, and they defined Studyholism as an Obsessive-Compulsive related Disorder (OCD-related disorder) made up by two components (i.e., obsessive-compulsive symptoms and high or low study engagement), which led to the proposal of two subtypes of Studyholics: Engaged Studyholics (students with high levels of both Studyholism and Study Engagement) and Disengaged Studyholics (students with high levels of Studyholism and low levels of Study Engagement). Moreover, Loscalzo and Giannini (2017) specified that not all the students with high time and energy investment in study (or Heavy Study Investors, HSIs) are studyholics. By crossing the high/low levels of Studyholism and Study Engagement, four kinds of student arise (three of which are HSIs): Disengaged Studyholics, Engaged Studyholics, Engaged students, and Detached students. The Detached student, which is characterized by low levels of both Studyholism and Study Engagement, is not an HSI. Though, he/she is a negative type of student anyway, as he/she is detached from one of his/her central daily activities. The Engaged student is instead the most positive kind of student, since he/she has low levels of Studyholism and high levels of Study Engagement (i.e., a positive attitude toward studying, such as intrinsic motivation toward studying). A comprehensive model was also proposed by Loscalzo and Giannini (2017) to include possible antecedents and outcomes of Studyholism. This model distinguishes between individual (e.g., perfectionism) and situational (e.g., overstudy climate) antecedents, as well as between individual (e.g., physical and psychological impairment) and situational (e.g., aggressive behaviors at school) outcomes. Further work of these authors led to a path analysis model (Loscalzo and Giannini, 2019)[1] where they found that, in line with their OCD-related model, the variable worry is a strong predictor of Studyholism.

There are few papers about problematic overstudying and there is disagreement concerning the proposed theoretical construct. Also, all the studies on this topic published since now, e.g. Atroszko et al. (2015) and Loscalzo and Giannini (2019), analyzed ordinal variables like metric-normal random variables, a widespread practice in psychology (Liddell and Kruschke, 2018). Hence, it is of paramount importance to assess the effect of less strong statistical assumptions,

[1]The dataset considered in this work has been discussed in the following work (2019): https://www.frontiersin.org/articles/10.3389/fpsyt.2019.00489/full

in particular by respecting the ordinal nature of qualitative variables. In this paper, a Bayesian ordinal probit model is developed for the key variable Studyholism considered in Loscalzo and Giannini (2019). Model fitting is performed by Markov Chain Monte Carlo (MCMC) simulation. The discussion of results from the Bayesian reanalysis emphasizes distinctive steps and achievements as compared to previous work.

2. A Bayesian ordinal regression model

Let Y_i, $i = 1, 2, \ldots$, be random variables obtained by adding all ordinal questionnaire items representing the i^{th} feature of the underlying theoretical construct. The sample space $\Omega_i = \{1, 2, \ldots, m_i\}$ for Y_i is made by natural numbers (after rescaling). Let $\{x_{i,k}\}$ be a collection of antecedents explanatory variables for Y_i, and $\{\beta_{i,k}\}$ a collection of model parameters, than $y^*_{i,r} = \sum_k x_{i,k,r}\beta_{i,k} + \epsilon_{i,k,r},\ r = 1, 2, \ldots, n$ (sample size) are latent variables of an ordinal probit model where $\epsilon_{i,k,r} \sim N(0, 1)$ and $-\infty = c_{i,0} < c_{i,1} < \ldots < c_{i,j} < c_{i,j+1} < \ldots < c_{i,m_i-1} < c_{i,m_i} = \infty$ is a sequence of unknown but sorted cutpoints with index $j = 1, 2, \ldots, m_i$. An observed ordinal variable is than defined as $y_{i,r} = j$ iff $c_{i,j-1} < y^*_{i,r} \leq c_{i,j}$. In the above regression of $y^*_{i,r}$, orthogonal polynomials were defined over the sample space of each ordinal explanatory variable and scaled to null mean and unit variance, although only linear terms were retained into the model. Sum-to-zero constrasts were adopted for nominal variables instead. Beliefs about cutpoints were defined through a Dirichlet distribution after eliciting the expected number of virtual observations belonging to each ordinal class (below equal to 2 for each ordinal class), so that a realization $(\theta_1, \ldots, \theta_{m_i})$ from the Dirichlet implicitly defined the location of cutpoints by the inverse of the normal cumulative distribution function, e.g. $c_j = \phi^{-1}(\sum_{s=1}^{j} \theta_s)$. The initial distribution for betas was defined as flat-non informative to better match estimates obtained in the cited (frequentist) paper, but informative alternatives could be used instead, for example a multivariate normal distribution for the vector of betas.

3. Results and discussion

A model for Studyholism (SH_TOT_ter, 12 ordered classes) was fitted to 1958 Italian college students aged between 18 and 60 years and quite heterogeneous about their year and major of study (details in cited work). The explanatory variables were (variable names among brackets): Study-related Perfectionism (SrPS_TOT.L), Perfectionistic Strivings (SAPS6_Standard.L), Perfectionistic Concerns (SAPS6_Discrepancy.L), Worry (PSWQ_TOT.L), Parents Overstudy Climate (OCS_Gen.L), Teachers Overstudy Climate - Overt Comments (OCS_InsComm.L), Teachers Overstudy Climate - Hard Study (OCS_InsHardStudy.L), Area of Study - technological, social, humanistic, medical, scientific (Area_Study). Computations were performed with the R packages *rstan, rstanarm, loo, ggmcmc*. [2]

In the first step of the analysis, the starting model defined by the theoretical construct was fitted by maximum likelihood with ordinal explanatory variables represented by the linear term (L) of orthogonal polynomials. Several further models were considered adding quadratic and cubic model terms but the starting model resulted in the smallest BIC value. In the second step, the starting model was fitted by MCMC simulation, with 3 chains where warmup was 1000 iterations after which 5000 realizations from the final (posterior) distribution were retained without thinning. Output diagnostics were calculated before obtaining inferences. All trace plots did not suggest problems of mixing. The Rhat statistic calculated from the three chains resulted equal to 1.0 for all model parameters. One-dimensional cross-validated Leave One

[2] https://cran.r-project.org/web/packages/

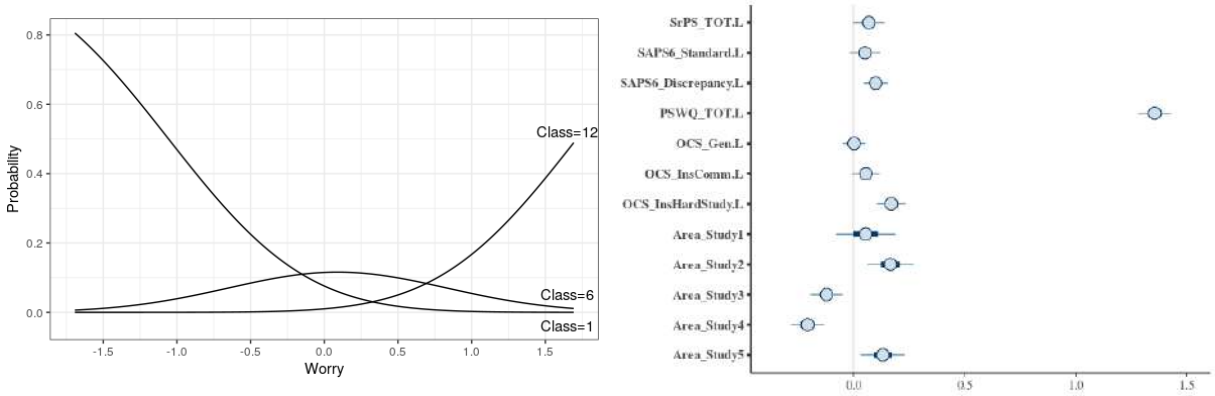


Figure 1: Point estimates of probability to belong to a Studyholism class as a function of Worry (left, other variables null) and credible intervals of betas in the linear predictor (right).

Out (LOO) predictive distributions $p(y_i \mid Data_{\setminus y_i})$ were sampled to produce further diagnostic plots. The Pareto-Smoothed Importance Sampling (PSIS) diagnostic plot (not shown) did not reveal any particularly difficulty in predicting observations: all values were below 0.3.

The MCMC simulation output produced a sample of 19500 realizations from the posterior distribution that were exploited to obtain marginal summaries like posterior Bayesian credible intervals (level $= 0.95$, Figure 1, right). In Table (1), statistical marginal summaries are shown. The posterior average of the beta parameter for Worry is about one order of magnitude larger than the other parameters; moreover, its (approximated) marginal posterior distribution is enterely located quite far from zero. It is also worth noticing that effects related to the area of study are relevant in almost all cases, and credibility intervals are also useful for informal Bayesian (Lindley) tests of the hypothesis: the null hypothesis of no effect is rejected when the null value is outside the credibility interval. These findings stand out as remarkable when compared with the original analysis of Loscalzo and Giannini (2019) enterely performed under normality: they found that Worry was the strongest predictor of Studyholism (beta $= 0.67, p < 0.001$). Other statistically significant predictors were Perfectionistic Concerns (beta $= 0.06, p = 0.001$), Teachers Overstudy Climate-Hard Study (beta $= 0.08, p < 0.001$), and Area of Study-Humanities (beta $= -0.08, p = 0.01$). Though, Loscalzo and Giannini (2019) highlighted that betas for these variables are too low for concluding that they are predictors of Studyholism since they are under their selected cut-off of 0.10. In line with this, the present study found that Worry is the strongest predictor of Studyholism, and that also Perfectionistic Concerns and Teachers Overstudy Climate-Hard Study are statistically significant predictors. About Area of Study, all the areas (except for the Technological one) resulted statistically significant.

However, the interpretation of parameters is not exactly the same. From one side, betas are regression coefficients of the latent normal variable underlying the ordinal Studyholism variable, thus parameters are comparable in magnitude: contrasts for ordinal explanatory variables are built from orthogonal polynomials with null mean and unit variance; on the other side, changes of conditional mean for the latent variable correspond to (non linear) changes of probability value over the ordinal classes of the manifest response, given the estimated cutpoints (Figure 1, left).

In the literature, a general advise for quantitative psychologists has been provided by Liddell and Kruschke (2018). The authors strongly advocate use of models on the ordinal scale because they better describe the data. They also make a case for Bayesian methods recognizing their flexibility, richness and accuracy in providing parameter estimates. In the original frequentist analysis, despite several questionnaire items were added and standardized following an aggregation provided by substantive reasons, the analysis of residuals on several path variables

Variables	Mean	St.Dev.	$Q2.5\%$	$Q25\%$	$Q50\%$	$Q75\%$	$Q97.5\%$
SrPS_TOT.L	0.07	0.04	−0.02	0.04	0.07	0.10	0.15
SAPS6_Standard.L	0.05	0.04	−0.03	0.02	0.05	0.08	0.13
SAPS6_Discrepancy.L	0.10	0.03	0.04	0.08	0.10	0.12	0.16
PSWQ_TOT.L	1.36	0.04	1.27	1.33	1.36	1.39	1.44
OCS_Gen.L	0.00	0.03	−0.06	−0.02	0.00	0.02	0.06
OCS_InsComm.L	0.06	0.04	−0.02	0.03	0.06	0.08	0.13
OCS_InsHardStudy.L	0.17	0.04	0.09	0.14	0.17	0.20	0.25
Area_Study1	0.05	0.08	−0.11	0.00	0.05	0.11	0.21
Area_Study2	0.17	0.06	0.04	0.12	0.17	0.21	0.29
Area_Study3	−0.12	0.04	−0.21	−0.15	−0.12	−0.09	−0.04
Area_Study4	−0.21	0.05	−0.30	−0.24	−0.21	−0.17	−0.12
Area_Study5	0.13	0.06	0.01	0.09	0.13	0.17	0.25

Table 1: Mean, standard deviation and percentiles of marginal posterior distributions (cutpoints not shown).

revealed that the adoption of the normal family for the manifest variables is not safe: while point estimates from least squares are not questioned, confidence intervals and statistical tests could be. Furthermore, the interpretation of parameters in the manifest-as-normal approach depends on the assumption that the adopted scale is not an artifact, thus the explicit introduction of latent normal variables does not constitute an entirely new element of the proposed model. Indeed, moving the normality at the level of latent variables also has the benefit of avoiding tail events of null probability be associated with non null tails of the normal distribution. Thus we conclude that, although results obtained in the original analysis of Studyholism were not overturned in this work (e.g. Worry is a strong predictor), Bayesian ordinal models are a rich class of models ready for being exploited by psychologists in routine work (e.g., see Figure 1).

References

Atroszko, P.A., Andreassen, C.S., Griffiths, M.D., Pallesen, S. (2015). Study addiction? A new area of psychological study: Conceptualization, assessment, and preliminary empirical findings. *Journal of Behavioral Addiction*, **4**(2), pp. 75–84.

Griffiths, M. D., Demetrovics, Z., Atroszko, P. A. (2018). Ten myths about work addiction. *Journal of Behavioral Addiction*, **7**(4), pp. 845–857.

Liddell, T.M., Kruschke, J.K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, **79**, pp. 328–348.

Loscalzo, Y., Giannini, M. (2017). Studyholism or study addiction? A comprehensive model for a possible new clinical condition, in *Advances in Psychology Research*, eds. A. M. Columbus, Nova Publishers, New York (USA), pp. 19–37.

Loscalzo, Y., Giannini, M. (2018a). Response to: Theoretical and methodological issues in the research on study addiction with relevance to the debate on conceptualising behavioural addictions: Atroszko (2018). *Psychiatria i Psychologia Kliniczna*, **18**(4), pp. 426–430.

Loscalzo, Y., Giannini, M. (2018b). Problematic overstudying: Studyholism or Study Addiction? Commentary on: Ten myths about work addiction (Griffiths et al., 2018). *Journal of Behavioral Addictions*, **7**(4), pp. 867–870.

Loscalzo, Y., Giannini, M. (2019). Heavy Study Investment in Italian College Students. An Analysis of Loscalzo and Giannini's (2017) Studyholism Comprehensive Model. *Frontiers in Psychiatry*, **10**(489), doi:10.3389/fpsyt.2019.00489/full

City Prosperity Index: a comparative analysis of Latin American and Mediterranean cities based on well-being and social inclusion features.

Alessio Surian[a], Andrea Sciandra[b]
[a] Department of Philosophy, Sociology, Education and Applied Psychology (FISPPA), University of Padova, Padova, Italy.
[b] Department of Communication and Economics, University of Modena and Reggio Emilia, Reggio Emilia, Italy.

1. Introduction

Paying specific attention to the different stages of modernization, this study analyses and compares recent data provided by the City Prosperity Index (CPI) concerning Latin American and Mediterranean cities. CPI provides baseline data to monitor Sustainable Development Goal 11 aiming at making cities and human settlements inclusive, safe, resilient and sustainable. It is an instrument based on UN-Habitat concept of urban prosperity as being composed by six dimensions: productivity; infrastructure; quality of life; equity and social inclusion; environmental sustainability, and governance and legislation. CPI is meant to support decision-making for multi-scale governance ranging from local and national urban policies to regional strategies by identifying opportunities and potential areas of intervention based on understanding the relations between different dimensions of urban development.

The study focuses on Quality of Life (QOL) and Equity and Social Inclusion (ESII) indicators as key data in order to monitor Sustainable Development Goal 11 with specific attention for urban inclusion. Specifically, it relates QOL and ESII indicators to two Goal 11 targets addressing to enhance inclusive and sustainable urbanization and capacity for participatory, integrated and sustainable human settlement planning and management in all countries (by 2030); and to substantially increase the number of cities and human settlements adopting and implementing integrated policies and plans towards inclusion, resource efficiency, mitigation and adaptation to climate change, resilience to disasters, and develop and implement holistic disaster risk management at all levels (by 2020).

Based on CPI's units of analysis, this study performs a comparative analysis of cities from different countries belonging to two areas: Latin America and Mediterranean (including both European and Northern African cities). This allowed us to explore the different stages of modernization (Inglehart and Welzel, 2005) in relation to the dimensions of social inclusion and well-being.

2. Theoretical approach

According to Adler and Kwon (2014) the basic social capital thesis has been broadly accepted, as social ties can be efficacious in providing resources, influence and solidarity. Consequently, the research in this field is shifting toward new directions dealing with more specific aspects, such as inequalities where social capital can be involved as an opponent but also as a factor of reproduction of inequalities of income, health, education and inclusion. Moreover, among many studies on the relationship between social capital and life satisfaction, Bjørnskov (2003) found that generalized trust and civic participation have a positive strong relationship with happiness at the national level and above their effects on income.

In addition, a World Bank Study (Narayan, 1999) specifies that bridging social capital proved to be essential for social cohesion and for poverty reduction. This study focuses on the role of cross-cutting ties between social groups and between society's formal and informal institutions,

ISBN: 978-88-5495-135-8

showing that an inclusive development could be a crucial factor for social well-being. Within this perspective Narayan (1999) identifies three key dimensions: inclusive participation, education, and decentralization. Such dimensions enhance the importance of engagement, including individuals' involvement in the public decisions making process (Helliwell, 2006). In this context social exclusion refers to the lack of access to a range of citizen rights (health, education, etc.) and a lack of societal integration in the ability to participate in political decision-making (Shortall, 2008).

From an inclusive participation perspective, based on education and decentralization, the city is a promising unit of analysis to address the above assumptions. To this purpose we chose a meso-level of analysis, focusing on the cities as distinct organizations and trying to identify significant connections between micro and macro levels.

Our empirical analysis aims at exploring whether economic development is involved in activating cultural changes (Inglehart and Welzel, 2005) in terms of individual autonomy, social and gender inclusion.

3. The City Prosperity Index: instrument and data

The City Prosperity Index serves as a monitoring and diagnostic tool providing evidence-based for policy-making and accountability. Cities participate in the initiative in two ways: (1) city governments interested in conducting the CPI approach UN-Habitat out of their own initiative; or (2) national governments join the initiative with a representative sample of cities. The national samples can be chosen with the support of UN-Habitat, considering the size of cities, functionality, location and other relevant attributes.

At present, CPI has been applied in over 400 cities across 46 countries. CPI includes a total of 72 indicators grouped into 6 main dimensions for each city. In particular this study focuses on 200 cities located in Brazil, Mexico, Northern African and European Mediterranean.

We chose to compare cities from the two most populated Latin American countries (Brazil and Mexico) with cities from the Mediterranean area (Northern African and European countries) in order to analyse different economic and social contexts including different stages of modernization. In fact, the chosen cities data show high variability in several productivity and social inclusion features, such as City Product per capita, Unemployment Rate and Poverty.

The City Prosperity Index is computed using city level data deriving from a set of commonly available indicators that exist among all cities. In this way, CPI acts as a platform for regional/national comparison purposes. The global CPI is a weighted mean of standardized indices from each of the six dimensions (productivity; infrastructure; quality of life; equity and social inclusion; environmental sustainability, and governance and legislation). Each dimension is based on three to five sub-dimensions, which include several indicators that allow for the calculation of the specific sub-index. Therefore, CPI emphasises the fact that urban prosperity, well-being and human development are broader than economic achievements or growth and that they are multidimensional categories that can only be measured more accurately by using a composite index (UN-Habitat, 2016). The index provides a 0-100 score for each indicator, sub-dimension and dimension. The data used in calculating each indicator come from the last available year, with all data collected between 2010 and 2017. So, CPI enables to identify which dimension is performing well or poorly in terms of prosperity of the city and to identify which features require specific intervention measures.

Our analysis focused on Quality of Life (QOL) and Equity and Social Inclusion (ESII) indicators. QOL index includes the following sub-dimensions: Health (indicators: Life Expectancy at Birth and Under-Five Mortality Rate), Education (indicators: Literacy Rate and Mean Years of Schooling) and Security (indicator: Homicide rate). In turn ESII index is segmented into three sub-dimensions: Economic Equity (indicators: Gini Coefficient and Poverty Rate), Social Inclusion (indicators: Slum Households and Youth Unemployment) and Gender Inclusion (indicator: Equitable Secondary School Enrolment).

ISBN: 978-88-5495-135-8

Although the CPI includes only 7 Mediterranean cities, we chose to compare the indicators of this group with those of Latin America through the t-test. Non-significant differences between the two groups, in terms of well-being and social inclusion features, would suggest a shift towards a post-materialist perspective of the cities of the two Latin American countries. Subsequently, through a cluster analysis we wanted to identify which cities have similar performances with respect to the CPI indicators, going beyond the geographical location. This analysis could also provide further insights into the well-being and the social inclusion dimensions for the sampled cities.

4. Analyses and results

We started by analysing the distributions of our selected indicators by comparing cities within their countries and within the whole sample. For each indicator we carried out statistical tests in order to identify significant differences between the Mediterranean area and Latin America. We made use of non-parametric tests and bootstrap sampling suitable for small samples and non-normal data.

An interesting result (Table 1) occurred using t-tests with 1000 bootstrap samples (Efron and Tibshirani, 1994) to compare the Latin American cities with the Mediterranean ones (Southern Europe and Morocco). The analysis does not show statistically significant differences concerning indicators relating to Health (Life Expectancy at Birth and Under-Five Mortality Rate) and Education (Literacy Rate and Mean Years of Schooling). Instead, the indicators about Security (Homicide rate), Economic Equity (Gini coefficient, Poverty rate), Social Inclusion (Slum Households) and Gender Inclusion (Equitable Secondary School Enrolment) are significantly different in the two groups, with more positive results for the Mediterranean area. It must be emphasized that Youth Employment turned out to be significantly higher in Brazilian and Mexican cities when compared to the cities belonging to the Mediterranean area. Since Moroccan cities accomplished better results than Southern European cities in Youth Unemployment indicator, we believe that CPI is showing the persistent problems of Southern Europe with respect to this issue.

Table 1: Boostrap (1000 samples) t-tests for CPI QOL and ESII standardized indicators comparing Mediterranean cities (n=7) and Latin American cities (192).

	Mediterranean cities		Latin American cities		
	Mean	Std.dev.	Mean	Std.dev.	sign. (*)
Life Expectancy at Birth	84.017	19.426	71.125	3.755	.113
Under-Five Mortality Rate (reversed)	72,883	22,167	56,972	5,338	.071
Literacy Rate	86,980	17,579	93,714	7,515	.366
Mean Years of Schooling	59,081	18,972	70,884	11,725	.076
Homicide rate (reversed)	96,129	5,338	60,771	10,447	.001*
Gini Coefficient (reversed)	74,433	11,136	47,272	12,927	.001*
Poverty Rate (reversed)	76,224	8,622	47,179	9,691	.001*
Slum Households (reversed)	94,430	7,378	83,424	20,520	.002*
Youth Unemployment (reversed)	23,321	19,085	69,046	11,420	.001*
Equitable Secondary School Enrolment	92,496	8,384	82,976	12,993	.004*

Additionally, an average linkage cluster analysis (Hair, Black, Babin, and Anderson, 2010) allowed comparison among Brazilian, Mexican, Moroccan, and Southern European cities (Figure 1). By grouping data by Quality of Life and Equity and Social Inclusion indicators, cluster analysis shows four main groups. Brazilian and Mexican cities have very similar performances (the biggest cluster), while Moroccan cities show a gap in QOL indicators when compared to Southern European cities, even though Moroccan cities accomplish better results in relation to Poverty Rate and Youth Unemployment. The last cluster is made up by Mexican cities which accomplished better results than the other Latin American cities in terms of Economic Equity (Gini coefficient and Poverty rate), Social Inclusion (Slum Households and Youth

ISBN: 978-88-5495-135-8

Unemployment) and Gender Inclusion (Equitable Secondary School Enrolment).

Figure 1: Cluster Analysis (average linkage method) based on Quality of Life and Equity and Social Inclusion indicators.

5. Conclusions

These analyses may suggest an incomplete but visible improvement of well-being and social inclusion features in the two most populous Latin American countries or, in other words, a shift towards a post-materialist perspective. This shift could be related to the socio-economic transformations that produce intertwined results in terms of both self-expression as well as throwback towards survival values.

A clear limitation of this research is the scarce representativity of the Mediterranean cities. As the number of cities joining CPI initiative increases and time series accumulate, it will be possible to learn more about what types of cities are likely to be successful in urban prosperity and under what mechanisms and circumstances. Future comparison might as well consider complementary indexes such as the City Resilience Index (CRI) in order to discuss the capacity of cities to adapt to social and environmental challenges.

References

Bjørnskov, C. (2003). The happy few: Cross–country evidence on social capital and life satisfaction. *Kyklos*, **56**(1), pp. 3-16.

Efron, B., Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press, Boca Raton, FL.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. (2010). *Multivariate data analysis. 7th Edition*, Pearson, New York.

Helliwell, J. F. (2006). Well- Being, social capital and public policy: What's new?. *The Economic Journal*, **116**(510).

Inglehart, R., Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.

Kwon, S. W., Adler, P. S. (2014). Social capital: Maturation of a field of research. *Academy of management review*, **39**(4), pp. 412-422.

Narayan, Deepa. 1999. *Bonds and bridges: social and poverty (English)*. Policy Research working paper; no. WPS 2167. World Bank, Washington, DC.

Shortall, S. (2008). Are rural development programmes socially inclusive? Social inclusion, civic engagement, participation, and social capital: Exploring the differences. *Journal of Rural Studies*, **24**(4), pp. 450-457.

UN Habitat (2016) *Measurement of City prosperity. Methodology and Metadata*. United Nations, Nairobi.

Invariance in the structural topic models

Emma Zavarrone [a], Maria Gabriella Grassia [b], Rocco Mazza [b]
[a] Department of Business, University IULM, Milan, Italy;
[b]Department of Social Sciences, University of Naples Federico II, Naples, Italy;

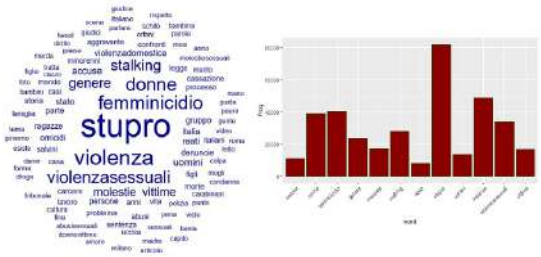
1. Introduction

The information creation process involving the mass media also uses the source of social networks. The indiscriminate use of this source is very dangerous as it could be exposed to the risk of generalizing the ideas expressed by few in a widespread opinion. However, what people comment on or express on social networks if properly analyzed can help to understand socio-political dynamics. Starting from the collection of tweets concerning the wide reported semantic spectrum of women, we focused on the violence against women and women hate speech (Davidson et al., 2017; Malmasi and Zampieri, 2018). In this paper we propose a invariance study for latent concept, starting from a non-automated keyword selection.

2. The Data

A development of Latent Dirichlet Allocation (LDA) has been applied to a dataset composed of 403612 Italian tweets. The tweets were extracted using specific search keywords, based on previous content analysis, from July 2018 to May 2019. The collection of tweets represents the raw corpus, after the cleaning and pre-treatment phases, a series of outputs have been extracted which allow us to frame the manifest semantic dimensions of the violence against women.Figure 1 shows the wordcloud that describes the most representative graphic forms of the extracted texts, in terms of frequency. It is possible to understand which are the main discussion themes about hate speech from this first result. Particularly we refer to to violence against women. We can schematically trace three dimensions to which the illustrated graphic forms refer: the first is banally referable to the cases of news commented by users on the social; the second to an institutional and regulatory dimension, in fact there are references to the need for stronger penalties; in the last one we find the ways in which violence against women can be realized. The opinions study of on social network users regarding a specific topic is very

Figure 1: Wordcloud and top 12 term frequency



difficult. For this reason it seemed appropriate to isolate and deepen with a specific reading also the hashtags that accompany the tweet. In figure 2 it is possible to observe the resulting word cloud. The hashtags provide a quick reading key: as well known, their use drives the discussions and the community users organize their observations and considerations. In fact

ISBN: 978-88-5495-135-8

we found further insights. There are references to television programs that have dealt with the topic and around which the discussion has been concentrated. Moreover references to events and moments of reflection on the phenomenon are present. This indicates a community that is very active and attentive to the evolution of the phenomenon, capable of cross-media reading and activity. Not only that, it is also possible to note that the same phenomenon has extended references to several media levels and not concentrated in a single space.

Figure 2: Hashtag wordcloud

3. Our proposal

This paper aims to study the invariance between latent topic distribution within documents extracted using no-automated keywords. The tools available for studying invariance within statistical models are many, we are interested in identifying a measure that allows us to calculate a divergence between parameters of the model selected for themes extraction. The starting point is the LDA model, applied for the extraction of latent themes within the collected documents, at the base of the LDA we find this assumption:

- Documents are mixtures of topics, where a topic is a probability distribution over words (Steyvers and Griffiths, 2007). It is a generative and Bayesian inferential model.
- Only documents and words have been observed. The topics have not considered , these are partially hidden, latent precisely, within the structure of the document.

The model goal is to infer the latent structure of topics, consisting of documents and words, it does this by recreating the documents in the corpus considering the relative weight of the topic in the document and the word in the topic, in an iterative way. Documents and topics distributions can be described by a Dirichlet distribution (Blei et al., 2003), as it is the conjugate distributions of the Multinomial, it is very convenient to use it as a priori, and it follows that it can be an excellent tool as regards the problems of model inference (Neapolitan, 2003; Balakrishnan and Nevzorov, 2003). Taking a function a posteriori from the Dirichlet distribution, parameterized by higher weights on a single topic, we can derive a distribution that goes to make up every topic mix. This is essentially what a Dirichlet distribution does, sampling probability distributions of a specific type. To define the process schematically: from a *Dir* (α) Dirichlet distribution we carry out a random sampling which represents the distribution of the topics of a particular document; this topic distribution is θ; from θ, we select a particular *topic Z* based on the distribution; then from another Dirichlet distribution *Dir* (β), we select a random sample representing the word distribution of the topic *Z*, this distribution isψ, from ψ we choose the word *w*. Formally, the process to generate each word from a document is as follows:

1. Extract $\theta_i \sim$ Dirichlet (α), where i $\in \{1,...,M\}$
2. Extract $\psi_k \sim$ Dirichlet (β), where k $\in \{1,...,K\}$
3. For each i , j of word, dove j $\in$ 1,...,N, e i $\in$ 1,...,M
   (a) Extract a topic from z_i,j $\sim$ Multinomial(θ_i)
   (b) Extract a word from w_i,j $\sim$ Multinomial($\psi_{zi,i}$)

With the study of the divergence between the topic distribution, we want to develop a measure for the costruct invariance (Cheung et al., 2002) among the tweets grouped by extraction keyword. The objectives are two:

- Identify in the literature a divergence measure between the parameters θ calculated by the model.
- Synthesize an unique measure that indicates invariance in terms of divergence.

For the first point, the literature review has highligthed few original studies, the Jensen-Shannon divergence measure (Lin, 1991; Endres and Schindelin, 2003) has been chosen because it is not negative and symmetrical.

4. Results

Due to computational limitations we decided to split the dataset into a training set extracted through a random sample. We have extracted 2% of the texts for each keyword. After the pre-treatment of the data and the application of the model, the goal is calculate the a posteriori probability distributions of the topics and study their invariance in terms of divergence by using a specific measure. The first objective has been satisfied. In Fig. 4 there is a representation of the divergence matrix obtained on the key #abusosessuale. The figure shows compact blocks (the red ones) in which the distributions tend to diverge, this highlights the need to identify a validation measure for the keywords. For the second point we will develop a synthesis indicator. This will capable of providing a synthetic measure of the divergence between probability distributions.

5. Discussion

The innovative proposal of this paper is to give a tool to study the relationships between topics distributions, thus allowing a validation for the keywords used for the extraction. In this way it will be possible to validate the selection of keywords made to extract texts that refer to the same concept.
In coclusion, at the time the work have three limits:

1. The tweets sampling does not allow a correct generalization of the results obtained.
2. There is a computational limit for the constructed model.
3. The constructed divergence indicator needs to be tested and compared with other measures.

ISBN: 978-88-5495-135-8

Figure 3: *a*, the figure show keywords selection for hate speech concept; *b*, the model proposed, the # mark the hashtags used for the extraction of the texts

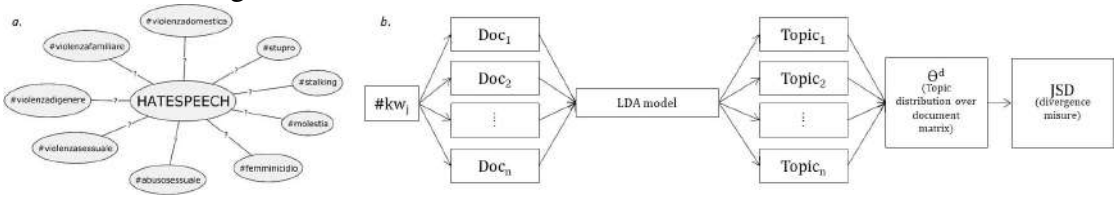
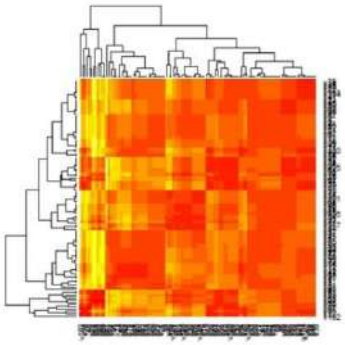


Figure 4: *a*, top 20 terms in each topic; *b*, the heatmap represents the divergence measured on keyword #abusosessuale

a.

Topic 1	Topic 2	Topic 3
stupro	violenzasessuale	violenza
vittime	stalking	donne
italia	molestie	femminicidio
ragazze	violenzadomestica	genere
gruppo	reati	uomini
vita	persone	casa
italiani	accusa	anni
storia	stato	omicidi
problema	parte	figli
colpa	legge	lavoro
tratta	confronti	mondo
parte	processo	carabinieri
italiano	polizia	paese
aggravante	carcere	marito
stuprata	abusi	morte
cultura	pena	uccisa
video	salvini	orfani
stato	denuncia	violenze
sesso	mesi	bambini
abusisessuale	accuse	parole

b.



References

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, **9**(2), pp. 233-255.

Blei, D. M., Ng, A. Y., & M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, pp. 993-1022.

Lin J. (1991). Divergence measures based on the Shannon Entropy. *IEEE Transactions on Information Theory*, **33**(1), pp. 145-151.

Endres M. & Schindelin J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, **49**(3), pp. 1858-1860.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder Luis, J., Gadarian, S. K., & Rand, D. G. (2014). Structural topic models for open ended survey responses. *American Journal of Political Science*, **58**(4), pp. 1064-1082.

R. E. Neapolitan (2003). Learning Bayesian Networks. *Prentice-Hall*,

N. Balakrishnan & V. Nevzorov (2003). A Primer on Statistical Distributions. *Wiley-Interscience*

Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, **30**(2), pp. 187-202.

Steyvers & Griffiths (2007). Probabilistic topic models, in *Latent Semantic Analysis: A Road to Meaning*, eds. T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Lawrence Erlbaum, page 427.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language, in *Eleventh international aaai conference on web and social media.*

ISBN: 978-88-5495-135-8

Inferring Twitter users home location based on trend topics

Paola Zola [a], Costantino Ragno [b], Paulo Cortez [c]
[a] Department of Economics and Management, University of Brescia, Brescia, Italy;
[b] School of Science and Technology, University of Camerino, Camerino, Italy;
[c] Department of Information Systems, University of Minho, Guimarães, Portugal.

1. Introduction

Inferring Twitter users home location is a growing interest research topic given its importance for many application. For example knowing the people location is fundamental in event detection studies, recommendation systems, friendship network analysis and so on. Some studies have been done in order to estimate the users' home country given the world distribution (Zubiaga et al., 2017) while other studies focused on more fine grain location prediction relatively to a specific country or region (Eisenstein et al., 2010).

In this work we propose a novel tool that aim to infer Twitter users home location at the finest grain (coordinates) considering all the world surface. The proposed approach follows the one in Zola et al. (2019) where for each Twitter account the nouns distribution is evaluated and, passing thought Google Trends (GT) data, the Twitter user location estimation is performed. However, while in (Zola et al., 2019) the prediction is determined at country level, in this research we apply clustering algorithm to assign to each user a probability distribution over the whole world area in order to identify an unique couple of coordinate associated to the user home location.

The dataset is composed by 2,880 Twitter account with a verified location, and, for each we consider his/her historical 3,200 tweets. For each account we extract the cities distribution from GT using as keyword the nouns (generic and proper) hold in the tweets collection. Having the cities distribution we define the cities polygons sampling points in order to respect the GT data distribution. Then the Gaussian Mixture Models clustering algorithm is performed. The results are evaluated using mean and median absolute error computed on the Haversine distance from the ground truth users home location derived in the work of Zola et al. (2019).

2. Proposed Approach

The proposed approach which follows the one adopted in Zola et al. (2019), uses only tweet nouns assuming that they are the most representative part of speech able to identify different countries. For user u, the proposed approach works by first identifying the sequence of all nouns $\mathbf{n}_u = \langle n_1, n_2, ..., n_{l_u} \rangle$. To obtain $\mathbf{n}_u$, the tweets are first preprocessed by transforming the text to lowercase and removing English stopwords.

For each noun $n_i \in \mathbf{n}_u$, a GT query is executed by using the `Pytrends` Python module. The GT query results for noun n_i is a sequence with integers confidence scores for the most frequent cities C that have typed the specific noun $n_i \in \mathbf{n}_u$.
Denoting by C_u the city distribution for a given user given all his/her nouns, we need to locate these cities on the World surface and find the bigger cluster. To perform this operation we computed each cities polygons $p(c) for c \in C$ as the World area on which the specific city c is located.

The $p(c)$ is defined as two-dimensional polygon where the edges corresponds to the physical borders of the city c. The aim is to sample uniformly a certain number of two-dimensional points

ISBN: 978-88-5495-135-8

(latitude and longitude) in the city polygon representing the GT integer confidence scores generated for each cities. To obtain the cities polygons we scrapered the data from OpenStreetMap website[1]. Whenever the city c in study is not present on openstreetmap website we considered as a proxy of the city polygon a circumference centred in the city coordinates and with a radius equal to $r = \sqrt{\frac{A}{\pi}}$ where A correspond to the city area. The city coordiantes centers (latitude and longitude) are derived from Wikidata[2]. Once each city polygons is complete we sampled a finite number of points from each city polygons representing the complete GT data distribution.

Gaussian Mixture Models The Gaussian Mixture Model (GMM) (Banfield & Raftery (1993)) is a probabilistic model to represent the distribution of a population as a linear combination of Gaussian. Usually, each Gaussian k in the GMM is called component. If each component in the mixture is interpreted as a cluster of the dataset then the GMM can be seen as a unsupervised learning algorithm for classification. Let's define K as the number of components in the GMM, α_k as the weight of the kth component, $\vec{\mu}_k$ and Σ_k as the location and the scale parameter of the component k, then the GMM model for the random variable X has the following distribution:

$$p(X) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(X|\vec{\mu}_k, \Sigma_k) \tag{1}$$

where $\sum_{k=1}^{K} \alpha_k = 1$.

The estimation is performed with the Expectation Maximization Algorithm (EM), an iterative procedure which alternate the computation of the expectation of the likelihood (E-step) with the maximization of the expectation of the likelihood (M-step) up to convergence. The only parameter that need to be fixed in a GMM is the number of cluster K. In this work we select the best K value according to the Bayesian Information Criterion (BIC) as performed in Bakerman et al. (2018).

Evaluation To evaluate the proposed approach we computed the Mean Absolute Error (MAE) and the Median Absolute Error (MdAE) on the Haversine distance mesures expressed in Kilometers as proposed in Eisenstein et al. (2010).
The Haversine formula measures the the great-circle distance between two points on a sphere given their longitudes and latitudes. The great-circle distance is the shortest distance between two points on the surface.

3. Results

This Section reports the results of the proposed method. We performed the GT query search for all 2,880 users according to the nouns extracted from their old tweets. Each user queries resulted in a city distribution from which we sampled random data respecting the GT queries distribution. Then, the GMM clustering algorithm is computed until an unique location (pair of latitude and longitude) is extracted. The Table 1 reports the MAE and MdAE, express in kilometers, computed between the GMM estimated location and the ground truth ones. The MAE indicate that the mean average error is around 4 thousand km given all the world surface. The median absolute error is of 1,7 thousand km which is, for example, the distance between

[1] www.openstreetmap.org
[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

Italy and Portugal.

Moreover, looking at the distribution of the Haversine distance between the true location and the estimated points we got the some users are correctly located with an error of few meters while, for others the estimation error is huge. In particular, we noticed that location estimation errors are frequent among anglophone countries and, in particular, Australian users are often located in United States. The higher difficulty of Anglophone country identification is well know in geolocation reseach filed (Han et al. (2012)). However, the GMM error that involves Australia/New Zeland and USA might be related to the low population density in the Oceanian continent that results in a lower number of Australian cities from GT queries.

Table 1: Evaluation Metrics of the proposed approach

Metric	**Kilometers**
MAE	4359.05
MdAE	1759.40
Min	0.01
Max	18717.46

Figures 2–1 show an example of two located users. Figure 1 represents, with different colours, clusters identified by the GMM at the first iteration while the yellow star is the final point estimated by the model. In Figure 1 the GMM predicted values are respectively 53.4807 for the latitude and -2.2426 for the longitude, identifying the city of Manchester (UK), which is the exact city of user origin in the ground truth data. Differently, Figure 2 represent a misclassified Twitter user. The true user origin is the city of Perth in Australia while the predicted one is Chicago (USA).

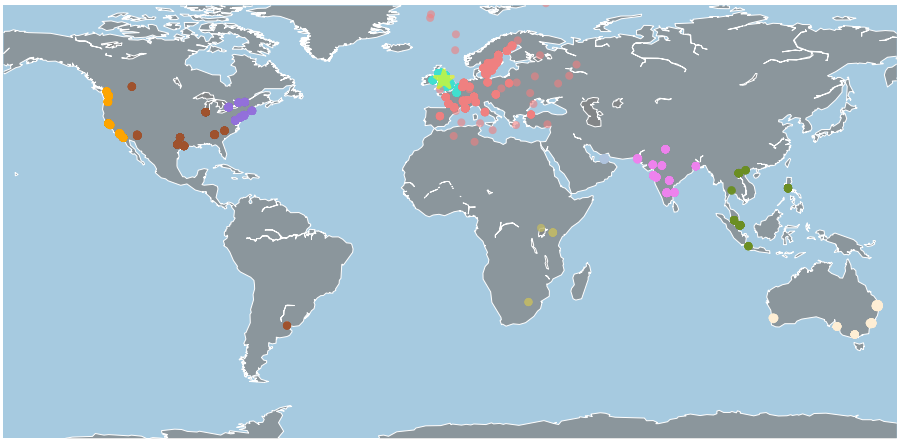


Figure 1: GMM corrected location estimation.

4. Conclusion

Inferring Twitter user geolocation is not a trivial task and it is fundamental for many social media analytics application. Estimating the user home location at coordinate level has been studied but, in general, focusing on specific are (as United States). In this work we proposed

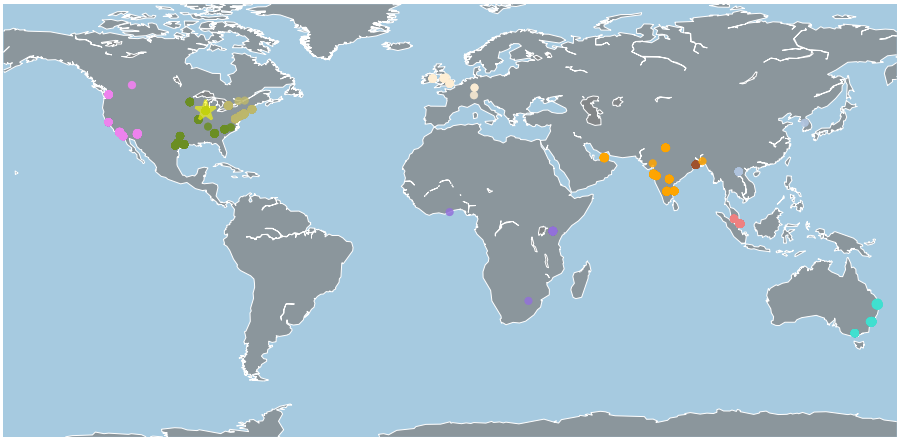


Figure 2: GMM wrong location estimation.

a novel approach to infer Twitter users location considering all the World surface using a textual based approach. The proposed model is based on Google Trends (GT) city distributions given the Twitter users nouns hold in their tweets, then the estimation of the coordinate points is performed by GMM clustering algorithm. The results show encouraging performance. However, some limits are evident for anglophone countries (Australia, United States and United Kingdom), thus, further studies will be done in order to increase the goodness of the proposed approach, for example, weighting the city distributions with citys' population density information.

References

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pp. 803-821.

Bakerman, J., Pazdernik, K., Wilson, A., Fairchild, G., & Bahran, R. (2018). Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **12**(3), pp. 34.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010, October). A latent variable model for geographic lexical variation. *In Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 1277–1287.

Han, B., Cook, P., & Baldwin, T. (2012, December). Geolocation prediction in social media data by finding location indicative words. *In Proceedings of COLING 2012*, pp. 1045-1062.

Zubiaga, A., Voss, A., Procter, R., Liakata, M., Wang, B., & Tsakalidis, A. (2017). Towards real-time, country-level location classification of worldwide tweets. *IEEE Transactions on Knowledge and Data Engineering*, **29**(9), pp. 2053–2066.

Zola, P., Cortez, P., & Carpita, M. (2019). Twitter user geolocation using web country noun searches. *Decision Support Systems*, **120** pp. 50–59.

ASA Conference 2019 - Book of Short Papers
Statistics for Health and Well-being
University of Brescia, September 25-27, 2019
Maurizio Carpita and Luigi Fabbris (Editors)

ISBN: 978-88-5495-135-8 October, 2019