ISSN 2611-0172

UNIVERSITÀ DEGLI STUDI DI PALERMO

dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

d/SEAS WORKING PAPERS

Editor in chief / Direttore *Calogero Massimo Cammalleri*
Vol 1 No 1 (2017): Inaugural issue

inside volume / nel volume

Dipartimento di Scienze Economiche Aziendali e Statistiche, publisher

Editorial/Editoriale

I am glad to announce publishing of the first issue of d/SEAS Working Papers; - the new journal of Dipartimento di Scienze Economiche Aziendali e Statistiche dell'Università degli Studi di Palermo [d/SAEAS for short]. It gathers eight papers in all reseach fields of the d/SEAS: economics, law, business & managment, sociology, statistics. The aim of the SWPs is to share with a wide audience of academics, practioners and policy makers original research in the above said research fields. I hope our Inaugural issue encourages the members of dSEAS, as well as all visitors, phD students, participants in conferences and workshops organized by the dSEAS to submit their papers to the SWPs and so being active part in the sharing of a free knowledge as it is in the spirit of academic community and that readers will find this new publication useful for their purposes. The d/SEAS Working Papers aims to be an active space for discussion as well: -for this reason it is possible, by registering, to interact through platform with the authors of the papers.

This Inaugural Issue publishes:

University student talent: the real driver for performance? by Giovanni Boscaino and Giada Adelfio about the university student performance, and its measurement. It investigates the role of a latent variable that can take into account the student motivation, aptitude, and abilities, here conveniently called talent and the results, they got through a random effect Quantile Regression on a new measure of Italian student performance, seems to highlight the main role of the talent).

Credit demand and supply shocks in Italy during the Great Recession, by Andrea Cipollini and Fabio Parla, uses Structural VAR analysis to disentangle credit demand and supply shocks and their effect on real economic activity in Italy during the 2008-2014 crisis period observinmg data at annual frequency for each of 103 Italian provinces. The empirical findings suggest a more important role of credit supply shocks in shaping the level of real economic

activity. Furthermore, the results show that credit crunch hits the North of Italy less than the remaining macro-regions, especially the South-Italy.

Saving process within a zero waste strategy in Sicily: a system dynamics approach, by Andrea Cuccia, shows implementation of a zero waste strategy in a small municipality in Sicily. By building up a SD model it figure out how saving process connected to the decrease of amount of waste piled up in the landfill, has been boosting this virtuous cycle, which is bound to cope with the citizens claims about environmental care and lower tax burden level.

Some features of deindustrialisation in EU15 during the period 1999-2004: a multivariate analysis, by Maria Davì, examines the causes and the different modalities of deindustrialisation process closely with reference to 13 of the first EU15 countries. According to this essay the estimates indicate that the main factors responsible of the dynamics of deindustrialisation have been efficiency and the scale of production processes in various manufacturing activities.

Leisure, Social Capital and Life Turns in Deviant Youth, by Fabio Massimo Lo Verde, investigates the issues of the production of social capital in a specific area of everyday life such as leisure time and the different sociocultural contexts in Italy. The conclusion it reaches offers three metaphors for understanding the trends of leisure time and sociability.

A recap on Linear Mixed Models and their hat-matrices, by Gianfranco Lovison and Mariangela Sciandra, has a twofold goal: provides a recap of Linear Mixed Models (LMMs focusing on the derivation of theoretical results on estimation of LMMs and, on the other hand, it discusses various definitions that are available in the literature for the hat-matrix of Linear Mixed Models, showing their limitations and proving their equivalence.

Does regulatory regime matter for bank risktaking? A comparative analysis of US and Canada is a working paper presented by Sana Mohsni and Isaac Otchere at a seminar of ours Department. Our guest authors compare banking structure in the US e in Canada in the the wake of the worst financial crisis in 2008.Their analysis shows that entry restrictions, which create concentrated banking structure, restrictions relating to capital, liquidity and activities, and strong supervisory power and discipline positively related to the z-score, suggesting that these factors constrain excessive risk taking by Canadian banks.

Fabrizio Piraino's Ideas for a new reading of the law regulation of internet service providers addresses the issue of the so called internet service provider's liability under artt. 12-14 dir. 00/31, with specific regard to the violations of copyright. The study aims at demonstrating that European Law on ISP is not a law on tort, but regulates a sphere of lawful action in favor of internet service providers. The analysis of the Court of Justice's case-law reveals that the primary remedy against offenses committed on the internet is an injunction, while damages are only a secondary relief. This confirms the hypothesis that artt. 12-14 dir. 00/31 draw the perimeter of the legitimate activity of the internet service provider. The essay ends with a re-interpretation of the Italian provisions set forth at artt. 14-16 d.l gs. 70/2003 in order to align them, with the European directive, especially with regard to the hosting performance rules.

The editor in chief

Calogero Massimo Cammalleri

Articles/Articoli

UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 'in fase di assegnazione', volume I, 2017

University student *talent*: the real driver for performance?

Giovanni Boscaino · Giada Adelfio

Abstract Investigation about the university student performance, and its measurement, are very crucial issues for any policy maker. Since the economic crisis, jobs market requires even higher skills and competences. Literature offers a lot of papers about the university student quality and performance, in order to identify the main determinants of them. Often, results are very different, and they seems to hold just in a specific context. This paper aims to investigate the role of a latent variable that can take into account the student motivation, aptitude, and abilities, here conveniently called *talent*. A random effect Quantile Regression on a new measure of Italian student performance has been adopted, and results seem to highlight the main role of the *talent*.

Keywords Student performance · Indicator · Random effects Quantile Regression

Riassunto *Gli studi sulla performance delle carriere degli studenti sono oggetto di grande attenzione sia da parte della comunità scientifica sia da parte dei policy maker. L'attenzione negli ultimi anni sembra essere cresciuta, a pari passo con quella della valutazione dei sistemi formativi: la valutazione dell'efficacia e dell'efficienza di un servizio di formazione passa anche dalla valutazione delle performance degli studenti. Pertanto, l'esigenza di formare studenti ben preparati e in tempo è diventata una priorità per tutto il mondo accademico. La letteratura*

G. Boscaino: Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo
viale delle Scienze ed. 13, 90128
E-mail: giovanni.boscaino@unipa.it
· G. Adelfio: Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo
viale delle Scienze ed. 13, 90128
E-mail: giada.adelfio@unipa.it

scientifica offre diversi spunti di riflessione circa le dinamiche e le determinanti di una buona o cattiva performance studentesca ma i risultati non sembrano convergere verso un'unica direzione. Quello che emerge è che da un lato il contesto (area geografica/economica) ha un ruolo fondamentale, dall'altro la non convergenza verso un insieme di fattori determinanti (ad eccezione del Genere che risulta significativo nella maggior parte degli studi) potrebbe suggerire che si stia guardando nella direzione sbagliata. Lo studio qui riportato si inserisce in questo contesto: utilizzando le "classiche" variabili esplicative disponibili per lo studio del successo universitario, è possibile trovare indicazione del fatto che queste non sono sufficienti? In altre parole, il successo universitario può essere spiegato da caratteristiche intrinseche dello studente, che qui chiameremo genericamente "talento", piuttosto che da quelle socio-demografiche? A tal proposito, si è deciso di adottare una misura della performance universitaria alternativa ai Crediti Formativi Universitari (CFU) accumulati e ai voti conseguiti: tale misura tiene conto contemporaneamente del voto e del CFU così che, sempre in una scala da 18 a 30, l'unico valore espresso sia portatore di entrambe le informazioni. Il principio è che guardare solo ai CFU accumulati nel tempo non permette di valutare quali CFU siano stati accumulati, e che conseguire un 30 per un insegnamento di 3 CFU non ha lo stesso valore che conseguire un 30 in un insegnamento da 12 CFU. La misura proposta in Adelfio et al. (2014) tiene conto di entrambi gli aspetti. Tale nuova misura è stata oggetto di un'analisi condotta secondo un approccio di Regressione Quantilica (RQ) ad effetti casuali, dove la variabile risposta considerata è la nuova misura rilevata per ogni esame di ciascuno studente mentre quelle esplicative sono state il Genere, il Tipo di Diploma, la Residenza, il Voto di Diploma, lo Status di "In corso"-"Fuori corso" dello studente alla laurea, e il Corso di Laurea di immatricolazione. I dati hanno riguardato la coorte di immatricolati nel 2002 che si sono laureati entro 7 anni ai Corsi di Laurea (CdL) in Economia e Finanza e in Biologia dellUniversità degli Studi di Palermo (Italia). L'approccio scelto è stato preferito per due motivi: da un lato la RQ consente di studiare dipendenze che non siano solo "in media" ma anche in altri punti della distribuzione; dall'altro, l'approccio per misure ripetute ha consentito di tenere in considerazione la variabilità della performance intrinseca di ogni studente. I risultati hanno evidenziato che la variabile più significativa è stata il CdL, piuttosto che quelle socio-demografiche. Secondo l'ottica descritta in precedenza, il CdL scelto dallo studente potrebbe essere considerato come una proxy delle personali inclinazioni e attitudini, della sfera di interessi, capacità, abilità e indole dello studente: il suo "talento". Probabilmente la selezione e l'orientamento degli studenti verso un Corso di Laurea dovrebbero tenere in maggiore considerazione le reali aspettative, capacità e attitudini dei soggetti: questa potrebbe essere la chiave per ridurre gli abbandoni, conseguire il titolo nei tempi regolari e con voti migliori. In futuro, il presente studio sarà ripetuto su altre coorti e Corsi di Laurea al fine di valutare la robustezza dei risultati osservati.

Parole chiave *Performance dello studente - Indicatore - Regressione Quantilica ad effetti casuali*

1 Introduction

Academic student performance is a crucial issue for the university policy makers, today more than ever. The global job market needs more and more competitiveness, and high skills and competences, therefore an improved quality of the graduates is required. Obviously, the definition of graduate's quality is a difficult matter. It could concerns the actual student ability in solving practical problems, or the wideness and depth of his/her knowledge, or the number of years to graduation (in those countries where there are no time limits to get the degree), etc. or a combination of them. In this paper the attention is more devoted to the student performance at university, instead of his/her quality. But its measurement is really challenging too. Someone uses the final grade, others refer to the distribution of examinations marks, or pay attention to the credits earned (usually related to time, or marks, or both), or, again, to the time spent to get the degree. Or all of them. There is not a shared and acknowledge measure (and measurement methodology) of the student performance. In fact, literature offers several studies about student performance, mainly devoted to find its determinants, and, despite they are often based on the same measure, results address sometimes to different directions. Just to quote some of most recent ones, Cheesman et al. (2006) applied a regression analysis to describe students performance – measured by four categories of graduation marks – singling out that Gender, Enrolment status, Faculty, Finance assistance, and Residence are likely determinants. Tattersall et al. (2006) measured the educational efficiency in terms of comparison between inputs and outputs. The output-input ratio was analysed including several aspects of the students path to graduation, e.g. Learning interruption and Changes of the Study Programme (SP). The influence of the "Change of SP" on the expected time to graduation was also analysed by Adelfio and Boscaino (2016), highlighting its significant and negative effect. Birch and Miller (2006) used a Quantile Regression approach identifying in Tertiary Entrance Rank, Gender, and High School the most important determinants of high and low performance. Boscaino et al. (2007) focused on students who never earned credits after four years, using a Zero Inflated model and singling out different social demographic profile for different performance levels, based on Gender, High School, and Income level. Van Bragt et al. (2011) followed a more social psychological root making a deeper analysis of performances: they also included an ad-hoc survey to investigate the impact of the Big Five personality characteristics, Personal learning orientations, and students study Approach on their performance, using a logistic model. Results show a positive effect of Conscientiousness and a negative one of Ambivalence and Lack of regulation. Horn, Jansen, and Yu (Horn et al.) performed an exploratory analysis on the determinants of success of second-year students. The authors asked if the factors that leads the success at the end of the first year could rule the performance of the second year: they discovered that Lectures and Tutorial attendance were still significant factors, and the most important determinant was the performance during the first year. Attanasio et al. (2013) highlighted the crucial role of the credits earned at the end of the first year as a good and simple predictor of the success, in a retrospective exploratory study. Grilli et al. (2013) introduced Pre-enrolment assessment test outcomes, together with some personal student characteristics, on earned credits at the end of

the first year, using different models (hurdle, binomial mixture model): they highlighted the poor role on the Pre-enrolment test as predictor of number of credits. Adelfio et al. (2014) proposed a new measure for student performance, and Quantile Regression results showed no significant effect of the social and demographic variables, but just of the Attended SP.

In conclusion, at first sight, the literature generally suggests that the impact of the determinants varies (in terms of extent and direction) according to the context (economic, social, political, demographic, etc.) and results should hold just in that context.

This paper considers the Italian context, and aims to highlight a new study perspective: what if the real determinants for the good performance of the students were their motivations or attitudes, instead of their socio-demographic characteristics? That question arises from some considerations about particular results (here reported) obtained by a Quantile Regression approach on a recent new measure of student performance (Adelfio et al., 2014), aimed to identify its possible determinants.

Therefore, the paper is organised as follow: Section 2 recalls the new measure for the student performance, introduced in Adelfio et al. (2014); Section 3 is devoted to a brief description of the Quantile Regression model, used to investigate the determinant of the performance; in Section 4 some consideration arose from the results showed in section 3 are reported, and the random effect QR adopted model is illustrated in Section 5; last two Sections (6 and 7) discuss the results of the analysis and some remarks.

2 The measure of student performance

Student performance at university is often simply measured by the exam marks. But, obviously, same marks got in different classes play a different role for the final grade due to the different workload of the classes. The workload is usually measured by the credits. For the sake of simplicity, and because this paper refers to the Italian university context, next example is based on the Italian university grading system. In Italy, universities use a 30-point scale system with 0-17 as non passing grades and 18-30 as passing ones, with 30 the maximum (for outstanding results, the "lode" is added to the maximum in order to praise the real deserving student). Therefore, the grade 28 got in a course with 10 credits weights more then a 28 got in a 1 credit course. Indeed, the final grade is a weighted mean of the exam grades, with credits as weights. If we want to study the student performance, we have to use the couple grade-credits. Adelfio et al. (2014) proposed a new unique measure of the student performance, still based on grades and credits, that takes values in the same quantitative grading scale used by a given Country C:

$$m'_{ij} = \frac{best_C - suf_C}{max_j(m^w_{ij}) - min_j(m^w_{ij})} \times (m^w_{ij} - min_j(m^w_{ij})) + suf_C \tag{1}$$

where $best_C$ and suf_C are the marks that correspond to the best and to the minimum passing mark stated in the country C system, respectively; $m_{ij}^w = \frac{m_{ij} Cr_j}{\sum_{j=1}^{J} Cr_j}$, with m_{ij} is the grade got by the student i and Cr_j is the credit for the course j.

In such a way, each country can measure the performance of its students accounting both for marks and credits in one only measure (m_{ij}^w), and adopting an indicator that still gets values in the same original scale for passing marks (m'_{ij}).

Figure 1 illustrates the effect of the new indicator. Our data concerns the cohorts of 131 and of 98 students enrolled at the First Level Degree in Economy and Finance (E) and in Life Sciences (L), respectively, of the University of Palermo (Italy) in 2002, and graduated from 3 to 7 years after. The top and bottom plots on the left report the distributions of marks conditioned by credits for the two different degree courses (E and L, respectively). The E box plots show a more variability for the mark means across the credits than L ones. While the distributions of marks for L are highly negative asymmetric for all the credits, the distributions of marks for E are very different, showing high mean grades in correspondence to the low credits and lower mean grades for the higher credits (with the exception of the courses with 9 credits). The top right and bottom right plots show the effect of the new indicator (1): marks are rescaled accounting for the workload of the courses penalizing high grades got in low credit courses with respect to the low marks got for the high credits courses. The distributions of new marks are more fair and easily comparable.

The new indicator is the object of the analysis reported in this paper. First of all a Quantile Regression model is adapted to our data in order to investigate the determinants of the performance. Briefly, results give evidence to consider a new perspective of analysis and a random effect Quantile Regression model is considered.

3 The Quantile Regression model

The crucial aim of the paper concerns the investigation of the determinants of the student performance. We refer to the Quantile Regression (QR) approach in order to investigate the influence of some determinants over the whole shape of distribution of the proposed indicator (1).

QR (Koenker, 2005) deals with the estimation of conditional quantile functions, for models in which quantiles of the conditional distribution of the response variable are expressed as functions of the observed covariates. Whereas the method of Least Squares results in estimates that approximate the conditional mean of the response variable, QR aims at estimating either the conditional median or other quantiles of the response variable; QR also provides more robust estimates than the usual OLS based regression. Unlike the ordinary linear regression, the QR parameters measure the change in specified quantiles of the response variable produced by one unit change in the predictor variables. This allows comparing how some percentiles of the variable of interest may be more affected by certain subject characteristics than other percentiles.

In particular, from a more formal point of view, let $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$ be a sample of size n from some unknown population, where $\mathbf{x_i} \in \mathbf{R^d}$. The conditional ϕth quantile function $f_\phi(x)$ is defined such that $P(Y \leq f_\phi(X)|X = x) = \phi$, for $0 < \phi < 1$.

Therefore, the ϕth conditional quantile function can be estimated by solving:

$$\min_{f_\phi \in \mathbf{R}} \sum_{i=1}^{n} \rho_\phi \left(y_i - f_\phi(\mathbf{x}_i) \right) \tag{2}$$

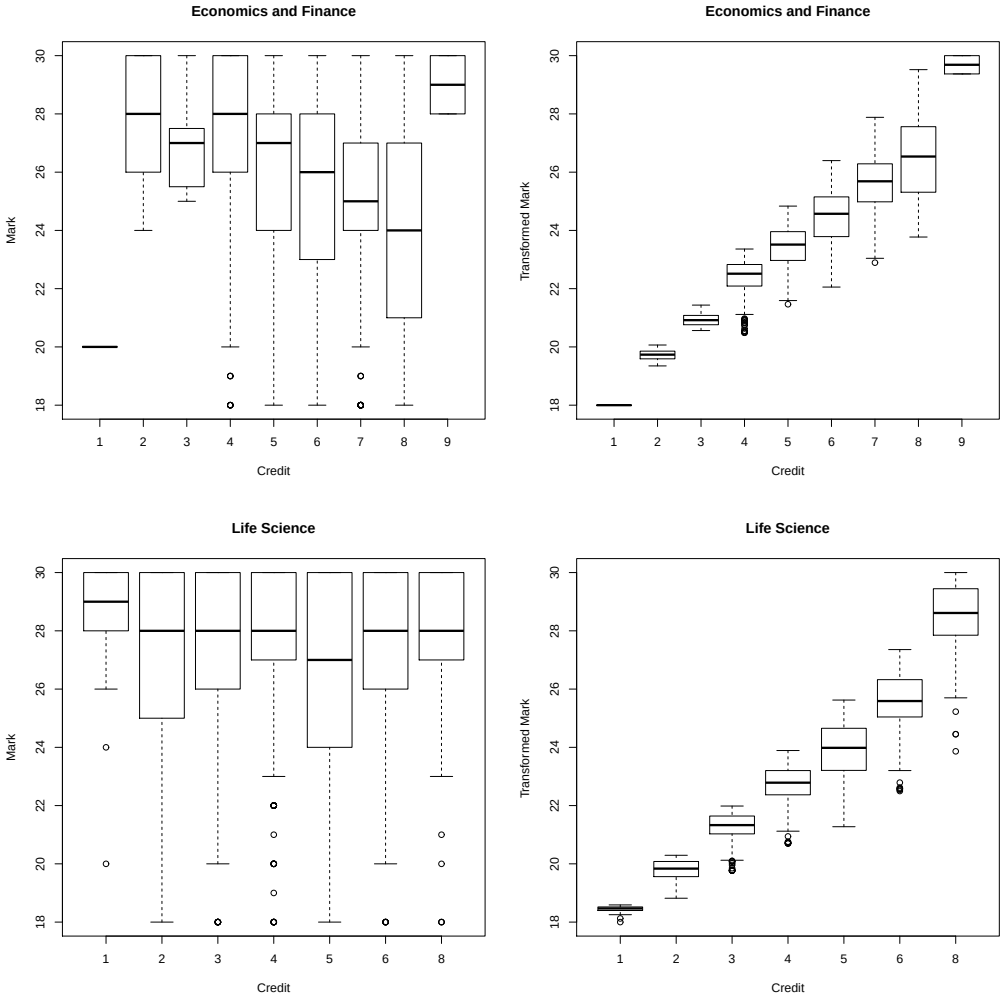


Fig. 1: Grades distributions MMM conditioned by credits, for E and L Study Programme

where the function $\rho_\phi(\cdot)$ is the tilted absolute value function, that yields the ϕth sample quantile as its solution and is defined by $\rho_\phi(r) = \phi r$ if $r > 0$, and $-(1-\phi)r$ otherwise (Koenker and Bassett, 1978).

Setting $f_\phi(\mathbf{x}) = \mathbf{x}^T\beta_\phi$ where $\beta_\phi = (\beta_{\phi,1}, \beta_{\phi,2}, \ldots, \beta_{\phi,d})^T$, such that the conditional ϕth quantile function $f_\phi(x)$ is a linear function of the parameters β, a linear quantile regression is considered.

The QR estimates are obtained by the R package `quantreg` (Koenker, 2012), that considers by default the modified version of the Barrodale and Roberts algorithm described in Koenker and d'Orey (1987) and Koenker and d'Orey (1994). We follow that approach because of the dimension of the analysed data set, since this approach is quite efficient for problems up to several thousand observations, computing also confidence intervals for the estimated parameters, based on inversion of a rank test (Koenker, 1994).

The Dataset introduced in Section 2 reports information about course credits and marks for each student, but it includes also other variables: Gender (Female vs Male), High School (Lyceum vs Not Lyceum), Residence (Palermo vs Not Palermo), and Diploma mark (centred at the mean). In addition, since in Italy the legal duration of the first level Study Programme (SP) is 3 years, but it is not compulsory (then, for example, a very unwilling student can get the degree 9 years after matriculation), we considered the Student Status at graduation as variable that classified as "On-Time" (OT) those students that get the degree in 3 years, and as "Out-of-time" (OOT) the others.

A preliminary linear QR model on $Me'_i = Me_i(m'_{ij})$ (not here reported for the sake of brevity) was performed on E and L separately, and it suggested that the only covariate with a significant effect was the Student Status at graduation. Therefore, we considered the linear QR model conditioned to the Status and results showed no significant variable with respect to the OT graduates. The performance of the students who get their degree on time do not seem to depend on the considered social and demographic features. It seems that good performer students are just good, and their positive performance may be ascribed to own motivation, inclination, method of study, etc.. Otherwise, the analysis of the OOT students (here reported for both E and L) showed some covariates are significant, only with respect to some percentiles. Conditioning to five quantiles $\tau = (0.05, 0.25, 0.50, 0.75, 0.95)$, the parameter estimations for the OOT students are reported in Table 1.

In short, the Intercept of the models represent the estimated conditional quantile of Me'_i distribution of students that are Female, with Lyceum diploma, living in Palermo, and with mean-centred diploma mark. As expected, the higher the quantile the higher the performance. The other values refer to the distribution of the estimated coefficients for different quantiles. Whereas the OLS results inform about the conditional mean of the response variable – given certain values of the predictor variables – QR aims to estimate the fixed quantiles of the response variable, using different measures of central tendency (and statistical dispersion), in order to obtain a more comprehensive analysis of the relationship between variables. Indeed, the QR analysis allows to interpret results also for the tails of the distribution (excellent and

Table 1: OLS and QR estimates for Me'_i for OOT graduates, in E and L cohort.

E cohort	OLS	τ_1	τ_2	τ_3	τ_4	τ_5
Intercept	25.28**	24.71**	24.95**	25.31**	25.56**	25.94**
High School	0.17	0.34*	0.27	0.07	0.06	0.13
Residence	-0.27**	-0.47**	-0.40**	-0.33*	-0.08	-0.10
Gender	-0.20	-0.61**	-0.39*	-0.12	-0.06	0.27*
Dipl. Mark	0.02**	0.02*	0.01	0.02**	0.02**	0.02**
L cohort						
Intercept	23.60**	22.66**	23.15**	23.80**	23.80**	24.23**
High School	-0.01	0.47**	0.032	-0.34	-0.10	-0.20
Residence	-0.35**	-0.31**	-0.28	-0.49*	-0.23	0.17
Gender	0.02	0.47**	0.11	-0.08	0.00	-0.18
Dipl. Mark	0.01	0.01**	0.01	0.00	0.00	0.02**

(** for $\alpha = 0.05$, * for $\alpha = 0.1$)

mediocre students), instead of focusing just on the "average student". More detailed comments about the QR results for E are in Adelfio et al. (2014), since in this section, we just highlight the comparison between E and L results. Focusing on L, the Intercept is steeper than the E cohort from the 5-th to 50-th percentile, but it gets always lower values. With respect to the covariates, for the L cohort all the variables are significant just around the 5-th percentile – that is the lowest performance students group – and the Diploma mark around 95-th percentile – the highest performance students group. L cohort shows a higher effect on performance than the E cohort and, in addition, a different role is played by the Gender: if for the E cohort males perform worse than females, for the L cohort we notice the reverse (but just conditioned to the significant 5-th percentiles).

Following a different point of view, we can study if there is a statistical difference among the estimated values of the coefficients for E and L. We restrict the comparison to the couples of significant coefficients, conditioned to the percentiles. In Table 2, the proportions of confidence intervals that do not overlap are reported for each covariate. As it is always true that if the confidence intervals do not overlap, then the statistics will be statistically different (Knezevic, Knezevic), it is possible to notice (tab. 2) that Gender has always a different effect (for E and L), conditional to the same significant percentiles. At the opposite, considering the Residence, it happens for just the 14% of the comparisons.

Table 2: Proportions of E and L non-zero-including confidence intervals that do not overlap, by covariate.

High School	Residence	Gender	Dipl. mark
0.50	0.14	1.00	0.25

4 Some reflections on QR results

Results showed in previous paragraph need some reflections. First of all, among the considered covariates, just the OOT Status seems to play a role on the student performance. Secondly, with respect to the OOT students, just a little effect of their characteristics is noticed, and in particular just for few quantiles. Therefore, performance could be affected by some not-here-considered social, demographic, economic, and/or by some latent student characteristics. In particular, E and L are two different SPs with respect to course subjects (in Italy, L courses subjects could be more specific than E ones) and to job market opportunities (in Italy, E gives more wide-ranging knowledge than L, hence E graduates can have access to a larger set of jobs than L graduates). Therefore, it is plausible that L and E students are different in motivation, basic knowledge, abilities, and aptitudes. We call these aspects with one unique convenient word: *talent*. Then, is the student *talent* a determinant for his/her university performance? In order to answer to this question, a different (from the usual regression) perspective has been followed. Instead of investigating the effect of the covariates on just one single synthetic measure of the students marks distribution, via the Quantile Regression approach it is possible to analyse the covariates effect on different parts (the quantiles) of that distribution. The *talent* will be highlighted by the comparison results between E and L students.

5 Quantile Regression for repeated measurements

In the light of the previous considerations, a a linear mixed QR approach is considered in order to take into account the whole mark distribution for each student. In such a way, it is possible to focus on the subject specific variability.

More formally, let $(\mathbf{x}'_{ij}, y_{ij})$, for $j = 1, ..., n_i$ and $i = 1, ..., N$, be repeated measurements data, where $\mathbf{x}'_{ij}$ are row p-vectors of a known design matrix and y_{ij} is the j-th measurement of a continuous random variable on the i-th subject.

According to the considered approach the linear mixed quantile functions of the response y_{ij} is:

$$G_{y_{ij}|u_i}(\tau|\mathbf{x}_{ij}, u_i) = \mathbf{x'}_{ij}\beta + u_i,\ j = 1, \ldots, n_i,\ i = 1, \ldots, N \tag{3}$$

where $G_{y_{ij}|u_i}(\cdot) \equiv F^{-1}_{y_{ij}|u_i}(\cdot)$ is the inverse of the cumulative distribution function of the response conditional on a location-shift random effect u_i (Geraci and Bottai, 2006). For this model,

the location-shift effects are assumed random and identically and independently distributed according to some density f_u, usually $u_i \sim N(0, \alpha)$, characterized by a τ-dependent dispersion parameter ($\alpha(\tau)$). Moving away from the penalized approach provided by Koenker (2004), Geraci and Bottai (2006) assume that y_{ij}, conditionally on u_i are independently distributed according as an Asymmetric Laplace Distribution (ALD):

$$f(y_{ij}|\beta, u_i, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right)\right\}$$

where $\mu_{ij} = \mathbf{x}'_{ij}\beta + u_i$ is the linear predictor of the τth quantile, fixed and known, and σ is the usual scale parameter. The random effects, that induce a correlation structure among observations on the same subject, are assumed to be independent. That is a likelihood-based approach to the estimation of the QRs based on the ALD and it is better then the penalized fixed effects based approach in terms of mean squared error of the QR estimators. Alternative models with non-normally distributed residuals were developed (Seltzer and Choi, 2002).

The ALD approach has been considered as it provides an automatic choice of the optimal level of penalization and also because it represents a suitable error law for the least absolutes estimator (and therefore a natural choice in QR).

6 Data Analysis via QR mixed model

We apply the linear QR model with a subject-specific random intercept that accounts for the within-group correlation (3) with respect to the two formerly considered cohorts of students graduated in E and L, managed together in a unique dataset. The analysis was performed using the R package lqmm: Linear Quantile Mixed Models (Geraci, 2014).

In Figure 2, results are reported with respect to the fixed coefficients estimates. To assess the suitability of (3), results are also commented in the light of those reported in the previous paragraphs. For each of the estimated coefficients we plot the QR estimates of the fixed parameters of (3), conditional to each quantile τ ($\tau = 0.05, 0.25, 0.50, 0.75, 0.95$), by the dashed curve with filled dots. These points may be interpreted as the impact of a unit-change of each covariate on the response variable, fixed the others. The grey area represents the 95% pointwise confidence band. The solid horizontal line, together with its 95% confidence intervals (horizontal dashed lines), refers to the estimate for a Linear Model with random intercept. The Intercept panel refers to the expected m'_{ij} (vertical axis) conditional to each quantiles (horizontal axis) for students that are Female, living in Palermo, with a Lyceum diploma, graduated On-time in Life Science, with a mean High school diploma mark equals to 90. We also added an interaction term between the student Status and the SP. The Intercept is steeper than the QR models without random effect (tab. 1): considering the m'_{ij} distribution rather than their median allows to appreciate the variation of the expected performance when we move between two consecutive quantiles. With the exception of the SP panel, other panels show no significant effect – in most of the quantiles – of the covariates. This could be a partial confirmation that

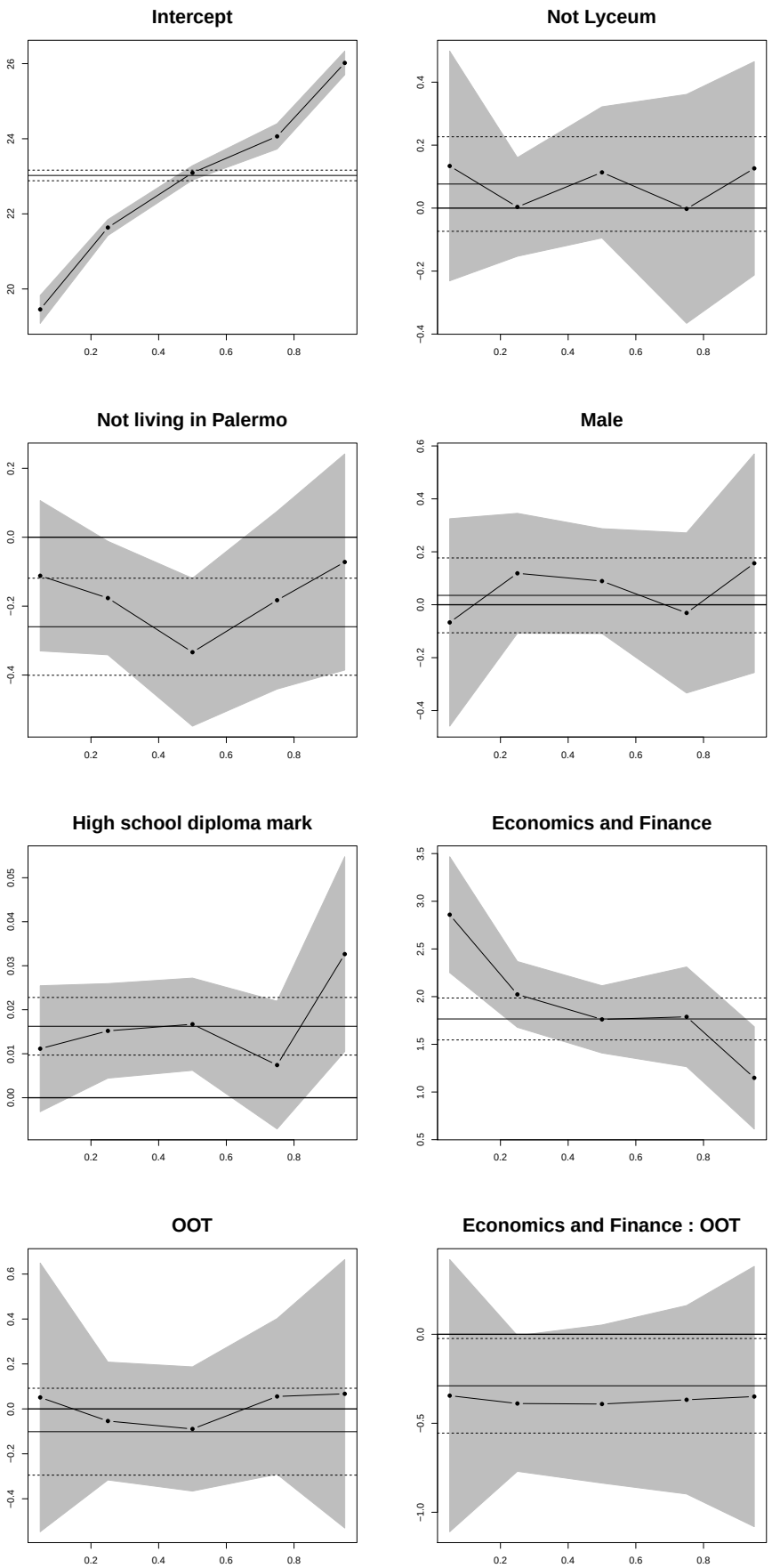


Fig. 2: Fixed coefficients estimates of random intercept QR model

Table 3: τ dependent estimated variance of u_i random intercept

τ	0.05	0.25	0.50	0.75	0.95
$\hat{\alpha}(\tau)$	0.367	0.391	0.398	0.574	0.613

student performance seems to be mainly due to students own *talent* rather than social and demographic characteristics.

The random intercept aims to catch the subject specific *talent* effect on performance. In fact, the τ dependent estimated variances of u_i random intercepts of (3) are all in $(0.36, 0.62)$ and they reflect the heterogeneity among students due to their own unobservable *talent* (tab.3).

Results from linear mixed QR, then, seem to confirm our previous assumption, that is the only variable that has a significant effect in the performance study is the SP choice. In other words the only relevant aspect from this analysis is the choice of the SP, easily ascribable to the personal inclinations and subjective sphere of any student (the *talent*).

7 Final remarks

The necessity of investigating the effect of possible social-demographical variables to the overall performance of university students is widely recognized, although the literature offers results that are dependent on the analysed context, both in space and time.

This paper is based on first results reported in Adelfio et al. (2014): these led to a new investigation perspective, for analysing the determinants of the students performance.

For this reason the choice of students about the SP has been considered in order to catch the effect of a latent variable for their subjective characteristics (such as aptitude, motivation, or inclination) on the overall performance.

In order to have comparable results, also among different SPs, we needed a measure like the one in (1) that accounts in a unique quantity both for the workload and marks. Results from the linear mixed QR seem to confirm our previous assumption: in fact the only variable that has a significant effect on the performance is the SP.

In other words, the chosen SP could be a proxy of the personal inclinations and subjective sphere of any student.

In the light of these, though ascribed to two big SPs of the University of Palermo, in order to improve the students performance, probably policy makers should look at the student personal abilities, aptitudes, motivation, and *talent*, instead of just at the usual social and demographic characteristics. Therefore, following this perspective, student orientation and tutoring activities, and proper selective University entry-tests could be useful tools to address students to the appropriate Study Programme (and, therefore, to help them in following their own *talent*).

As future work, more Sps will be considered to assess the robustness of the current results.

References

Adelfio, G. and G. Boscaino (2016). Degree course change and student performance: a mixed-effect model approach. *Journal of Applied Statistics, 43*, 3–15.

Adelfio, G., G. Boscaino, and V. Capursi (2014). A new indicator for higher education student performance. *Higher Education 68*(5), 653–668.

Attanasio, M., G. Boscaino, V. Capursi, and A. Plaia (2013). May the students career performance helpful in predicting an increase in universities income. pp. 9–16. series in studies in classification, data analysis, and knowledge organization, Switzerland Springer International Publishing: Statistical models for data analysis.

Birch, E. R. and P. W. Miller (2006). Student outcomes at university in australia a quantile regression approach. *Australian Economic Press 45*(1), 1–17.

Boscaino, G., V. Capursi, and F. Giambona (2007). The careers' performance of a university students' cohort. Technical report, DSSM Working paper, n. 2007.1.

Cheesman, J. S., N. Simpson, and G. Wint (2006). *Determinants of student performance at university Reflections from the Caribbean.*

Geraci, M. (2014). Linear quantile mixed models the lqmm package for laplace quantile regression. *Journal of Statistical Software 57*(13), 1–29.

Geraci, M. and M. Bottai (2006). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics 8*, 140–154.

Grilli, L., C. Rampichini, and R. Varriale (2013). Predicting students academic performance a challenging issue in statistical modelling. CLEUP: Cladag 2013 Book of abstracts.

Horn, P., A. Jansen, and D. Yu. Factors explaining the academic success of second-year economics students an exploratory analysis. *South African Journal of Economics 79*(2).

Knezevic, A. *Overlapping Confidence Intervals and Statistical Significance. StatNews n.73.* Cornell Statistical Consulting Unit. Cornell University.

Koenker, R. (1994). Confidence intervals for regression quantiles. In M. Huskova (Ed.), *Asymptotic Statistics Proceedings of the 5th Prague Symposium on Asymptotic Statistics .*, pp. 349–359. Heidleberg Physica-Verlag.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*, 74–89.

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

Koenker, R. (2012). quantreg quantile regression. *R package version 4.9*.

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

Koenker, R. and V. d'Orey (1987). Computing regression quantiles. *Applied Statistics 36*, 383–393.

Koenker, R. and V. d'Orey (1994). Remark as r92. a remark on algorithm as. *Applied Statistics 229*, 410–414.

Seltzer, M. and K. Choi (2002). Model checking and sensitivity analysis for multilevel models, in n. duan and s. In N. Duan and S. Reise (Eds.), *Multilevel modeling Methodological advances, issues, and applications*. Hillsdale, NJ Lawrence Erlbaum.

Tattersall, C., W. Waterink, P. Hoppener, and A. Koper, R. (2006). case study in the measurement of educational efficiency in open and distance learning. *Distance Education 27*, 391–404.

Van Bragt, C. A. C., A. W. E. A. Bakx, T. C. M. Bergen, and M. A. Croon (2011). Looking for students personal characteristics predicting study outcome. *Higher Education 61*, 59–75.

UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 'in fase di assegnazione', volume I, 2017

Credit demand and supply shocks in Italy during the Great Recession

Andrea Cipollini · Fabio Parla

Abstract In this paper, we use Structural VAR analysis to disentangle credit demand and supply shocks and their effect on real economic activity in Italy during the 2008-2014 crisis period. The three endogenous variables considered are the loan interest rate, the loans growth rate and the employment to population ratio. The data are observed at annual frequency for each of 103 Italian provinces. The structural shocks are identified through heteroscedasticity, by letting the variance of the shocks to switch across four Italian macro-regions: North, Centre, South and Islands. Sign restrictions are used to interpret ex post the structural shocks. The empirical findings suggest a more important role of credit supply shocks in shaping the level of real economic activity. Furthermore, the results show that credit crunch hits the North of Italy less than the remaining macro-regions, especially the South-Italy.

Keywords Structural VAR · Identification through heteroscedasticity · Credit shocks · Regional economic activity

Riassunto *Una caratteristica predominante della Grande Recessione è stata una contrazione prolungata del credito al settore privato in diversi paesi. Lo scopo di questo studio, basato su dati disponibili a livello provinciale durante il periodo di crisi 2008-2014, è duplice. In primo luogo, ci concentriamo sull'identificazione degli shocks dal lato della domanda e dell' offerta di credito.*

Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo
Viale delle Scienze, Ed. 13, 90128 Palermo
E-mail: andrea.cipollini@unipa.it

Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo
Viale delle Scienze, Ed. 13, 90128 Palermo
E-mail: fabio.parla@unipa.it

In secondo luogo, siamo interessati ad analizzare l'impatto degli shocks dal lato dell'offerta di credito sull'attività reale per l'economia italiana.
Numerosi studi empirici basati su micro-dati studiano l'impatto di una contrazione del credito sull'attività reale dell'economia. Per quanto riguarda l'economia italiana, lo studio di Barone et al. (2016) usa una strategia di identificazione che si basa su dati che catturano le relazioni banca-province e necessita di un periodo campionario che consenta di confrontare un periodo di tranquillità con uno di crisi. Il periodo campionario usato nel presente studio è relativo solo ad un prolungato periodo di crisi. Inoltre, contrariamente allo studio di Barone et al. (2016), l'analisi non si basa su una stima a due stadi, dove, nel primo, viene identificato un indicatore di offerta di credito a livello provinciale e, nel secondo stadio, si esamina l'impatto sull'attività reale dell'economia.
In questo studio, viene utilizzato un modello Vettoriale Autoregressivo Strutturale considerando tre variabili endogene: il tasso di interesse sui prestiti, il tasso di crescita dei prestiti ed il tasso di occupazione osservati annualmente, tra il 2008 ed il 2014, per 103 provincie italiane. Gli shocks strutturali di domanda e offerta di credito ed il loro impatto reale sono simultaneamente identificati attraverso la presenza di eteroschedasticità osservata nei dati a livello di quattro macro-regioni italiane: Nord, Centro, Sud e Isole. Le restrizioni di segno vengono utilizzate per interpretare ex-post gli shock strutturali. I risultati empirici suggeriscono un ruolo più importante degli shock dal lato dell'offerta di credito sul livello di attività economica reale. Inoltre, i risultati mostrano che la crisi del credito colpisce il Nord dell'Italia meno delle rimanenti macro regioni, in particolare del Sud-Italia.

Parole chiave *Modello Vettoriale Autoregressivo Strutturale - Identificazione tramite eteroschedasticità - Shocks al mercato creditizio - Attività economica regionale*

1 Introduction

A predominant feature of the Great Recession has been a prolonged contraction of credit to the private sector in a number of countries. The aim of this paper, based on provincial level data, is twofold. First, we focus on the identification of credit demand and supply shocks in explaining the credit contraction in Italy. Second, we are also interested in analyzing the effects of the identified credit supply shock on real activity for the Italian economy.
The slowdown in bank lending which occurred in many advanced economies has led to a debate about the effects of disturbances in credit markets on business cycles. In spite of the increasing importance of capital markets, the Euro financial system is typically bank-based. Furthermore, bank loans play a non-negligible role in the financing of private investment and consumption in the European countries. The Italian financial system has been dominated by banks: the ratio of loans to non-financial private sector to Italian banks total assets was 76.8 percent in 2014. Hence, bank lending might play an important role in explaining fluctuations of economic cycle. In the aftermath of the financial crisis, the Italian banking system has seen a slackening growth

of bank loans to non-financial corporations and households. The Italian year-on-year growth rate of loans to private sector fell from 9.2 percent in the first quarter of 2008 to 1.1 percent in the first quarter of 2010. After a sharp upturn, the growth rate has become negative since the second quarter of 2012.

A number of empirical studies based on micro-data informing on bank-firm relationship employs the methodology proposed by Khwaja and Mian (2008) to identify credit supply shocks (see Bonaccorsi di Patti and Sette, 2016, for the Italian economy, among the others). The Khwaja and Mian (2008) methodology exploits a sample which includes observation for a pre and post crisis period, and it is based on the estimation of a regression of the change in the loans provided by each bank to its borrowing firms after an exogenous shock (e.g. a crisis event) as a function of bank exposure to that shock. Del Giovane et al. (2017) use Bank Lending Survey (BLS) to identify, through zero exclusion restrictions, the simultaneous equation system fitted to interest rates and loans data for Italy.
While the previous studies are only interested in the identification of a credit supply factor, some authors are also concerned with their real effect using a two-stage estimation analysis. Cingano et al. (2016) use data on bank-firm relationship and they identify credit supply shocks through the variation in bank reliance on the interbank market at the end of 2006, leading to different bank exposure to the July 2007 liquidity shock. The proxy used by Cingano et al. (2016) for the real activity is the private investment. The study of Barone et al. (2016) use bank-province relationship data for the Italian economy to identify a local (province) credit supply indicator. In a second stage of the analysis they assess the impact of credit supply shock on investment, value added and employment.

The Khwaja and Mian (2008) methodology employed by Cingano et al. (2016) and by Barone et al. (2016) relies on an individual bank exposure to an exogenous shock (e.g. crisis event) switching from a no crisis period to one characterized by turmoil. Since we focus only on a prolonged crisis time span, we exploit the heterogeneity in the data across Italian macro-regions. More specifically, ex-ante, we employ identification through heteroscedasticity (see Rigobon, 2003; Lanne and Lütkepohl, 2008) and, ex-post, we give an economic interpretation to the shocks through sign restrictions (see Mumtaz et al., 2015, for a review). In particular, we follow the suggestion of Kick (2016) in setting the sign restrictions: a credit supply (demand) shock moves the price and quantity of credit in opposite (same) directions.
Both methods are popular for the identification of a Structural VAR, which is the model used in this paper. Moreover, we argue that, contrary to the previous studies which rely on a two-stage analysis, our study, based on an estimation in one-shoot, does not suffer from a measurement error affecting the use of an estimated regressor in the second stage regression.

The empirical findings show that credit supply shocks play a more important role than innovations to demand for credit. Furthermore, there is evidence that credit crunch hits the

North of Italy less than the remaining macro-regions, especially the South-Italy.

The paper is structured as follows. Section 2 provides a literature review on identification of shocks to credit markets; Section 3 describes the empirical methodology; Section 4 describes data and the empirical findings and Section 5 concludes.

2 Literature review

As mentioned in the introduction, a number of empirical studies on the Italian credit crunch is only interested in the identification of credit supply shocks. Presbitero et al. (2014) relies on the identification of constrained Italian firms, using firms survey data containing information on loan applications and bank decisions. The authors main focus is the role played by functional distance between the loan office and the headquarters where final lending decisions are made to explain the tightening in lending conditions in Italy. For this purpose, the authors combine survey data on firms with aggregate data on banks informing on the openings and closures of branches at the bank-province level. The authors use a sample of monthly observations from 2008:1 to 2009:4 and the empirical findings show that the credit crunch experienced in Italy after Lehman Brothers collapse has been more severe in provinces with larger shares of branches owned by distantly managed banks. Moreover, there is evidence of a home bias, given that the credit crunch has not been harsher for small and economically weak firms.

The identification methodology put forward by Khwaja and Mian (2008), which is based on Credit Register data for firms that have multiple lenders, has been applied by a number of studies. This approach consists of estimating a regression of the change in the loans provided by each bank to its borrowing firms after an exogenous shock (e.g. a crisis event) as a function of bank exposure to that shock. For this purpose, they use firm fixed effects to capture shifts in the demand for loans and other unobservable borrower characteristics, such as changes in their balance sheet conditions. The identification methodology provides an estimate of the differential change in credit supply for the same firm, associated with a different exposure of the lending banks to the exogenous shock. Albertazzi and Marchetti (2010) present evidence of a contraction of credit supply associated to low bank capitalization and scarce liquidity, over the 6-month period following the Lehman bankruptcy. Bofondi et al. (2013) exploit the differential exposure to the sovereign risk between domestic banks and foreign banks operating in Italy. The authors find that the lending of domestic banks grew less (and their interest rates were higher) than that of foreign banks, after the outbreak of the sovereign debt crisis. Bonaccorsi di Patti and Sette (2016) link banks' balance sheet conditions to the provision of credit and show that Italian banks that relied heavily on securitization prior to the subprime crisis curtailed lending more than other banks.

Del Giovane et al. (2017) estimate a system of two simultaneous equations regarding the interest rates and loan amounts of 11 Italian banks. The authors use demand and supplies dummies obtained from Eurosystem Bank Lending Survey (BLS).[1] In order to identify the simultaneous equations system, the demand dummies are excluded from the equation where the dependent variable is the price and the supply factor dummies are excluded from the equation involving quantity. After a number of robustness checks, the authors acknowledge that they cannot exclude the possibility that their findings are affected to some extent by some residual endogeneity. The authors find that the effects of the supply restriction on both the cost and the availability of credit were, on average, stronger during the sovereign debt crisis than during the Lehman global crisis. Moreover, the authors find that credit crunch was mostly related to the banks' risk perception during the global crisis, whereas funding conditions became predominant during the sovereign debt crisis.

The second empirical issue regarding the impact of the identified credit supply shock on real activity of the Italian economy has been addressed by the following studies. The study of Cingano et al. (2016) use in the first stage the Khwaja and Mian (2008) identification methodology. In particular, the authors use data on bank-firms relationships and they identify credit supply shocks through the variation in bank reliance on the interbank market at the end of 2006, leading to different bank exposure to the July 2007 liquidity shock. The authors' findings show that although credit tightening was homogeneous across firms, investment fell by a much larger amount among smaller and younger firms, and those with higher bank dependence. Bottero et al. (2015) show that the Greek bailout in 2010 led to a fall in loan supply in Italy, which depressed investment and employment for smaller Italian firms.
The methodology suggested by Greenstone et al. (2014) is used to identify and assess the real effects of a credit supply indicator by Barone et al. (2016). The authors use confidential data over 2008-2011, obtained from the Bank of Italy Supervisory Report, on total outstanding loans extended by Italian banks to the private sector (firms and households) aggregated into local credit markets corresponding to provinces. The identification strategy employed by Barone et al. (2016) is based on data capturing bank-provinces relationships (hence it is similar to Khwaja and Mian, 2008). More specifically, the authors focus on the identification of a local (province) credit supply indicator by, first, using a panel regression. The dependent variable is the change in credit granted by one of the 650 banks to households and firms located in a given province and operating in a given economic sector and the explanatory variables are two dummies. The first dummy measures province-year fixed effects that capture the variation in the change of lending due to local economic factors (capturing local demand). The second

[1] Ciccarelli et al. (2015) use Bank Lending Survey (BLS) for the Euro area and the Senior Loan Officer Survey (SLOS) for the U.S. Contrary to Del Giovane et al. (2017) study which employs BLS data for Italy only to identify credit demand and credit supply shocks, Ciccarelli et al. (2015) are also interested in the real effects of credit supply shocks. The qualitative data are transformed into quantitative and treated as endogenous variables together with proxies of output, prices and monetary policy rates in a Vector Autoregression model, VAR, fitted to the Euro area and to the US separately.

dummy measures bank-year fixed effects which identifies nationwide bank lending policies. The authors, then, use the coefficient associated to the second dummy and pre-crisis bank market shares in the province (as weights) to aggregate and to construct a province-year credit supply index. In a second stage, the credit supply real effect are estimated regressing either value added, or investment, or employment (observed for each province) on the estimated local credit supply variable. The empirical findings show that the most severe effect of the credit crunch occurred in the North and Central Italy which have firms relatively more dependent on external finance. The methodology suggested by Greenstone et al. (2014), is also employed by Berton et al. (2017), using a matched data set of job contracts, firms and banks in one Italian region (Veneto). The authors, first, identify and construct a credit supply factor at firm level, and, in a second stage, they assess the impact on employment. The empirical findings (for Veneto region) show that the effects of the credit crunch have been particularly severe for smaller, younger and less productive firms, and those with higher debt overhang and weaker bank-firms relationships have been more vulnerable to the (negative) impact of the credit crunch.
Dörr et al. (2017) use information on loans by individual banks to firms that borrow from multiple Italian banks, which are exposed to foreign borrowers in distressed countries (Greece, Ireland, Portugal and Spain). The authors use a novel identification method suggested by Amiti and Weinstein (2017) which does not rely on a comparison between access to credit during pre-crisis and a crisis period (as in Khwaja and Mian, 2008), but only on loan data over the 2010-2012 period characterized by Euro sovereign debt crisis. The credit supply and demand components are recovered by imposing an additional constraint. The adding-up constraint states that changes in individual loan growth between banks and firms must add up to the overall, economy-wide change in loan growth. After establishing that credit supply shocks reduce firms' loan growth, Dörr et al. (2017) show that credit supply rationing had significant real effects on firms' investment and employment decisions, as well as total factor productivity. Italian firms with higher exposure to troubled banks reduced their investment and employment and they experienced a significant fall in productivity.

Recent empirical studies on the Italian economy (together with Euro area countries) employ macro-time series data and they identify credit supply shocks and their impact on the real economy by imposing sign restrictions to identify a Structural Vector Autoregression model, SVAR. In particular, Bijsterbosch and Falagiarda (2014) use time-varying parameter Vector autoregression model with stochastic volatility, producing results for Euro area countries, including Italy. The studies of Hristov et al. (2012), based on Panel VAR and the study of Kick (2016), based on Global VAR, analyze the dynamic effect of credit supply on real economic activity in Italy as well as a number of Euro area countries.

3 Structural VAR

In this section, we first describe the identification through heteroscedasticity methodology. The first study of identification of structural shocks via changes in volatility is due to Rigobon (2003). Recently, the studies of Lanne and Lütkepohl (2008), Lütkepohl (2012) and Lütkepohl and Netsunajev (2015) show that heteroscedasticity in residuals provides over-identifying restrictions (which can be tested) to traditional SVAR models employed to study the effect of monetary policy shocks. Lütkepohl (2012) identifies shocks by considering changes in volatility in given time periods (with breakpoints specified exogenously). The author considers also a vector generalized autoregressive conditional heteroscedasticity (MGARCH) to model for changes in volatility of residuals. Finally, a third specification model examines changes in volatility by using a Markov regime switching process. Lütkepohl and Netsunajev (2015) use a SVAR to estimate the interaction between US monetary policy and stock market where the identification is obtained by modelling heteroscedasticity in a way similar to Lütkepohl (2012), considering also smooth transition in the variances.

Following Lütkepohl (2005), we carry out with a SVAR analysis, estimating a structural B-model VAR(1) for pooled data, which has the following reduced form representation:

$$Y_{n,i,t} = \delta + AY_{n,i,t-1} + u_{n,i,t} \tag{1}$$

where $Y_n = (interest\ rate_{i,t}, \Delta loans_{i,t}, empl.ratio_{i,t})$ is a vector of three variables in province i at time t, namely interest rate on loans ($interest\ rate$), a log transformation of loans first order difference ($\Delta loans$) and the employment to population ratio ($empl.ratio$), δ is a 3×1 vector of constant terms, A is a 3×3 parameter matrix and u is a vector of residuals with covariance matrix $E(u_t u_t') = \Sigma_u$, which is not assumed to be diagonal.
According to Lütkepohl (2005), the relationship between the white-noise process and the structural disturbances has the following representation:

$$u_{n,i,t} = B\varepsilon_{n,i,t} \tag{2}$$

where B is a non-singular 3×3 matrix including the contemporaneous interactions between the endogenous variables and ε is a vector of uncorrelated structural shocks.
Hence, the structural form of VAR is:

$$Y_{n,i,t} = \delta + AY_{n,i,t-1} + B\varepsilon_{n,i,t} \tag{3}$$

We estimate a reduced-form of a VAR(1) by Ordinary Least Squares (OLS) for each equations separately.
In an attempt to estimate the model, we need to establish different regimes of volatility. This allows the determination of the covariance matrix structures as well as identifying the system of equations.
Regimes of volatility are selected on the basis of geographical discrimination. Particularly, four

heteroscedastic regimes are defined, corresponding to different Italian macro-areas: North Italy, Central Italy, South Italy and Insular Italy.
The sample of observations is divided into 4 sub-samples, based on geographical characteristics, $S = (S_{North\ Italy}, S_{Central\ Italy}, S_{South\ Italy}, S_{Insular\ Italy})$.
Constructing the covariance matrix structures is carried out by choosing the North Italy as the first regime, whereas the other regimes are: *(i)* Central Italy, *(ii)* Southern Italy and *(iii)* Insular Italy.

The covariance matrix structure has the following representation:

$$\Sigma_1 = BB', \quad \Sigma_i = B\lambda_i B', \qquad i = 2, \ldots, 4 \tag{4}$$

where

$$\Sigma_1 \quad \text{for } i \in S_{North\ Italy} \quad \text{and} \quad \Sigma_i = \begin{cases} \Sigma_2 & \text{for } i \in S_{Central\ Italy} \\ \vdots & \\ \Sigma_4 & \text{for } i \in S_{Insular\ Italy} \end{cases} \tag{5}$$

Once the reduced form of VAR(1) model is estimated by OLS estimation, the corresponding residuals are used in order to estimate the unknown parameters.
The set of unknown parameters includes matrix B coefficients and the variances of the structural error terms.
Assuming normality of the error terms, the structural parameters are obtained by Maximum Likelihood (ML) estimation. The Multivariate Gaussian log-density function at time t and for macro-region i is:

$$log\ l(\beta, \sigma) = -\frac{KT}{2}\ log(2\pi) - \frac{1}{2}\sum_{i=1}^{4} |log(\Sigma_i)| - \frac{1}{2}\sum_{i=1}^{4}(u_i'\Sigma_i^{-1}u_i) \tag{6}$$

where Σ_i is the covariance matrix of the reduced-form residuals, expressed in terms of the structural form coefficients as described in (4) and (5)[2]. As mentioned above, identification through heteroscedasticity is only a statistical tool, and to give an economic interpretation of the structural form shocks we use, ex post, sign restrictions on each column of the impact multiplier matrix B (see Table 1).
The economic identification of credit demand and supply shocks is based on a minimal set of identifying restrictions in line with previous studies (Peersman, 2011; Barnett and Thomas, 2013; Kick, 2016). More specifically, a negative credit demand shock reduces both credit price and the amount of loans. Conversely, a negative credit supply shock produces an increase of

[2] The log density functions are generated by using the **mvtnorm** package in R. The optimization problem is solved by minimizing the negative of the sum of the log densities by using the "BFGS" method. The "BFGS" method is a quasi-Newton method which uses function values and gradients to build up a picture of the surface to be optimize.

the loan interest rate as well as reducing the quantity of bank lending (see Hristov et al., 2012). The real variable is affected by credit supply and demand shocks negatively. Following Kick (2016), we do not expect any prior sign restriction from the responses of the credit variables to the real shocks.
Since the number of unknowns is equal to eighteen and the number of moment conditions (see (4) and (5)) is equal to twenty-four equations, a Likelihood Ratio test is employed to test for the six over-identifying restrictions:

$$LR = -2[ln(\hat{\theta}_R) - ln(\hat{\theta}_{UR})] \tag{7}$$

where $\hat{\theta}_R$ is the ML estimator of the restricted model and $\hat{\theta}_{UR}$ is the ML estimator of the unrestricted model. Under the null hypothesis, the Likelihood ratio statistic has an asymptotic χ^2 distribution with degree of freedom equal to the number of the over-identifying restrictions. We also compute the cumulative standardized impact of each structural shock over two year horizon by estimating $B + AB$. While the standard errors of the parameters of the standardized impact multiplier B are retrieved from inversion of the Hessian of the maximized log-likelihood function, the confidence intervals for the cumulative impulse response are generated through bootstrap. In particular, for each regime, we resample 1000 times the estimated residuals of the VAR(1)[3]. For each draw, we estimate the parameters of the structural form model by maximizing the log likelihood function.
Finally, for each of 103 Italian provinces, we compute the historical decomposition of the endogenous variables as follows:

$$Y_t = \delta \sum_{j=0}^{t-2} A^j + A^{t-1} Y_1 + \sum_{j=0}^{t-2} A^j B \varepsilon_{t-j} \quad , \quad \text{for } t > 1 \tag{8}$$

where $Y_1 = Y_{2008}$ in our analysis. Constructing the historical decomposition allows us to compute the anticipated and unanticipated components of each series.

4 Empirical analysis

4.1 Data

We use a panel data set of observations which contains information on credit aggregates and a real variable for 103 Italian provinces.

For the purpose of disentangling credit supply shock from the demand-side one, we consider two credit market aggregates and one real activity variable. Hence, as endogenous variables, we use as proxies of price and quantity of credit the loan interest rate and the amount of

[3] We keep only the replications (which are 421) in line with the ex post identification of the shocks according to the point estimation results.

loans, respectively; the employment to population ratio is the proxy of real economic activity. The data are at annual frequency, from 2008 to 2014, for each of 103 provinces, for a total of 2163 observations. We use low-frequency data because of the availability of the employment to population ratio: for each province, data are only made accessible with annual frequency. The shortness of the sample period used is due to the loan interest rate series which starts from 2008.

Information on credit aggregates are from the Statistical Database of Bank of Italy. As for the price of credit, we use the lending rates on loans facilities (stock) series for non-MFI resident sectors. Particularly, we consider the interest rate charged by banks at the end of the fourth-quarter as annual observation.
As for the quantity of credit, we consider the first-order difference of loans to non-MFI resident sectors as endogenous variable[4]. In an attempt to include in our model annual observations instead of quarterly data, we consider the value of loans registered at the end of each fourth-quarter. Taking into account the first difference allows us to avoid stationarity problems.
The real aggregate is the employment rate which is defined as the ratio between employed people (aged 15-64) and the corresponding overall resident population. The data are collected from statistical database of the Italian National Institute of Statistics (ISTAT).
Actually, ISTAT makes available Gross Value Added (GVA) data which might be used as a proxy for real economic activity at provincial level. Nonetheless, the value added series is not available for 2014.

Since we seek to identify credit supply and demand shocks, and a real shock, through cross sectional heteroscedasticity, we consider four macro-regions: North Italy, Central Italy, South Italy and Insular Italy.
Fig. 1-2-3 show the boxplots series of the three endogenous variables for each Italian macro-area from 2008 to 2014. The boxplots provide information on each province which belongs to different macro-regions.
Focussing on the mean values of Figure 1, all the Italian macro-regions exhibit the same pattern in the loans interest rate. After a twofold decrease over the 2008-2010, the loan interest rates stabilize around values ranging from 2.8 and 3.5 percent, before exhibiting a temporary upturn in 2011. Afterward, the interest rate on loans values do not exceed 3.7 percent in the 2012-2014 period.
Figure 2 shows a more heterogeneous evolution over time of the loan growth rates by inspection of the boxplots for the macro-regions. Whilst the highest loan growth rate is in South and Insular Italy at the beginning of the crisis, these regions experience the strongest slowdown in the growth rates, starting from 2011. During the last two years of the sample, there is a clear

[4] According to the definitions provided by the Bank of Italy, the loans aggregate is defined as the loans disbursed by banks to non-bank sectors. This variable includes mortgage loans, current account overdrafts, loans secured by pledge of salaries, credit card advances, discounting of annuities, personal loans, leasing, factoring, other financial investment (e.g. commercial paper, bill portfolio, pledge loans, loans granted from funds administered for third parties), bad debts and unpaid and protested own bills.

evidence of a recovery in the loans growth rates.
In Figure 3, we can observe that during the period 2008-2014, all the four macro-areas exhibit a relevant decline of the employment to population ratio, with different levels of decrease in the territorial areas. Whilst North and Central Italy experience a moderate reduction in the employment to population ratio until 2012 and a moderate upturn in 2013-2014, the South and Insular Italy manifest a significant negative trend during the whole period.

4.2 Empirical Evidence from structural VAR

The estimated parameters of the standardized impact multiplier are shown in Table 2 (panel A).
Whilst residuals heteroscedasticity is a statistical tool to identify structural form shocks, ex post interpretation is obtained using the sign based restriction suggested in Table 1. Therefore, according to the sign based restriction, the first, second and third column show the standardized impact of a negative shock to credit demand, credit supply, and real economy, respectively. While the credit demand shock plays a bigger role than the credit supply shock on the loan interest rate, the reverse is true as for the impact on the loan growth rate. Although, on impact, the only statistically significant effect of an innovation to credit demand and credit supply is the one on the interest rate on loans (at 1 percent and 10 percent level of significance, respectively), results from Table 3 show a statistically significant cumulative effect of credit demand and credit supply shocks to both credit aggregates. Moreover, the empirical findings show that credit supply shocks plays a more important role than those to credit demand in reducing the employment to population ratio. In particular, a one standard deviation shock to credit supply implies, on impact, a 1.3 percent change in the employment to population ratio (see Table 2 panel A) and a cumulative impact over a two year horizon equal to 2.4 percent (see Table 3).
The real shock, interpreted as a negative one, due to its marginal depressing effect on the employment to population ratio, raises both the interest rate on loan and the growth in lending. The impact of the real shock on the employment rate is statistically significant over a two year horizon.
Table 2 (panel B) shows that the identification assumption is satisfied because all the estimated parameters, λ_i, measuring the estimated relative variances, are distinct and statistically significant.
The interpretation of the results in Table 2 (panel B) is based on the square root of the relative variances, in order to focus on the magnitude of shocks relative to the one for the North of Italy. The innovation hitting credit supply in Central, South and Insular Italy are all above unity suggesting that credit crunch hits the North of Italy less than the remaining macro-regions. In particular, the largest relative magnitude is observed for the credit supply shock hitting the South of Italy, and the magnitude of the innovation to credit supply in Central and Insular Italy is almost the same. While the largest relative magnitude of the credit demand shock ob-

served is for Insular Italy, the South and Central Italy exhibit credit demand innovation with magnitude lower than the North. Finally, the largest magnitude of the real shock is observed in Central Italy (almost twice than the one for the North). Both South and Insular Italy exhibit a magnitude of the real shock above the corresponding one for the North (although much lower than the one for Central Italy).
Finally, Table 2 (panel C) shows the results of the over-identifying test restrictions using the LR statistic. The value of the log likelihood of the restricted model is equal to 2382.14, whilst the unrestricted log likelihood is equal to 2383.10. Therefore, the over-identifying restrictions are not rejected at 90 percent confidence level.

Following Lütkepohl (2011), we carry out with a historical decomposition (see Fig. 4 and 5) in order to analyse the effects of credit market shocks on the real variable in the Italian provinces. Our main focus is on the contribution of credit demand and supply shocks to the dynamic of the employment to population ratio (de-meaned at provincial level) for each macro-region. A credit demand shock seems to play a non-relevant role (with the exception of Insular-Italy) in explaining the downturn in the employment to population ratio.
We can observe, from historical decomposition, that credit supply shock plays an important role in tracking the dynamics of the employment rate in each macro-region, especially the slackening in employment rate in South and Insular Italy over 2013-2014.

To summarize, contrary to the empirical findings of Kick (2016), we find that credit supply shocks play a more important role than innovations to demand for credit for the dynamics of real economic activity in Italy. Our results are in line with previous papers which focus on the credit crunch effect on real economy across Italian provinces (see Presbitero et al., 2014; Barone et al., 2016; Cingano et al., 2016; Berton et al., 2017). In particular, our findings about regional differences of credit crunch are in line with the ones of the study of Presbitero et al. (2014) who find that the real economy of North Italy is more resilient to credit rationing, since, especially in the Southern regions, banks retracted disproportionally from markets that are more distant from their headquarters. Since our study shows that the Centre and South of Italy exhibit a relative higher magnitude of the credit supply shock, this contrasts the findings of Cingano et al. (2016) related to the territorial impact of rationing in lending. The authors find that the credit cut has been relatively homogeneous across borrowers and the firms with easier access to external finance or with a stronger liquidity position were more able to contain the negative consequences for investment (and, to less, extent on employment) of the drop in credit. Moreover, our findings contrast those from Barone et al. (2016) who find that the most severe credit rationing impact on real value added growth, during the recent financial crisis, occurred in the North and Central Italy which have firms relatively more dependent on external finance.

5 Conclusions

In this paper, we have investigated the role of credit market shocks in explaining the downturn of the Italian economic activity using data at provincial level over 2008-2014. A number of studies of the Italian credit crunch are based on a two-stage estimation approach where in the first stage a credit supply indicator is identified through the Khwaja and Mian (2008) method which requires data on either bank-firms or bank-provinces relationships, observed in a pre and post crisis period. However, since our dataset is constrained only to a period of prolonged recession, our identification scheme is based on the changing variance of the structural shocks to a VAR fitted to interest rates, loans growth rates and employment ratio observed in the Italian macro-regions. Heteroscedasticity is only a statistical tool for the purpose of identification, therefore we have used ex post sign restrictions suggested by theory to identify demand and supply of credit shocks.
Differently from the empirical findings of Kick (2016), we find that credit supply shocks play a more important role than innovations to demand for credit. Our findings related to a sizable and significant effect of credit supply on employment are in line with the studies, based on loans to Italian firms, of Barone et al. (2016), Cingano et al. (2016) and Berton et al. (2017).
Moreover, the empirical evidence shows that credit crunch hits the North of Italy less than the remaining macro-regions, especially the South-Italy. This findings are consistent with those of Presbitero et al. (2014) who find that the real economy of North Italy is more resilient to credit rationing, since, especially in the Southern regions, banks retracted disproportionally from markets that are more distant from their headquarters.
An implication of these findings for Italy is that a key policy priority should therefore take into account the significant role of the credit supply. Taken together, these findings support the implementation of the recent Quantitative Easing adopted by the ECB to stimulate the economy.

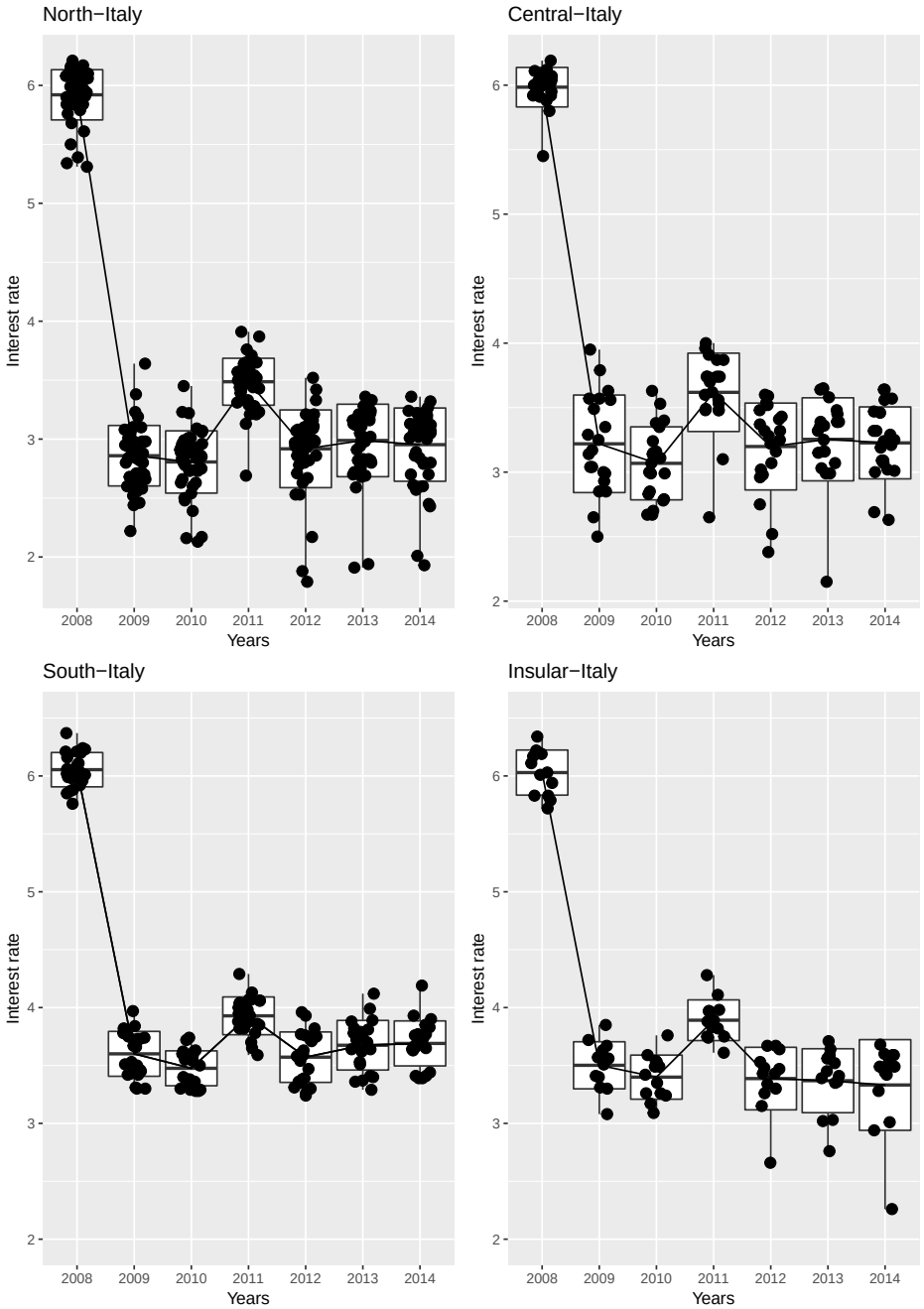


Fig. 1: Boxplots for Loans interest rates, percent, for the Italian macro-areas, 2008-2014.

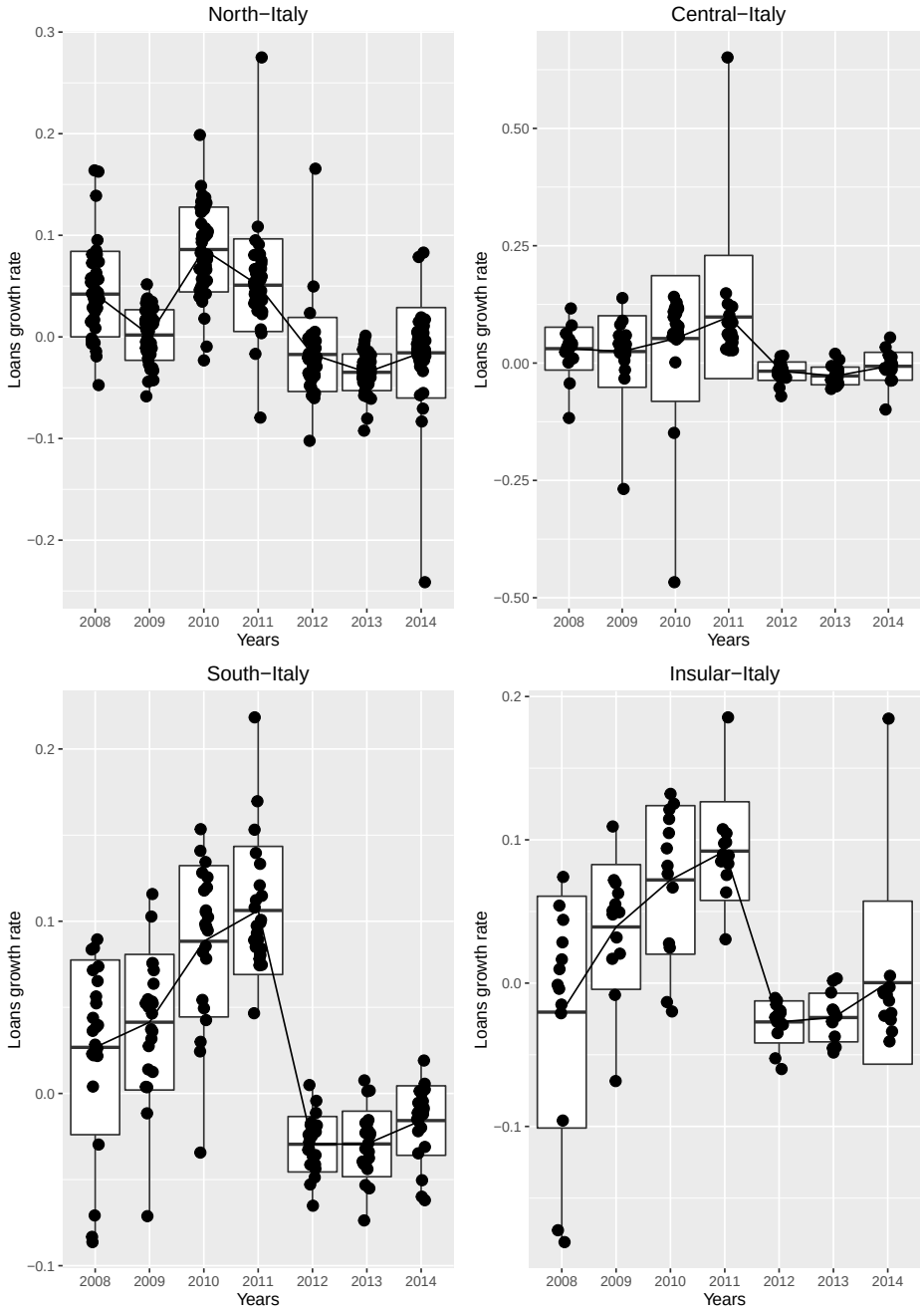


Fig. 2: Boxplots for Loans growth rates, for the Italian macro-areas, 2008-2014.

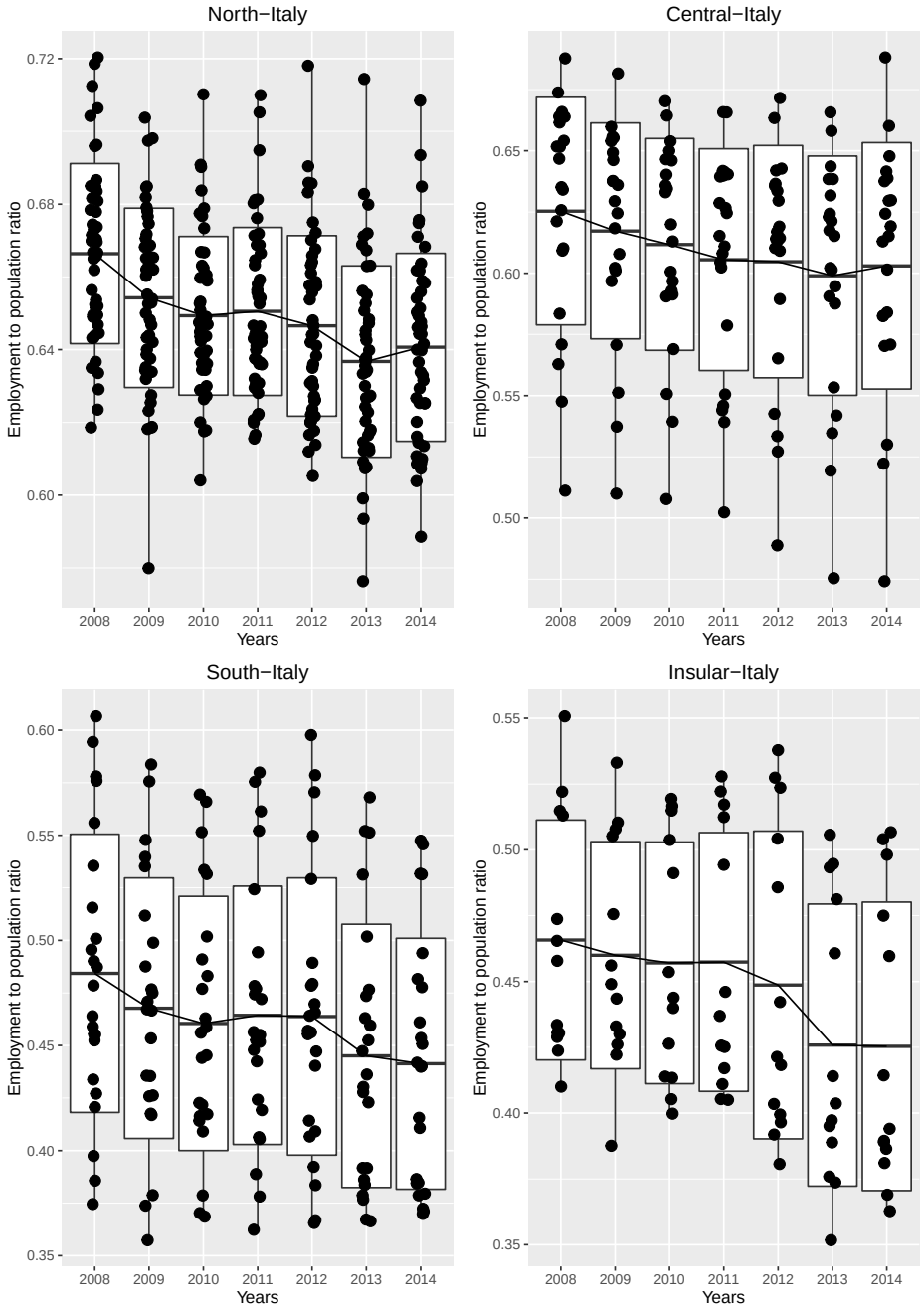


Fig. 3: Boxplots for Employment to population ratio, for the Italian macro-areas, 2008-2014.

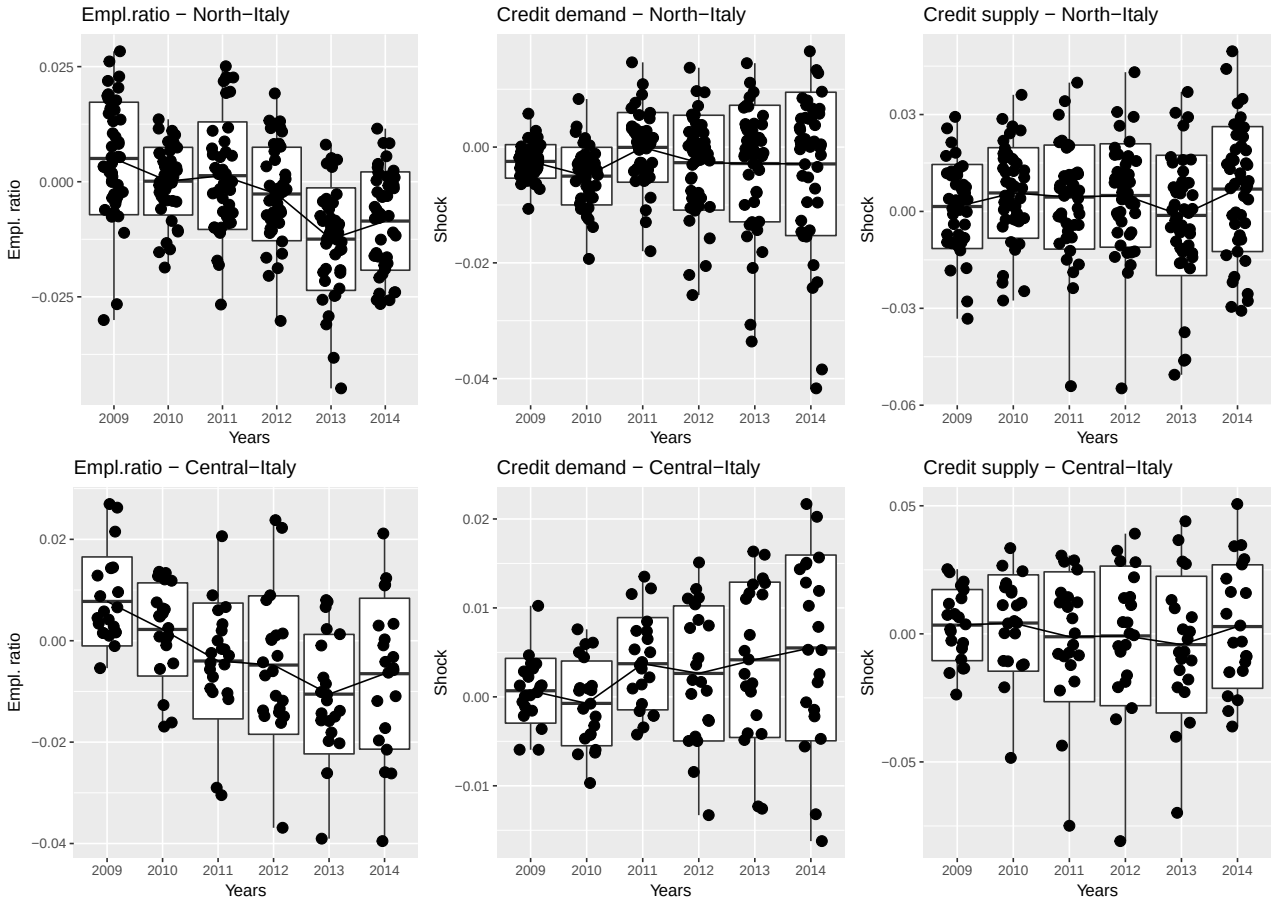


Fig. 4: Contribution of credit demand and supply shocks on historical decomposition of Employment rate, North and Central Italy, 2009-2014.

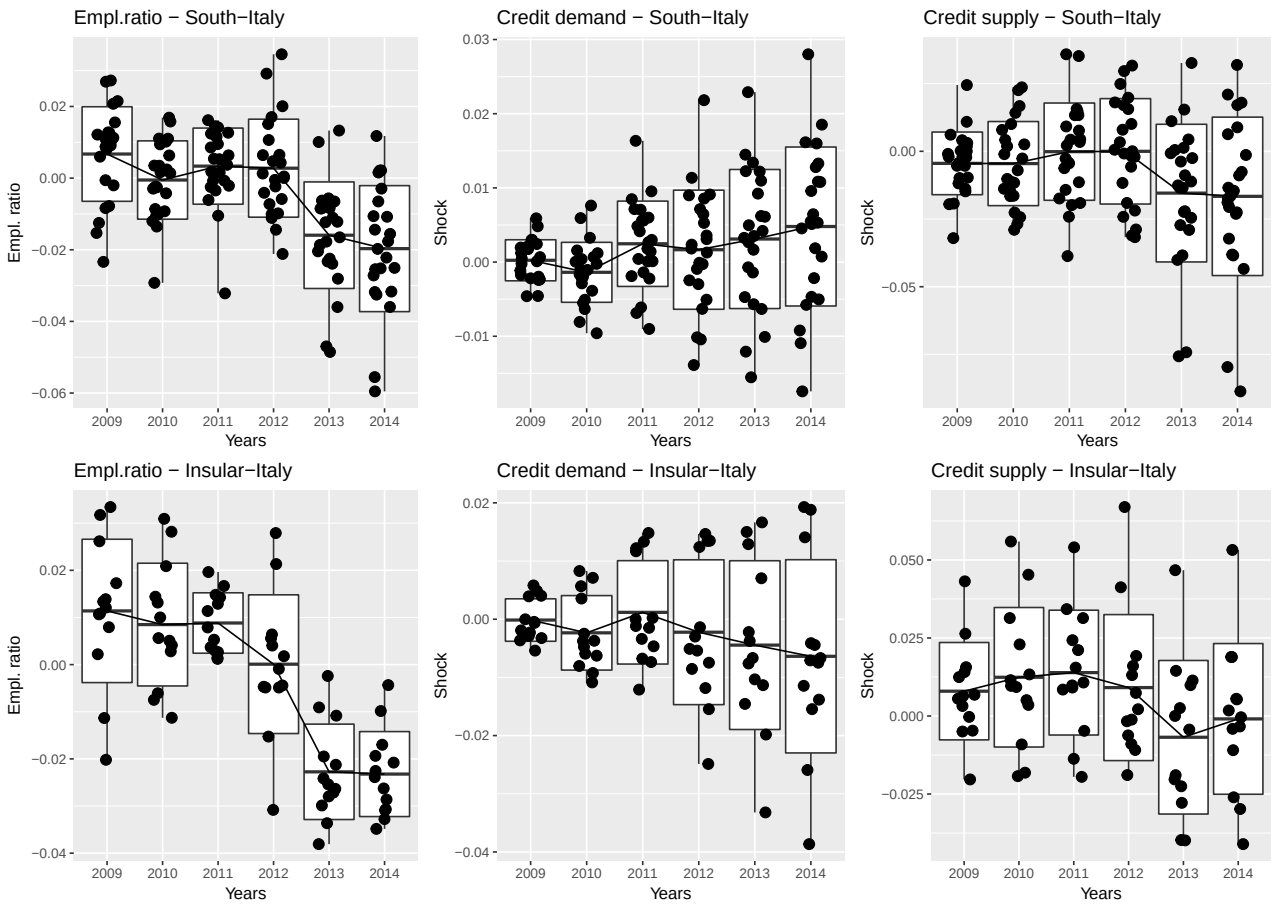


Fig. 5: Contribution of credit demand and supply shocks on historical decomposition of Employment rate, South and Insular Italy, 2009-2014.

Table 1: Theory-driven ex post sign restrictions on B matrix

Impact on	*Credit demand shock*	*Credit supply shock*	*Real shock*
interest rate	-	+	***n.a***
$\Delta loans$	-	-	***n.a***
empl. ratio	-	-	-

Note: Here the sign restrictions are related to negative shocks

Table 2: Maximum Likelihood Estimation results of B and λ matrices.

Panel A: Standardized Impact multiplier (B matrix).

	Credit demand shock	Credit supply shock	Real shock
interest rate	-0.323^{****}	0.073^{*}	0.044^{*}
	(0.019)	(0.042)	(0.024)
$\Delta loans$			
	-0.001	-0.008	0.054^{****}
	(0.005)	(0.006)	(0.002)
empl. ratio			
	-0.004^{*}	-0.013^{****}	-0.002
	(0.002)	(0.0013)	(0.001)

Panel B: Relative variances and magnitude of the shocks.

	Parameter	Magnitude
Central Italy		
CREDIT DEMAND SHOCK	0.933^{****}	0.966
CREDIT SUPPLY SHOCK	1.238^{****}	1.113
REAL SHOCK	2.963^{****}	1.721
South Italy		
CREDIT DEMAND SHOCK	0.657^{****}	0.810
CREDIT SUPPLY SHOCK	1.404^{****}	1.185
REAL SHOCK	1.234^{****}	1.111
Insular Italy		
CREDIT DEMAND SHOCK	1.405^{****}	1.185
CREDIT SUPPLY SHOCK	1.207^{****}	1.099
REAL SHOCK	1.075^{****}	1.037

Signif. codes: 0 $'****'$ 0.001 $'***'$ 0.01 $'**'$ 0.05 $'*'$ 0.1 $' '$ 1

Panel C: Likelihood Ratio Test.

LR Test		Log-likelihood		Value
	Unrestricted model	2382.138	LR statistic	1.931
	Restricted model	2383.104	p-value	0.926

Note: All the parameters are estimated by ML. Asymptotic standard errors are provided in brackets. The relative variances (see panel B) are obtained setting to unity the elements on the first regime structural covariance matrix main diagonal, here refereed to the Northern Italy. The magnitude are obtained by taking the square root of the relative variances (e.g. the parameters in the second column).

Table 3: Cumulative Impact over a two year horizon.

	Mean	Lower bound	Upper bound
Credit demand shock			
interest rate	-0.291	-0.346	-0.262
$\Delta loans$	-0.008	-0.013	-0.002
empl. ratio	-0.010	-0.014	-0.004
Credit supply shock			
interest rate	0.123	0.066	0.168
$\Delta loans$	-0.013	-0.021	-0.005
empl. ratio	-0.024	-0.027	-0.022
Real shock			
interest rate	0.097	0.067	0.117
$\Delta loans$	0.060	0.055	0.066
empl. ratio	-0.004	-0.008	-0.001

Note: The First column is the mean value of bootstrapped distribution of the Cumulative Impact over a two year horizon, the last two columns are 16 percent and 84 percent bootstrapped confidence interval bounds.

Acknowledgements We would like to thank participants to the Seventh Italian Congress of Econometrics and Empirical Economics - ICEE 2017 (University of Messina, 2017) and the Essex Finance Centre (EFiC) 2016 Conference in Banking and Finance (University of Essex, 2016) for helpful comments and suggestions. The usual disclaimer apply.

References

Albertazzi, U. and D. J. Marchetti (2010). Credit supply, flight to quality and evergreening: an analysis of bank-firm relationships after Lehman. Temi di discussione (economic working papers) 756, Bank of Italy.

Amiti, M. and D. E. Weinstein (2017). How much do idiosyncratic bank shocks affect investment? evidence from matched bank-firm loan data. Journal of political economics, *forthcoming*.

Barnett, A. and R. Thomas (2013). Has weak lending and activity in the United Kingdom been driven by credit supply shocks? Bank of england working paper no. 482, Bank of England.

Barone, G., G. de Blasio, and S. Mocetti (2016). The real effects of credit crunch in the great recession: evidence from Italian provinces. Temi di discussione della banca d'italia no. 1057, Bank of Italy.

Berton, F., S. Mocetti, A. F. Presbitero, and M. Richiardi (2017). Banks, firms and jobs. Working paper no. 136, Mo.Fi.R.

Bijsterbosch, M. and M. Falagiarda (2014). Credit supply dynamics and economic activity in Euro area countries: a time-varying parameter var analysis. Working paper series no. 1714, European Central Bank.

Bofondi, M., L. Carpinelli, and E. Sette (2013). Credit supply during a sovereign debt crisis. Temi di discussione (economic working papers) 909, Bank of Italy.

Bonaccorsi di Patti, E. and E. Sette (2016). *Did the securitization market freeze affect bank lending during the financial crisis? Evidence from a credit register*. Journal of Financial Intermediation 25, 54–76.

Bottero, M., S. Lenzu, and F. Mezzanotti (2015). Sovereign debt exposure and the bank lending channel: impact on credit supply and the real economy. Temi di discussione (economic working papers) 1032, Bank of Italy.

Ciccarelli, M., A. Maddaloni, and J.-L. Peydro (2015). *Trusting the bankers: A new look at the credit channel of monetary policy*. Review of Economic Dynamics 18(4), 979–1002.

Cingano, F., F. Manaresi, and E. Sette (2016). *Does Credit Crunch Investment Down? New Evidence on the Real Effects of the Bank-Lending Channel*. Review of Financial Studies 29(10), 2737–2773.

Del Giovane, P., A. Nobili, and F. M. Signoretti (2017). *Assessing the Sources of Credit Supply Tightening: Was the Sovereign Debt Crisis Different from Lehman?* International Journal of Central Banking June 2017.

Dörr, S., M. Raissi, and A. Weber (2017). Credit-supply shocks and firm productivity in italy. IMF Working Paper No 17/38, International Monetary Fund.

Greenstone, M., A. Mas, and H.-L. Nguyen (2014). Do credit market shocks affect the real economy? quasi-experimental evidence from the great recession and normaleconomic times. Working Paper No 20704, NBER.

Hristov, N., O. Hülsewig, and T. Wollmershäuser (2012). *Loan supply shocks during the financial crisis: Evidence for the Euro area*. Journal of International Money and Finance 31(3), 569–592.

Khwaja, A. I. and A. Mian (2008). *Tracing the impact of bank liquidity shocks: Evidence from an emerging market*. The American Economic Review 98(4), 1413–1442.

Kick, H. (2016). *Spillover effects of credit demand and supply shocks in the EU countries: Evidence from a structural GVAR*. Unpublished manuscript.

Lanne, M. and H. Lütkepohl (2008). *Identifying monetary policy shocks via changes in volatility*. Journal of Money, Credit and Banking 40(6), 1131–1149.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer-Verlag Berlin Heidelberg.

Lütkepohl, H. (2011). Vector autoregressive models. Economics working papers eco 2011/30, European University Institute.

Lütkepohl, H. (2012). Identifying structural vector autoregressions via changes in volatility. Discussion papers, DIW Berlin.

Lütkepohl, H. and A. Netsunajev (2015). Structural vector autoregressions with heteroskedasticity - a comparison of different volatility models. Sfb 649 discussion paper 2015-015, Humboldt University.

Mumtaz, H., G. Pinter, and K. Theodoridis (2015). What do vars tell us about the impact of a credit supply shock? Working papers 739, Queen Mary University of London, School of Economics and Finance.

Peersman, G. (2011). Bank lending shocks and the Euro area business cycle. Working papers 11/766, Ghent University, Faculty of Economics and Business Administration.

Presbitero, A. F., G. F. Udell, and A. Zazzaro (2014). *The home bias and the credit crunch: A regional perspective*. Journal of Money, Credit and Banking 46(s1), 53–85.

Rigobon, R. (2003). *Identification through heteroskedasticity*. Review of Economics and Statistics 85(4), 777–792.

UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 'in fase di assegnazione', volume I, 2017

Saving process within a zero waste strategy in Sicily: a system dynamics approach

Andrea Cuccia

Abstract Disposable society represents a way of living that is not sustainable anymore, mainly in a context like our planet where resources are really scarce. As theorized by Paul Connett, the guru of zero waste strategy, there is no waste until that material still contains a drop of usefulness. The core point of the paper is the implementation of a zero waste strategy in a small municipality in Sicily. By building up a SD model we can figure out how saving process connected to the decrease of amount of waste piled up in the landfill, has been boosting this virtuous cycle, which is bound to cope with the citizens claims about environmental care and lower tax burden level.

Keywords saving process · zero waste strategy · Dynamic Performance Management · System Dynamics · accountability

Riassunto

La sostenibilità può essere vista come la più importante sfida che l'umanità si appresta ad affrontare in questo millennio; infatti, guardando agli attuali livelli di consumo, sarebbero necessari almeno due pianeti se si assumesse a riferimento il modello di consumo europeo, quattro pianeti, invece, se si considerasse il modello di consumo americano. In questo contesto, "Rifiuti Zero" si configura come un obiettivo ad un tempo pragmatico e visionario in grado di spingere la collettività a rispettare la sostenibilità immanente nei cicli naturali e a valorizzare i materiali di scarto destinandoli ad ulteriori utilizzi. Il programma "Rifiuti Zero" è stato concepito da Paul Connett, professore di chimica presso la Santa Lawrence University, come alternativa all'incenerimento quale tradizionale metodo di trattamento dei rifiuti. In particolare, facendo

Phd student in "Model based public planning, policy design, and management",
University of Palermo E-mail: Missing

leva sulla responsabilità industriale e sul coinvolgimento attivo della collettività – previa sensibilizzazione della stessa grazie ad opportune campagne condotte dalle Amministrazioni Locali – il programma mira a ridurre alla fonte la produzione di rifiuti e ad estendere il più possibile il ciclo di vita dei prodotti grazie a pratiche di riciclo e di riuso. L'adozione di tale strategia, a partire dal 2012, da parte del Comune di Palazzo Adriano va interpretata come una scelta dettata da vincoli di bilancio sempre più stringenti e da obblighi normativi comunitari e nazionali sempre più incalzanti. Tale effetto combinato ha spinto il Comune a comprimere il più possibile i costi di smaltimento dei rifiuti attraverso la ricerca della conformità ad un obbligo comunitario-nazionale in merito alla frazione di raccolta differenziata da conseguire, nella prospettiva di soddisfare, al contempo, un interesse collettivo quale quello alla salubrità ambientale. La soglia del 65% come percentuale di rifiuti urbani differenziati raggiunta dal Comune di Palazzo Adriano nel 2015, era già stata formalizzata nell'art. 205 del D.Lgs. n. 152/06 come obiettivo da perseguire entro il 2012 e va oggi inquadrata in combinato disposto alla Direttiva Comunitaria n.98/2008 che ha fissato un nuovo obiettivo relativo alle pratiche di riciclo e riuso dei rifiuti urbani, espresso in termini di peso, da raggiungere entro il 2020. La Dynamic Perfomance Management (DPM) Chart, articolata in strategic resources, perfomance drivers ed end-results, diventa cornice ideale per la costruzione di un modello di system dynamics che si pone come integrazione ai tradizionali supporti informativi patrimoniali-finanziari nella valutazione dell'efficacia dell'azione pubblica. Il modello sviluppato illustra il funzionamento del sistema di smaltimento e le diverse possibili ramificazioni del ciclo di vita del prodotto "rifiuto" (la "waste management chain"), eleggendo l' area di raccolta ottimale (ARO) "Valle del Sosio" quale soggetto incaricato di gestire il servizio, sebbene quest'ultimo debba ancora entrare in funzionamento in sostituzione dell'ATO PA2, l'azienda pubblica responsabile del servizio fallita nel 2014. La DPM Chart si configura come plancia di comando volta a porre ad ordinamento le diverse variabili in gioco in termini di punti nodali di un reticolo di relazioni causali (causal loops); tale reticolo è destinato a rappresentare la performance dell'Amministrazione Pubblica, intesa sia come processo sia come risultati finali prodotti, enfatizzando, nel caso di specie, il ruolo critico assunto dalla pubblicizzazione da parte del Comune del risparmio conseguito – in ragione della riduzione di rifiuti conferiti in discarica – nel raggiungimento della sovrammenzionata soglia del 65%. Infine, l'avvento dell'ARO, come realtà consortile che coinvolge i Comuni dell'area "Valle del Sosio", unitamente al sistema di consorzi (quali COREPLA e COREVE) preposti alla lavorazione dei diversi materiali di imballaggio, potrebbe sancire l'inossidabile saldatura fra le due leve del programma "rifiuti zero", la responsabilità industriale e il coinvolgimento della collettività. Se da un lato, infatti, le lavorazioni effettuate dai singoli consorzi diventano prerequisito indefettibile per riconsentire uno sfruttamento industriale dei materiali di imballaggio; dall'altro lato, le compensazioni – alimentate dai contributi CONAI versati da produttori e utilizzatori per gli ulteriori imballaggi immessi nel mercato – per i maggiori costi legati alla raccolta differenziata, conferite dal sistema di consorzi all'ARO e poi traslate sui singoli Comuni in termini di minori tasse per i cittadini, finirebbero per alimentare ulteriormente la persistenza nel tempo di tale schema virtuoso.

Parole chiave *risparmio, strategia rifiuti zero*

1 Introduction

Sustainability might be seen as the most important challenge mankind is going to face in this millennium; in fact, given the ongoing consumption rate, we ought to live at least in two planets assuming the European consumption model, whereas if we assumed the American one, we would need at least 4 planets[1]. By comparing with this backdrop, *« Zero Waste is a goal that is both pragmatic and visionary, to guide people to emulate sustainable natural cycles, where all discarded materials are resources for others to use*[2]*»*. Zero Waste program might be seen asa philosophy, a strategy, and a set of practical tools seeking to eliminate waste, not to manage it. It was conceived for the first time by Paul Connett, professor of chemistry at the Saint Lawrence University, who has been very renowned throughout the world since he has stood against incineration as the widely-accepted treatment method of waste by proposing at the opposite a different way of planning resources life cycles so that all the products are intended to be reused (cradle to cradle scheme, instead of the overwhelming linear cradle to grave scheme).

By following this perspective, it has been obtained the main purpose of zero waste strategy: not to figure out more sophisticated methods to destroy waste, vice vice versa, to encourage production of products and packaging materials that will never be destroyed[3].

The change in the paradigm this strategy is trying to foster, is essentially summed up in the 3-R scheme: reduction at the source and recycling and reuse, just to extend the life cycle of products.

More deeply, this program is conventionally made up of 10 steps:

1s Source separation, made by citizens on their own just to let the intrinsic value of each product persist and not to make them contaminated by the blend with other different items;

2d Door to door collection, made by employers paid by the municipality;

3 Composting, arising from the organic waste fraction and intended to be use as natural fertilizer for the soil;

4 Recycling;

5 Ruse, Repair& Community Center;

6 Vaste reduction initiatives, for example forbidding the selling in the supermarket of disposable plastic dishes;

7 Economic incentives, to make people aware that making waste the least possible is first of all useful to keep the cost of waste management service down;

[1] To get further information, read J. Moore and W. E. Rees, "Getting to One-Planet Living", chapter 4 in "State of the World: is sustainability still possible?", The WorldWatch Institute, Island Press, 1 edition, 15th april 2013, p.41.

[2] The whole definition is available on the official websitewww.zerowasteeurope.eu/about/principles-zw-europe/[2017].

[3] P. Connett, "*Rifiuti zero, una rivoluzione in corso",* Dissensi Editor, 2012, p.11.

8 Residual separation & research center, that is supposed to act on what has been impossible to retrieve in the earlier steps;

9 better industrial design (namely industrial responsibility);

10 Temporary landfill, given that disposal infrastructure such as landfills or incinerators should no longer be built and be progressively phased out as prevention & recycling rates increase[4].

Surely the visionary aim of Zero Waste within 2020, requires two pillars to be pursued:

• **Engaging community**, since it becomes crucial to undertake public campaign just to make people receptive to this kind of subject and to invite them to adopt waste free practices;

• **Industrial responsibility**, as just a guarantee of designing long-lasting, easily maintainable and repairable products, of reducing packaging and redesigning those products that cannot be safely reused, recycled and composted. Lastly, industrial responsibility consists also of reusing parts and material coming from discarded products and material in line with a circular economy where every "waste" output of one process becomes an input for another such that the utility of the material is maximized. In this sense it might be seen as a further R to add to the traditional 3-R scheme, that boosts both recycle and reuse rate by improving the traditional design of products[5].

Adopting such a policy for a small municipality like Palazzo Adriano[6], that is made up of less than three thousands of in habitants, might be seen as a wise choice tackling the issue of intergenerational equity.

Recently municipalities in Italy have been involved in the grip of the fiscal compact, in line with European diktats and the tough economic and financial situation. As a consequence, they have been approaching the need of a spending review aimed at squeezing the management costs as much as possible. Fiscal compact, as fiscal chapter of the Treaty on Stability, Coordination and Governance in the Economic and Monetary Union, signed on 2 march 2012 formalized the need for governments to keep as sustainable their public finances and to prevent a general government deficit just to safeguard the stability of the euro area as a whole; accordingly, it also requires the introduction of specific rules, including a "balanced budget rule" and an automatic mechanism to take corrective action. This rule has been cascaded on municipalities forcing them to hold down expenditures level just to stabilize their financial equilibrium.

Obviously transfer of waste in landfill implies an incurring cost, referred to the incineration system as last treatment stop for the waste piled up in the landfill. In the last years, in accordance with the fiscal compact and with a view of achieving scale economies, it has been spread the idea of conferring waste management system to public consortia of services called ARO (Optimal Collection Area), whose costs will be splitted up among the different partic-

[4] P. Connett, "*The Zero Waste Solution: untrashing the Planet One Community at a Time*", Joni Praded Editor, 2013, p.15 and following pages.

[5] J.M. Simon, "*Stirring paper*"; Second Conference on Economic Degrowth for Ecological Sustainability and Social Equity March 26-29th 2010, Barcelona, pp. 1-2

[6] Palazzo Adriano is a little town in Sicani Mounts, western Sicily, belonging to the Metropolitan Area of Palermo

ipating municipalities. Specifically, the municipality of Palazzo Adriano will be merged with the municipalities of Bisacquino, Prizzi, Chiusa Sclafani and Giuliana, in the "ARO-Valle del Sosio", assuming that «*they generally represent a rural and mountainous area with a low population density*[7]». This consortium has been introduced by Regional Law n. 3/2013 but it has not begun to operate yet because of some bureaucratic quibbles about absorbing the employers referred to ATO PA 2, former public company owned by municipalities and responsible for waste collection, that has been failed in 2014. Therefore waste management will be outsourced in favor of a public consortium that will take care of workforce management, devices and facilities, maintenance investments and processing waste in landfill. In the meanwhile, from 2014 waste management has been assigned to a private company that made a tender for managing this service and it won.

Obviously, respect to the private one, Public sector performance has a broader impact on the quality of life of people and may constitute either an acceleration factor or a constraint for the growth of the local area[8].

In this sense, the decision about which waste management system municipality is prone to adopt is a clear demonstration of what it has been already said. Reducing the amount of waste in landfill by implementing a zero waste strategy, as well as an economic benefit (given the saving connected to the refusal of the incineration system as the main waste treatment system) also turns into a driver simultaneously of direct and indirect benefits for society: on the one hand it satisfies the collective interest in the environmental care; on the other hand it creates the conditions to generate redundant positive externalities that are bound to strengthen the image and attractiveness of the reference area.

2 AIM OF THE PAPER

Recycling has always been a challenge for municipalities. In particular, reviewing the results attained throughout Italy, it is known that the main causes of the failure of such initiatives are:

[7] Studio di progettazione e consulenza aziendale Dott. V. Marinello, "*Progetto Area di Raccolta Ottimale (ARO): comuni di Palazzo Adriano, Prizzi, Bisacquino, Giuliana, Chiusa Sclafani",* 2014, p.1.

[8] The greater impact is proven by the broader array of products Public Sector is used to provide to the citizenry respect to the private one. Specifically, Public Administration is conventionally used to provide these products: [F0B7?] Laws and administrative deeds;
•Collective goods;
•Individual goods of collective interest;
• contributions;
•Rules, programs, guidelines.
E. Borgonovi, "*Principi e sistemi aziendali per le amministrazioni pubbliche",* Egea, 2005, p.62.

- inability of municipality to raise awareness of the population, if the municipality either has not achieved satisfying fraction of waste recycled level or it has not been able to be accountable to the citizens for the results of such a policy[9];
- lack of sufficiently strong control mechanisms, ready to ensure the continuation of a similar program, which of course, at least in the short term, encounters strong resistance, given that it threats a more convenient but also more polluting model of life.

Obviously, little municipalities are more likely to get a target fraction of waste recycled sooner respect to the greater ones. In particular, according to Ispra[10] recycling rate of over 60% is more likely in municipalities including both between 2,501 and 5,000 inhabitants and between 5001 and the 15,000 inhabitants (respectively 52,6% and 55,9% of the corresponding class of municipalities have reached the target fraction). Over 60% of recycling has been achieved also in the 33,3% of municipalities comprising between 100,001 and 200,000 inhabitants. Instead, there's no municipality with a population of more than 200,000 inhabitants that has been reaching such a threshold[11].

However, it is useful to take into account the troubling datum about level of fraction of waste recycled in Sicily: in 2014[12] one third of all the municipalities did not go beyond the 5% and in 2015 the regional recycling rate has not recorded so many improvements, given a passage from 12,5% of 2014 to 12,8%[13].

Recalling the success of the San Francisco waste system[14] as successful pioneer of this new treatment method of waste, the aim of this research is to explain how the benefits for the

[9] *«How will who hold whom accountable for producing whose results? The question of democratic accountability for performance is really four distinct but interrelated questions: who will decide what results are to be produced? Who is accounting for producing these results? Who is responsible for implementing the accountability process? How will that accountability process work? [...] Thus accountability for performance requires some explicit expectations about what results will be produced by when».*R. D Behn, "*Rethinking Democratic Accountability*", 2004, Brookings Institution Press, Washington DC, pp. 62-63.

[10] Public Agency supervised by Ministry of the Environment, responsible for technical scientific activities connected to environmental care, protection of water and soil conservation.

[11] Ispra, "*Rapporto rifiuti urbani",* Ed. 2016, pp. 64-65. Available on: http://www.isprambiente.gov.it[2017].

[12] For more information: http://meridionews.it/articolo/44305/differenziata-i-risultati-dei-390-comuni-siciliani-unterzo-sotto-il-5-regione-impone-prescrizioni[2017].

[13] Surely there is a remarkable delay about aggregation of data referred to the separated waste collection level achieved in each municipality. Governor of Sicily, has recently declared that in 2016 Sicily has reached 21 % of fraction of waste recycled. In particular, he stated it has been observed locally an increase of 1% on average respect to the year before. Actually, to increase this datum massively, it is necessary to focus on the big cities, where fraction recycled is still low (around 10% in 2015, according to Ispra). To read more, Ispra, *op.cit.,* p. 63 and following pages and http://palermo.repubblica.it/politica/2017/01/27/news/rifiuti_crocetta_annuncia_differenziata_al_21_per_cento_ottimrisultato_-156996772 [2017].

[14] Currently, San Francisco diverts 80% of its waste away from landfills. According toNew York *Times*reporter Matt Richtel, *«San Francisco also has a world-class reputation for its composting processes, which turns food waste into fine, coffee-like grounds that is sent to farms as fertilizer».*And he observed that San Francisco has been becoming the "Silicon Valley of recycling".Reference: http://www.nytimes.com/2016/03/29/science/san-

citizens have sprung up from the start of this program (with its ten steps) since 2012 in Palazzo Adriano, a small village just 80 km far from Palermo. To do that, it has been crucial to integrate the 3-R scheme with the S of saving money, just to highlight what has been strengthening this strategy over time, bringing out a virtuous waste management system, embedded into a tough regional situation.

Time horizon coincides with the lag of time whose data have been provided by municipality, namely a period of five years, including the year before the implementation of this strategy. This choice is justified by desire of understanding the strength of that change implementation of the policy has been unleashing over time.

To summarize, this paper is intended to answer the following research questions:

a) How does the ensuing waste management system work?

b) Why saving process might be seen as a significant push towards zero waste goal, influencing Behn's accountability paradigm?

c) Which are the most conducive causal loops to explain behavior of waste cumulated in landfill?

3 METHODOLOGY

Assuming that unpredictability and dynamic complexity are the main enemies for our better understanding of a system – in the classical meaning developed by Von Bertalanffy as a complex of tightly intertwined elements – and given that all is change, policy-maker ought to oversee changes occurring within a certain system at many time scales, and trying to keep track of interactions among these different scales[15]. A system without a time-oriented perspective is not a system. Therefore, if a municipality wants to succeed in managing a system, it should cater for performance, defined both as outputs (and mainly outcomes, as projections of the outputs in the outer local system in the long run) it is supposed to get at the end of its policy, and the process that is in charge for those end-results[16].

Combining SD models with "information feedback support" models based on financial perspective and static performance management, should be a profitable trick to foster mental models' elicitation and improve organizational capabilities in assessing performance through a sustainable development perspective (dynamic performance management)[17].

francisco-the-siliconvalley-of-recycling.html?_r=0[2017]. To still lift this threshold up to 100% in line with the ambitious Zero Waste Strategy goal within 2020, herewith listed a practical guideline for citizenry on: http://sfenvironment.org/zero-waste[2017].

[15] J.D. Sterman, "*System thinking and Modeling for a complex World",* McGraw-Hill, 2000*, p.22.*

[16] Nowadays citizens are demanding better results from government at a time when resource constraints are increasing, and level of trust in government at all levels is at an historic low. So to get more accountability rejecting black box concept and making all the process transparent as much as possible has become a priority. M.B.Sanger, "*Does measuring performance lead to better performance?",*Journal of Policy Analysis and Management, 1-18, 2012, p. 1.

[17] C. Bianchi, "*Dynamic Perfomance Management",* Springer International Publishing, 2016*,* p. 37.

Precisely, Dynamic Performance Management systems do not only intend to carry out a quantification of effects arising from the implementation of specific public policies, but also have the benefit of providing appropriate information to decision makers who can use them to influence in a targeted manner the reference environments and to evaluate the results achieved[18]. In fact, SD modeling techniques, embedded in a DPM perspective, allow to implement a double loop learning process since on the one hand the modeling process is rooted on the elicitation of decision makers' perceptions of the real world; on the other hand, decision makers' mental models are challenged through model validation (namely, the search for a consistency between the model hypothesis on the system structure and the simulated behavior).Validation of a SD model might be seen as an input that enables policy maker to deploy an action that can rely on a basement with an acceptable level of scientific rigor[19]. Furthermore, SD modeling challenges mental models through simulation and it should also be considered as a practical way to test in a "protected" environment the consistency – in terms of robust trade-offs perception – of their own decisions [20].

In order to provide decision-makers with proper lenses to interpret such phenomena like unpredictability and dynamic complexity, to understand feedback structure underlying performance, and to identify alternative strategies to change of the structure for performance improvement, SD modeling has been used to support an understanding of:

•how end-results can be affected by performance drivers;

• how performance drivers can, in turn, be affected by the use of policy levers aimed to influence strategic resource accumulation and depletion processes;

•how the flows of strategic resources are affected by end-results[21].

After all, Zero Waste strategy is a shift of paradigm that has required time to be understood and accepted by citizens, and initiatives which take up time and can thrive over time in a certain way, depending on how the action started out from the beginning (path dependence[22] as a peculiar feature of SD) typically represent the natural area of SD approach[23] applicability.

[18] C. Bianchi, W.C. Rivenbark, "*Alla ricerca dei fattori rilevanti nell'adozione dei sistemi di gestione della performance nelle amministrazioni pubbliche territoriali. L'analisi di due casi di studio*", Azienda Pubblica, n. 1, 2013, p. 36.

[19] System dynamics might push away the performance paradox risk. This phenomenon is caused by the tendency of performance indicators (especially when they are conceived as static) to run down over time. They lose their value as measurements of performance and can no longer discriminate between good and bad performers. A typical process that can present this risk is positive learning, according to which, as performance improves, indicators lose their sensitivity in detecting bad performance. M. W. Meyer & V. Gupta, "*The performance paradox. Research in Organizational Behavior*",

[20] *«Microworlds (otherwise stated as: "interactive learning environments", or "management flight simulators") are SDbased simulation models aiming to foster policy debate. The use of such simulators, supported by a learning facilitator, can help policy makers understand the dynamic relationships between strategic resources and performance variables».* To read more, C. Bianchi, "*Dynamic Performance Management",op. cit.,*p. 199.

[21] C. Bianchi, *op.cit.,* p. 72.

[22] *«Taking one road often precludes taking others and determines where you end up».* J.D. Sterman, *op.cit., p.22*

[23] This methodology has been conceived by J. Forrester since 1950s at MIT.

System Dynamics gives the possibility to build up an exploratory model looking into the dormant dynamics that would justify the adoption of such a policy and reasoning in terms of stocks (states) and flows (changes) in accordance with the principle of accumulation[24].

Assuming that stocks have four important characteristics (they have memory; they change the time shape of flows; they decouple flows; they create delays[25]), stock and flows structure might be seen as a completion of another pillar of SD approach, the causal loops diagramming, namely the possibility of mapping the system in terms of cause and effect relationships between individual system variables; the latter ones[26], when linked, form closed loops that feed back to the structure altering the relative importance of each of the variables listed in the system[27].

By taking resort of system dynamics it becomes possible to explicit the patterns of casual loops hidden in a traditional statistical tool[28], trying to distinguish, in the present case, the commitment both at source and at the end of the life cycle time of each product to reduce the stock of waste in landfill.

Therefore, a Dynamic Performance Management Chart based on SD modelling techniques is a useful dashboard to keep track of performance, both as process and end-results.

DPM chart, as structured below, acknowledging first the strategic role of population as the main factor that affects waste collection, emphasizes the importance of saving process, shown as a ratio between the actual waste management cost level and a target. This ratio might be seen as a performance driver that leads operationally, as end-results, to a reduction of waste not recycled and finally to a reduction of waste destroyed; this happens thanks to the concurrent

[24] L. Booth Sweeney and J.D. Sterman, "*Bathtub dynamics: initial results of a systems thinking inventory*", System Dynamics Review, volume 16(4), 2000, pp. 252-253.

[25] D.M. Buede & W.D. Miller, "*The engineering design of systems: models and methods",* Wiley, 3rd Edition, 2016, p. 464.

[26] There is a need of a framework to frame any attempt of modeling management control system. To cope with this necessity, a good starting point is surely sorting out a useful conceptual framework:

- to define an appropriate set of independent variables related to the firm and, variables that explains environment and its actual influence over the control process;
- to select features of managing control system;
- to find out the links between any facets of the process listed above.

F. Amigoni, "*Planning management control systems",* 1978, Journal of Business Finance and Accounting, 5(3), 279291, p, 1.

[27] *«Learning from performance measures, however, is tricky. It isn't obvious what lessons public managers should draw about which factors are contributing to the good or poor performance, let alone how they might modify such factors to foster improvements. Improvement requires attention to the feedback»* R.D. Behn, *"Why Measure Performance? Different Purposes Require Different Measures",* Public Administration Review, 63(5), 2003, p.8.

[28] For example a multiple regression analysis whose aim would have been trying to explain how the criterion variable (in this case, the tons of waste permanently conferred to the landfill) is affected by the predictor variables, and to what extent its change over time is due to the influence of each predictor variable, individually considered. Therefore, recalling the 3R scheme integrated with C (composting) the author would have built up a multiple regression model drawing on some proxy variables representing the pillars of a common zero waste strategy.

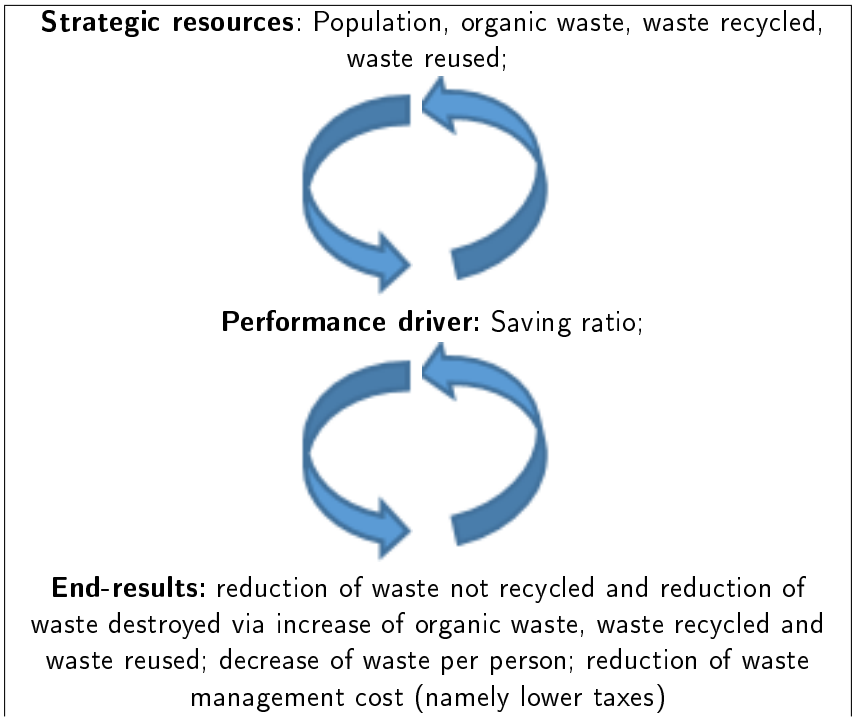


Table 1 Dynamic Perfomance Management Chart applied to a zero waste strategy

increase of organic waste, waste recycled and waste reused that flow into the corresponding stocks, feeding strategic resources ready to be exploited or marketed. But saving process is also intended to decrease the amount of waste made by each person, diminishing as a consequence the total amount of waste produced each year. From an economic point of view, these end-results determine lower taxes motivating citizens to fuel this pattern persistently.

4 Results of the empirical survey

Reason that pushed the municipality of Palazzo Adriano to increase the fraction of recycled waste, with all kinds of benefits (economic, environmental and externalities) identified before, lies in the alarming datum of the amount of waste piled up in the landfill for the year 2011, namely 693,84 tons.

To reverse this dangerous trend, the mayor of Palazzo Adriano, once elected, decided to adopt in 2012 a zero waste strategy; specifically, he pledged to reach at the end of the mandate

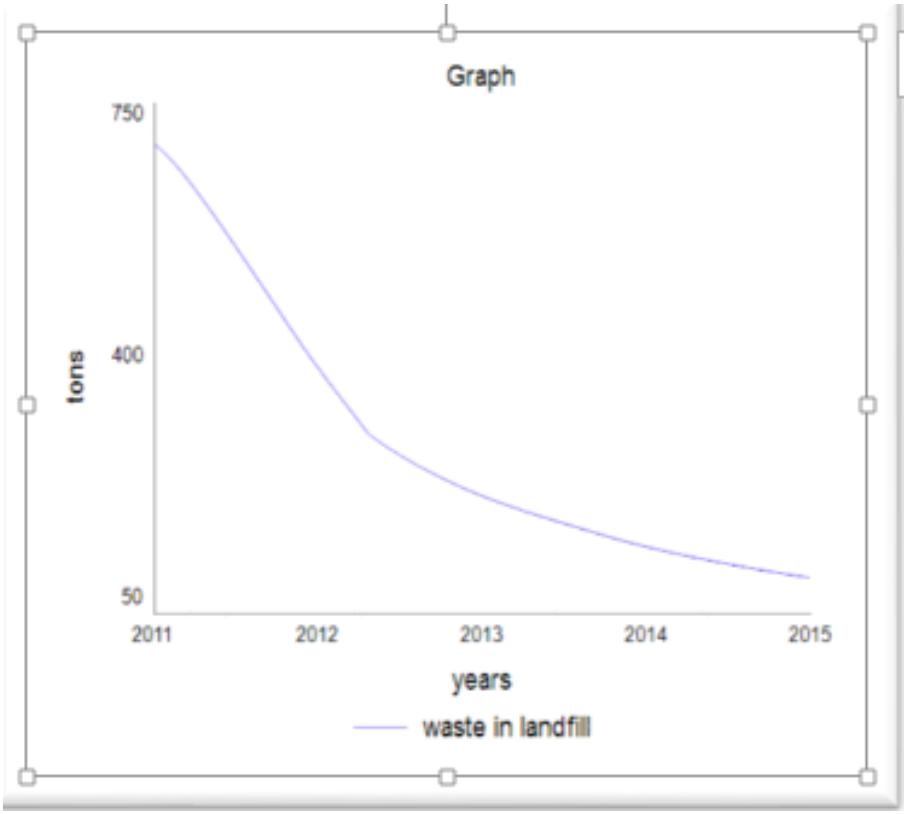


Fig. 1 Tons of waste piled up in landfill during the period 2011-2015

the ambitious threshold of 65% of recycled waste[29], which represents tons of waste diverted away from landfill and incineration system as last treatment stop.

As a consequence, municipality has achieved the following decreasing pattern of behavior related to the amount of waste irreversibly conferred to landfill, getting to the level of 99 tons in 2015:

At the same time it has been seen during the period 2011-2015 an increasing pattern of behavior of the fraction of waste recycled, up to the threshold of 65% that municipality set as target at the beginning.

Even though depletion of towns due to the juvenile migrations abroad or in the north looking for a job a common issue spreading in the South of Italy and particularly in Sicily[30] it would be

[29] Directive n. 98/2008 has established that by 2020, preparing for re-use and recycling of waste such as at least paper, metal, plastic and glass from households and possibly from other origins as far as these waste streams are similar to household ones, shall be increased to at least 50% in terms of weight. Truly, threshold of 65 % had been already listed in the Italian judicial system in art. 205 Legislative Decree n. 152/06, as an objective to be pursued by 2012. To read more, go to AA VV, "*Libro verde per la sostenibilità ambientale delle infrastrutture nodali di trasporto*", Franco Angeli, 2016.

[30] *«The highest long-term unemployment ratios were principally concentrated in southern and peripheral regions of the EU. There were 11 Greek regions, seven Italian regions (including the island of Sicilia), four French*

Fraction of waste recycled
2011 4%
2012 24%
2013 62,77%
2014 63%
2015 65%

Table 2 source: Municipality of Palazzo Adriano

year	population
2011	2227
2012	2200
2013	2178
2014	2155
2015	2135

Table 3 Source: Istat

biased to ascribe to the demographical issue such a progression of waste piled up in the landfill, since the demographical situation in Palazzo Adriano has been shaped in the last five years in the following way:

By comparing the significance of data related to the decrease of waste in landfill with the decrease of population over time, it is easy to realize that surely there have been other aspects that might be considered as more significant to clarify the pattern of behavior observed during the period of simulation.

To realize which causal loops are the most conducive to explain the dynamics underlying the implementation of such a policy, it is useful to take resort of the whole waste chain built up in the SD model[31]. Its development has been grounded in the idea of splitting up this chain in different steps just to figure out the destination of the waste produced each year.

There are two key conveyor stocks in this chain: waste produced that represents the amount of waste produced that is going to be subdivided, in accordance with its destination in three different outflows (organic waste, waste recycled, waste not recycled); the other one, waste in landfill expands upon one of the three possible destination, and in turn is going to be depleted by two outflows, namely waste destroyed and

waste reused which represent respectively the ultimate passage or a soft brake respect to the default scheme.

départements et territoires d'outre-mer (no data available for Mayotte), three regions from each of Bulgaria, Portugal (including the islands of Madeira and the Açores) and Slovakia, the two autonomous Spanish cities, both of the Croatian regions, and the Belgian capital city Région de BruxellesCapitale/Brussels Hoofdstedelijk Gewest». Eurostat, *"Eurostat Regional Yearbook"*, 2016, p.112.

[31] The variable "Equilibrium Switch" warns that the model has been initialized in equilibrium.

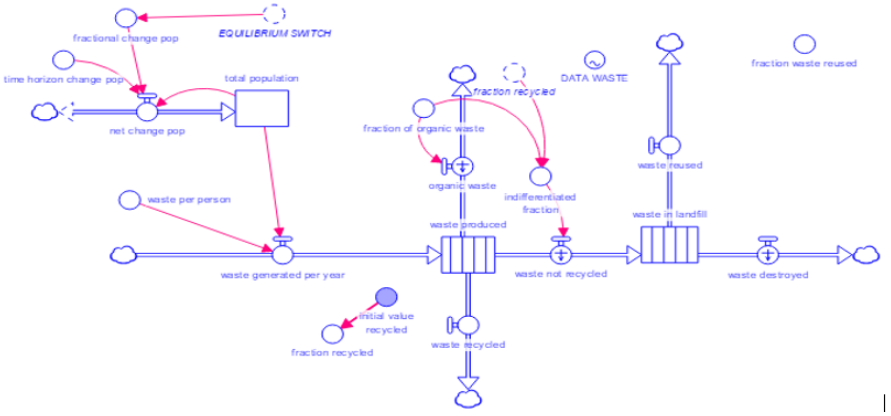


Fig. 2 The waste management chain

The conveyor stock waste in landfill accumulates the inflow waste not recycled, which comprises the amount of waste collected that has not been undergone by zero waste strategy. For this reason the leakage fraction referred to waste not recycled, that flows out of the previous conveyor stock (waste produced) has been determined as completion of the sum of fraction recycled and fraction of organic waste, namely the leakage fractions respectively referred to the outflow waste recycled and organic waste (compost).

Waste not recycled passed through – in accordance with a First in First out scheme as a typical feature of the conveyor scheme[32] – the stock to fuel the outflow waste destroyed, namely the amount of waste bound to be sent to the incinerator. Leakage fraction referred to waste destroyed has been set as exponential just to sketch a nonlinear discharge. Transit time related to the conveyor stock waste in landfill embodies somehow the constrain of capacity, since it takes time to treat waste not recycled and to destroy definitely what it has been not possible to recycle before.

Waste in landfill has been depleted also by another outflow that takes into account one of the three pillars of the zero waste strategy, namely reuse. Reused waste comprises all that stuff like aluminium or steel that was not possible to recycle in the steps before. In this case the inflow waste reused has been defined by selecting as leakage fraction of waste destroyed the completion of fraction waste reused. Therefore, waste reused has been represented by everything else is not going to flow out of the stock as waste destroyed. Positive correlation between the amount of waste in landfill and the amount of waste reused might be seen as a proof of the schedule referred to the treatment needed to make those products ready to be used again. In

[32] *«The conveyor allows for a first in, first out (FIFO) behavior, with allowance for leakage and for probabilistic delay times»,* T.K. BenDor & S.S. Metcalf, *"Conceptual Modeling and Dynamic Simulation of Brownfield Redevelopment",* 2005, p.11. Available on: http://www.systemdynamics.org/conferences/2005/proceed/papers/BENDO191.pdf [2017].

fact, what it has been impossible to recover in the steps before, it is going to be transferred to a provisional landfill just to undergo some interventions in order to refresh its usefulness.

On the other side of the waste chain, waste produced is boosted by an inflow that witnesses the role of population as the main responsible for waste production. Broadening the scope, waste produced is a conveyor stock that has been depleted by three outflow. One is referred to the compost production loop; another one is referred to the recycling loop, while the last one is related to the transportation to the landfill.

About compost, renowned as common natural fertilizer used locally for farms, it is useful to know that the corresponding outflow organic waste turns out to be related to the leakage fraction "fraction of organic waste" that has been kept constant (2,3% of total waste produced) during the period as well as the fraction of waste reused (3,6% of the amount of waste in landfill), because by looking at the data gleaned from municipality it has been seen a quasi-constant pattern of behavior. About reuse and recycled waste (which in turn comprises fractions of glass, paper and cartoon diverted from the transportation to the landfill) joining CONAI system[33][34], producers and users are obliged to pay a contribution depending on the type of packaging waste brought out in the market (paper, plastic, cartoon, aluminium, steel or glass). CONAI withdraws a minimum fraction of this contribution for the fulfillment of its bureaucratic tasks, while the greater part of it is bound to be sent to the Consortia responsible for processing each type of packaging waste in order to refresh their usefulness[35].

These consortia, in turn, according to the ANCI[36]-CONAI framework agreement (Accordo quadro ANCI-CONAI), assign to municipalities a compensation to cover higher costs connected to the differentiated collection of waste.

Currently, municipality are used to outsource waste management to private companies, but in the future, Palazzo Adriano together with the other municipalities, as said before, will manage this service by empowering a consortium of public services.

For this reason, model has been completed by internalizing the ARO system that will start out soon, replacing the entrustment to the private company.

In the past, as a guarantee for the success of recycling programs developed by each municipality, it has been decided to set up a recovery center (Centro Comunale di Recupero) for recycled waste in Bisacquino, chosen because it is equidistant for all those municipalities joining "Valle del Sosio" area. This center dealt with temporary storage of recycled waste, awaiting to be sent to each consortium designated for processing each type of packaging (COREPLA, COREVE).

[33] No profit management system for packaging waste in Italy introduced by Law Ronchi (Legislative Decree n. 22/97).

[34] For more information, go to http://www.conai.org/chi-siamo/cose-conai. CONAI is a member of EXPRA, the alliance for the extended producer responsibility, the European organization that represents the no profit management system for packaging waste [2017].

[35] For example COREPLA is the consortium delegated for processing plastic, while COREVE is that one responsible for processing glass.

[36] The category association including all the municipalities in Italy.

As said before, from 2014, given the failure of the ATO PA2, waste disposal service has been outsourced in favor of a private company, Traina srl. Assuming that this private company is equipped with a transference station in Cammarata (a municipality around 40 km far from Palazzo Adriano), currently recycled waste is stored there, until quantities piled up in the ecological island – which is located in each of the municipalities of the reference area – become commercially interesting; whereas undifferentiated waste is ordinarily sent to a landfill specifically identified by the Regional Government (currently Bellolampo, the main landfill of Palermo, whereas in the past it was Siculiana). Obviously, with the emersion of ARO system, as shown in the above mentioned "Progetto ARO", it is expected to reactivate the center of Bisacquino.

Therefore, to sum up, recycled waste is first:

- conferred to the ecological island, located in each municipality joining ARO-Valle del Sosio;
- then transferred to the transference station of the Traina company, when ecological island has accumulated such quantities to motivate transport (obviously, transporting small quantities would be uneconomic);however, whenever load exceeds 20 cubic meters, it is required a permission from the Province Government[37].

Getting back to the end of the waste chain, amount of waste destroyed is bound to determine a total cost due to the treatment of waste in landfill required. This cost is equal to the tons of waste destroyed per year multiplied by the cost per ton treated. Precisely, it has been set equal to 102 euro per ton conferred, using an average value since cost per ton treated has been kept quasi stable over time. Consequently, the more waste municipality has carried to the landfill awaiting to be destroyed, the more cost it needs to bear. This cost determines participation cost ARO, together with devices and maintenance investments and workforce cost, given that there are some employers appointed for the waste collection referred to Consortium "ARO -Valle del Sosio"[38]. Also, zero waste strategy starting investment represents a sort of sunk cost that Municipality bore at the beginning just to let zero waste strategy program start out. Therefore, it is a sort of umbrella term that comprises all the investments referred to disclosure in favor of the citizenry and building or purchase of the facilities connected to the treatment of waste in accordance with this new regime or devices referred to this program (for example the baskets related to each type of waste just to allow people to differentiate the waste produced at the beginning of our whole waste chain). As a consequence, the whole waste management cost for municipality is equal to the cost of investments bore at the beginning just to make the program start out plus the partecipation cost ARO.

Every year municipality has gathered a saving due to the lower level of waste piled up in the landfill thanks to the implementation of this program, which is intended to hold the waste

[37] Art. 183, letter m, Legislative Decree n. 152/2006. Available on: http://www.camera.it/parlam/leggi/deleghe[2017].

[38] As said before, hypothetically, they will be the same ones referred to the ATO PA2, the former public company responsible for the waste management system until 2014

management cost down more and more. This value has been compared with the initial waste in landfill cost referred to 2011, namely the year before the start of the program. The underlying idea is that the less waste municipality is used to confer to the landfill the more saving respect to the year 2011 it is bound to get year by year and at the end the lower taxes citizenry is supposed to pay as a reflection of expenditures downsizing.

By comparing the saving achieved thanks to the implementation of a zero waste strategy with a target waste management system cost to be pursued during the mandate and defined as 25% less than the waste management cost referred to 2011 (almost 400,000 euro), it is possible to build up a ratio that symbolizes the attempt to get to the target as much as possible. *Sic stantibus rebus*, according to financial statements referred to 2015, waste management cost has diminished to 322,100 euro[39]. Obviously once this virtuous cycle has started out, the more it lasts, the more difficult it will be to get a further improvement, both in terms of saving and increase of the fraction of waste recycled[40].

Saving ratio, in terms of policy, internalizes efforts of municipality to make citizens more and more aware of benefits emerging from such ongoing virtuous cycle, by periodically exposing – personally encountering citizens or resorting to social network – an in-depth analysis of the state of art. By way of example, two meetings have been mentioned: the meeting "zero"[41], where municipality for the first time explained the program, with its ten steps, to citizens on 10 July 2012; the other one relating to the visit in Palazzo Adriano on September 26, 2012 of Paul Connett, the guru of zero waste strategy[42].

The above-mentioned ratio becomes the input of a graph function that embodies the idea that the more saving municipality gets the more citizens are inclined to recycle, looking at the benefits they might get thanks to recycling, This graph function, that is bound to spring up from 2012 in accordance with the beginning of the program, is going to influence the fraction recycled, since it is bound to be added to the initial value of fraction of waste recycled (referred to 2011), enhancing or diminishing it.

Herewith represented the pattern of behavior assumed by the graph function during the simulation period:

Introducing the above mentioned graph function gives the possibility to show the first major causal loop (recycling loop) that may explain why there has been a decreasing pattern of behavior for the amount of waste discharged in the landfill.

Recycling loop is a reinforcing loop which refers to the recovery of usefulness of the amount of waste that it has been already circulated.

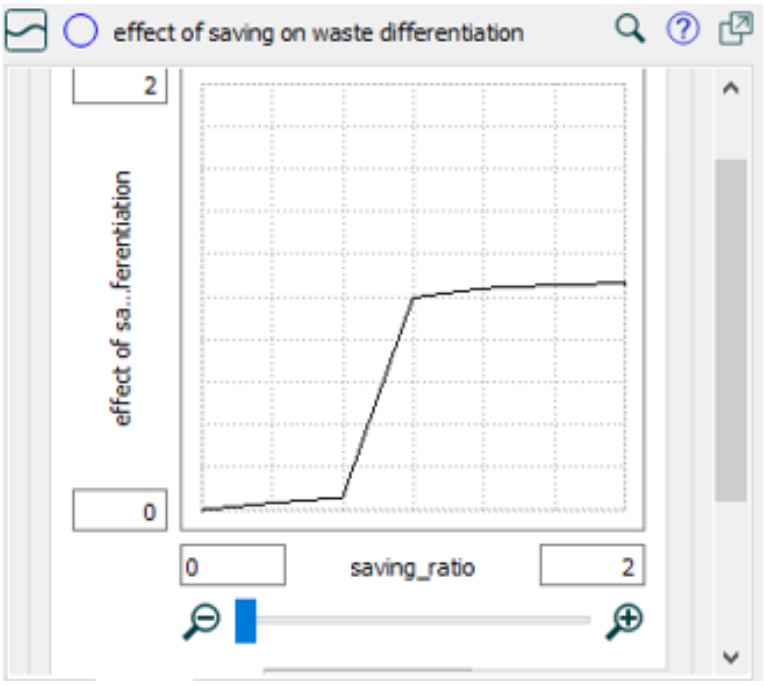
Implementing such a program activates a key factor of a typical zero waste strategy, namely recycling, causing waste in landfill and waste destroyed to decrease; this means lower cost for

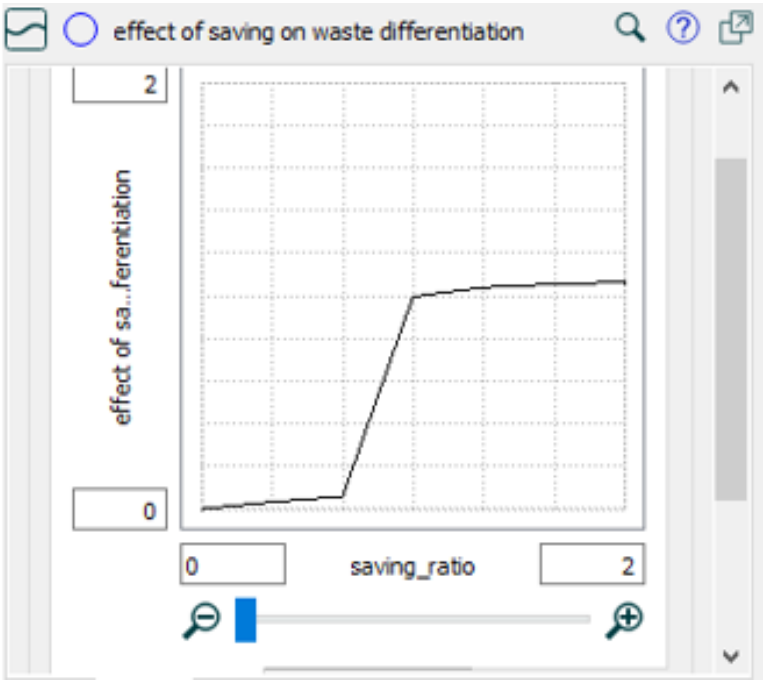
[39] http://www.comune.palazzoadriano.pa.it/it/DelibereConsiglio/2016/Luglio/CONTO%20DEL%20BILANCIO%202015%20-%20PARTE%20II%20SPESA.pdf [2017].

[40] This reasoning has been embodied in the decreasing decreasingly pattern of behavior referred to the amount of waste in landfill.

[41] http://www.magaze.it/wps/2012/07/04/parte-il-16-luglio-la-campagna-per-la-raccolta-differenziata-a-

[42] http://www.magaze.it/wps/2012/09/26/palazzo-adriano-con-paul-connett-verso-rifiuti-zero-una-rivoluz


effect of saving on waste differentiation
2
effect of sa...ferentiation
0
0
saving_ratio
2


effect of saving on waste differentiation
2
effect of sa...ferentiation
0
0
saving_ratio
2

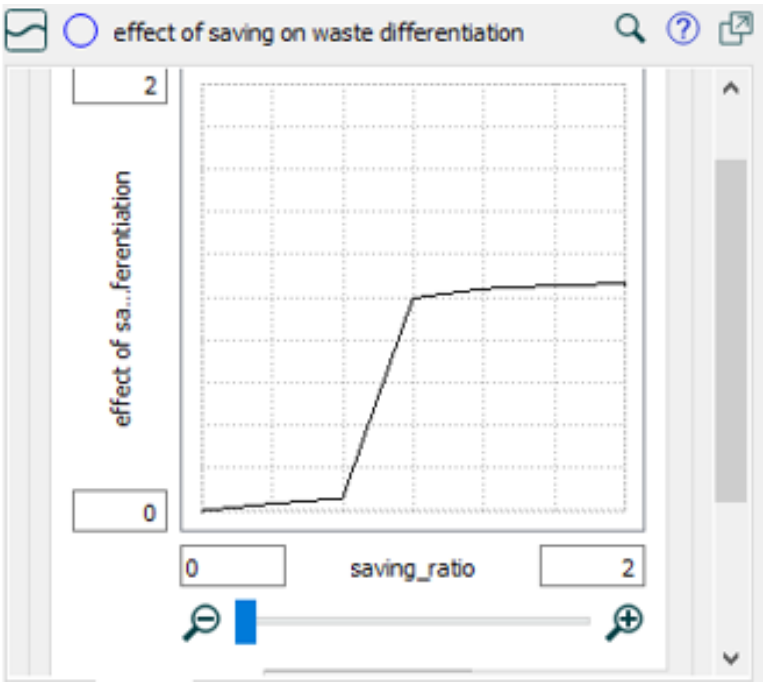


Fig. 3 Graph function "Effect of saving ratio on waste differentiation"

municipality and conversely an higher level of saving respect to 2011, the year before the start of this program. The more municipality is close to its target (25% less than the waste management cost referred to 2011), the more people feel themselves as encouraged to differentiate waste already brought out in the system as much as possible, causing waste piled up in the landfill to decrease further. Influence of saving process about boosting the differentiation of waste already circulated, might be seen as a remarkable impulse to make the three minor balancing loops (composting, recycling, reuse[43]) persist over time.

Finally, attitude of saving thanks to the development of such a policy over time is going to address also the habit of people to make waste at source. So saving ratio acts also at the start of the waste management chain encouraging people to make waste as little as possible, whenever they see that their efforts are rewarded by a lower level of actual cost, namely taxes. Otherwise, if the level of saving is lower and lower, they are less encouraged to produce less waste. This explanation constitutes the basis of legitimacy for the following graph function:

In particular, waste per person is modelled as a default value equal to 0,22 tons per person –a sort of estimation of how much waste a person is used to make on average by looking at the

[43] As seen before, the more compost and the more recycling people are used to make, the less waste municipality is going to carry in the landfill because somehow citizens succeed in retrieving the intrinsic value those products still embodied. The same is for waste reused, since the more waste they succeed in reusing the less waste piled up in the landfill is going to be pass through incineration system.

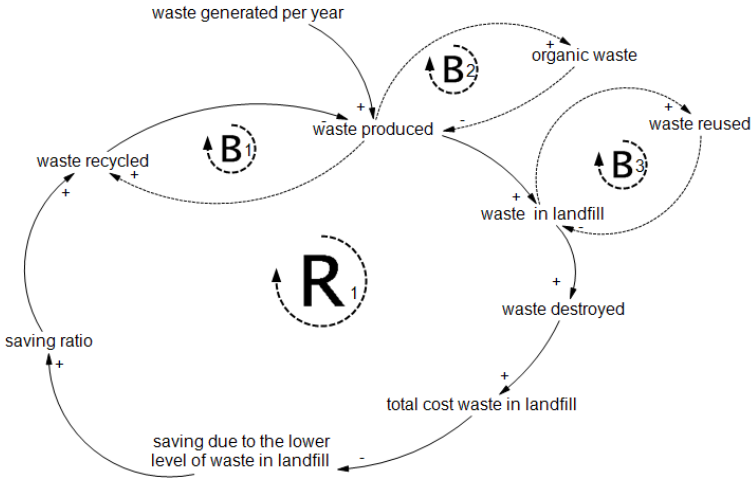


Fig. 4 Recycling loop (R1)

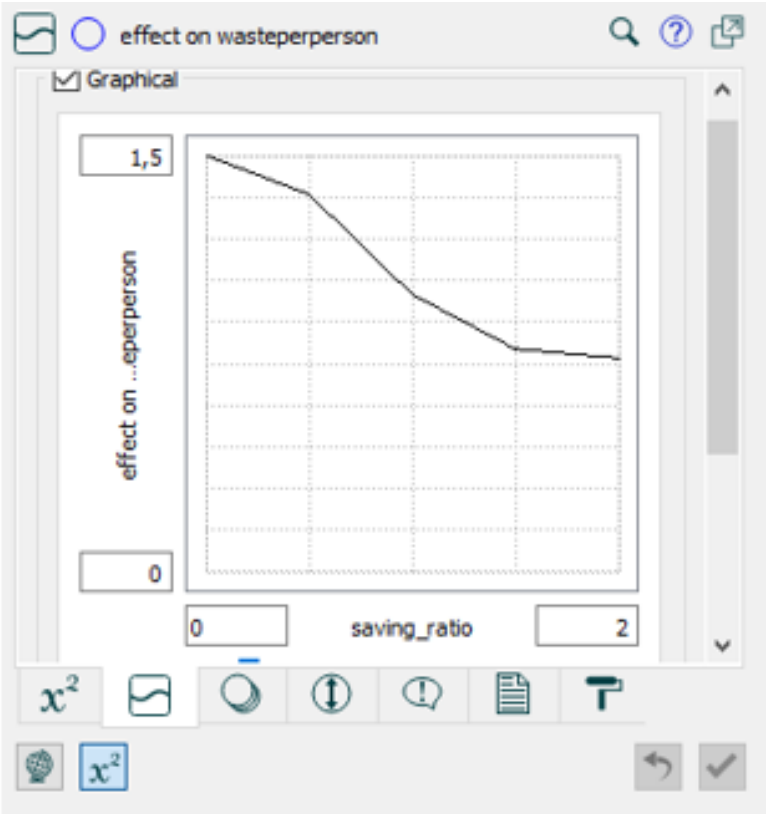


Fig. 5 Graph function "Effect on waste per person"

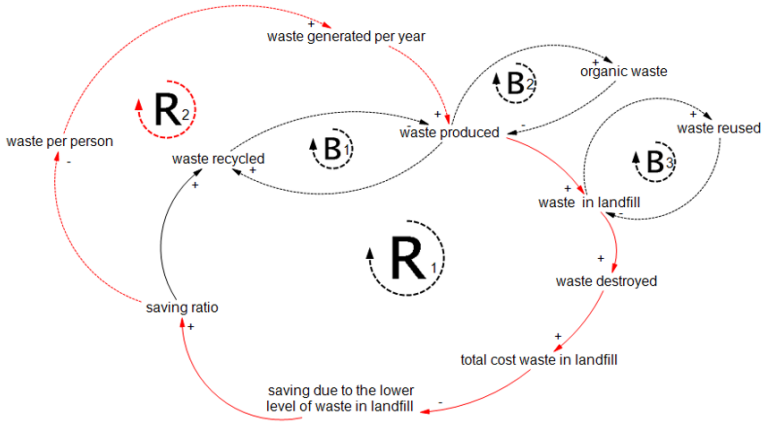


Fig. 6 Waste production loop (R2)

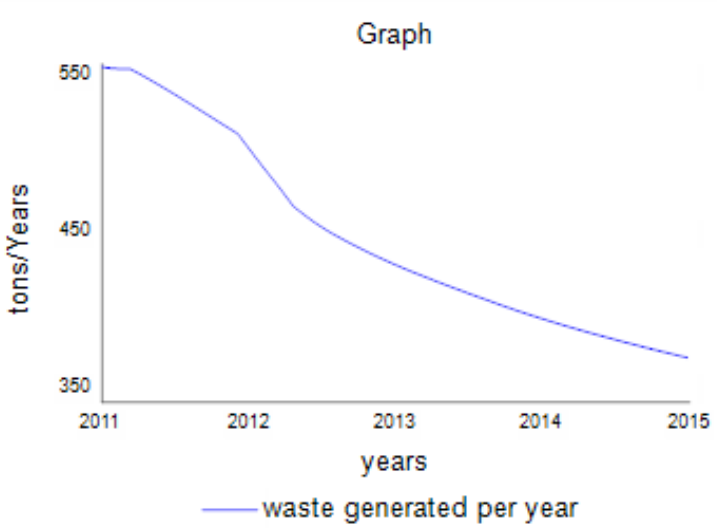


Fig. 7 Waste generated per year during the period 2011-2015

official data gleaned thanks to the municipality –multiplied by the effect that saving ratio can exert on the habit of recycling of each person.

This explanation constitutes the backdrop for introducing the second major causal loop (waste production loop) that, as well as the recycling loop, make the three balancing loops – designated to squeeze the amount of waste irreversibly conferred to landfill – strengthen over time.

Therefore, by diminishing the amount of waste per person, saving process is intended to decrease consequently the whole amount of waste produced each year.

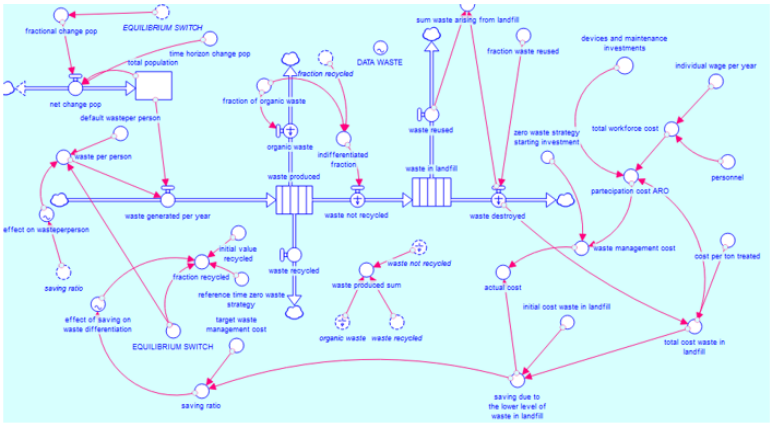


Fig. 8 The Whole SD model

5 CONCLUSIONS

Zero Waste Strategy has been a shift of paradigm that has required time to be understood and accepted by citizens. It might be seen as a double loop learning (Argyris and Schon) answer to sustainability challenge, given that by implementing this strategy, municipality seems to be more willing to look forward, to reveal an explicit mental model, to assess its consistency over time and to improve it[44] by collecting all the weak signals of change arising from the outer systems. Recalling the democratic accountability paradigm conceived by Behn (2004), accountability process requires people to form expectations about performance; specifically, accountability process works letting citizens approach directly the end-results of their ongoing good practices. Actually, lower tax burden level has been crystalized as the reward of their effort to cause waste to decrease both at source and at the end of the waste chain depicted in the SD model.

That being said, given that the threshold of 65% of fraction of waste recycled might be considered as a national diktat (art. 205 Legislative Decree n. 152/06) framed within the boundaries of a preeminent European objective (Directive n. 98/2008), saving process might be seen as a facilitator, since its advertising allows to achieve this goal as soon as possible, encouraging people to still follow zero waste program.

The latter one has been becoming the input of a virtuous cycle – fostered by saving process – which has jeopardized the previous regime of living of citizens, since for example the usage of plastic dishes or disposable paper cups, traditionally reckoned as comfortable, has been discouraged. Anyway, there should be still room for fine tuning, starting for example reverse logistics initiatives as well as in Scandinavia[45]. These initiatives might engage community more

[44] J. W. Forrester, " *Some Basic Concepts in System Dynamics"*,Sloan School of Management Massachusetts Institute of Technology, 29thJanuary 2009, p. 10.

[45] ICA (but also other supermarkets like Meny or Rema 1000), one of the biggest supermarkets in the Northern Europe and the biggest one in Sweden is equipped with a compelling reverse logistics grounded on an equipment

and also they might fuel industrial responsibility, since on the one hand firms would be forced to think further about a sustainable production; on the other hand consumers will be more and more sensitive to the environmental care. In this perspective, consortium like ARO-Valle del Sosio might be seen as a crucial bridge between community and industrial responsibility thanks to a door to door collection inspired by recycling criteria. It is important not to frustrate citizens commitment at source (by making waste as little as possible) and at the end (by differentiating waste as much as possible) avoiding to compact again waste and to send it indistinctly to landfill. Therefore, ARO together with consortia delegated for processing each type of packaging waste, will constitute the idealistic bond to let the two zero waste strategy levers (namely engaging community and industrial responsibility) match. In fact, compensations given to ARO are bound to be shifted downstream to the municipalities determining lower taxes for the citizens; this aspect will foster people more and more to make the above mentioned reinforcing loops persist over time. In addition, treatments carried out by consortia are the unavoidable starting point to enables industries to exploit these materials again.

REFERENCES

AA.VV., 2016. Libro verde per la sostenibilità ambientale delle infrastrutture nodali di trasporto. Franco Angeli, Milano.

Amigoni F., 1978. Planning management control systems. Journal of Business Finance and Accounting, 5(3), 279-291.

Argyris C. & Schon D.A., 1978. Organizational learning: a theory of action perspective. Addison-Wesley Pub.Co.

Behn R.D., 2004. Rethinking Democratic Accountability. Brookings Institution Press, Washington DC.

Behn R.D., 2003, Why measure performance? Different purposes require different measures. Public Administration Review.

BenDor T.K. & Metcalf S.S., 2005. Conceptual Modeling and Dynamic Simulation of Brownfield Redevelopment.

Bianchi C., 2016. Dynamic Perfomance Management. Springer International Publishing.

Bianchi C. & Rivenbark W.C., 2013. Alla ricerca dei fattori rilevanti nell'adozione dei sistemi di gestione della performance nelle amministrazioni pubbliche territoriali. L'analisi di due casi di studio. Azienda Pubblica, n. 1, 35-59.

system in front of the gate, which could recover second-hand bottles and cans. To recover the used bottles, people can get 2 kr for each big plastic bottle, 1 kr for each small plastic bottle and beverage can. Finally, customers will get a ticket with the return money from the equipment and be available to buy things with this ticket in the supermarket. People can also choose to donate the money to charity if they want. M. GONG, Y. KONG, "*The implementation of green logistics in supermarkets in Sweden and China — A case study for ICA MAXI and JIA JIAYUE*", 2013, p.29. Available on: http://www.diva-portal.se/smash/get/diva2:722929/FULLTEXT01.pdf [2017].

Borgonovi E., 2005. Principi e sistemi aziendali per le amministrazioni pubbliche. Egea, Milano.

Booth Sweeney L. & Sterman J.D., 2000. *Bathtub dynamics: initial results of a systems thinking inventory*. System Dynamics Review, volume 16, n.4, 249-286.

Buede D.M. & Miller W.D., 2016. The engineering design of systems: models and methods. Wiley, 3rd Edition.

Conai, Accordo Quadro Anci-Conai 2014-2019. Available on: http://www.conai.org/download-documenti.

Connett P., 2012. Rifiuti zero, una rivoluzione in corso. Dissensi Editor, Milano.

Connett P., 2013. The Zero Waste Solution: untrashing the Planet One Community at a Time. Joni Praded Editor.

European Union, 2th March 2012. Treaty on Stability, Coordination and Governance in the Economic and Monetary Union.

Eurostat, 2016. Eurostat Regional Yearbook.

Forrester J.W., 29th January 2009. Some Basic Concepts in System Dynamics. Sloan School of Management Massachusetts Institute of Technology.

Gong M. & Kong Y., 2013. The implementation of green logistics in supermarkets in Sweden and China — a case study for ICA MAXI and JIA JIAYUE. Available on: www.diva-portal.se.

Ispra, 2016. Rapporto rifiuti urbani. Available on: http://www.isprambiente.gov.it.

Meyer M. W. & Gupta V., 1994. The performance paradox. Research in Organizational Behavior.

OECD, February 2015. OECD Economic Surveys: Italy 2015. OECD Publishing, Paris.

Sanger M.B., 2012. Does measuring performance lead to better performance? Journal of Policy Analysis and Management, 1-18.

Simon J.M., 26-29th march 2010. Stirring paper. Second Conference on Economic Degrowth for Ecological Sustainability and Social Equity, Barcelona.

Sterman J.D., 2000. System thinking and Modeling for a complex World. McGraw-Hill.

Studio di progettazione e consulenza aziendale Dott. V. Marinello, 2014. Progetto Area di Raccolta Ottimale (ARO): comuni di Palazzo Adriano, Prizzi, Bisacquino, Giuliana, Chiusa Sclafani.

Van Thiel S. & Leeuw F.L., 2002. The performance paradox in the public sector.Public Performance & Management Review, 25(3), 267-281.

Von Bertalanffy L., 1968. General System Theory. Foundations, Development, Applications. Braziller, New York.

WorldWatch Institute, 15th april 2013. State of the World: is sustainability still possible? Island Press, 1st edition.

UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 'in fase di assegnazione', volume I, 2017

Some features of deindustrialisation in EU15 during the period 1999-2004: a multivariate analysis

Maria Davì

Abstract During the past 40 years, employment in manufacturing, as a share of total employment, has steadily fallen in the world's most advanced economies. In this paper the causes and the different modalities of deindustrialisation process are examined closely with reference to 13 of the first EU15 countries. The aim of the analysis is to empirically explore if deindustrialisation is primarily a natural outcome associated with the development of modern societies or, on the contrary, can be defined a symptom of the failure of a country manufacturing sector. To this end, the Analysis of Principal Components (PCA), applied on a n-way matrix, is used to obtain a synthesis of the factors influencing deindustrialisation. The multivariate analysis was carried out on the Eurostat data for the period 1999-2004, at a disaggregated level of manufacturing (i.e. Nace divisions), considering simultaneously the time, the regional and the sectorial effects.

The estimates indicate that the main factors responsible of the dynamics of deindustrialisation have been efficiency and the scale of production processes in various manufacturing activities.

Keywords European Union · Industry Studies and Structural Change · Deindustrialisation · Principal Component Analysis

Riassunto *Il fenomeno della deindustrializzazione, che ha connotato i sistemi economici più avanzati è stato analizzato, fin dagli anni '70 del secolo scorso, dal punto di vista macroeconomico, ma i mutamenti che avvengono a livelli disaggregati dell'attività produttiva consentono di distinguere le modalità in cui si è manifestato questo processo. Gli economisti hanno manifestato una certa preoccupazione di fronte ai cambiamenti avvenuti nella struttura produttiva*

Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo
viale delle Scienze ed. 13, 90128
E-mail: maria.davi@unipa.it

ritenendo che il fenomeno, attribuibile a cause diverse (perdita di competitività, crowding out ad opera del settore pubblico, aumenti della produttività del lavoro nel manifatturiero), possa determinare una perdita di benessere nei Paesi interessati. In letteratura sono state formulate altre interpretazioni che convergono nel considerate la deindustrializzazione una conseguenza naturale dello sviluppo di un'economia postindustriale che riflette differenti condizioni della domanda e dell'offerta nel lungo periodo. In generale, negli studi effettuati sono state impiegate analisi univariate a livello macro, per valutare la portata e gli effetti della deindustrializzazione nelle economie "mature" mancando di produrre una visione complessiva delle dinamiche che si sono registrate. Ma un'analisi condotta mediante singole proxy non consente una disamina congiunta di tutti gli elementi pertinenti, né può spiegare come il processo avvenga presso realtà territoriali e periodi storici differenti. Per quanto rilevato, nel presente lavoro, è stata condotta un'analisi multivariata per ottenere una descrizione del processo di deindustrializzazione valutando il fenomeno nella sua interezza e, altresì, la sua dipendenza dai fattori interni al settore manifatturiero. In una prima fase, mediante l'Analisi delle Componenti Principali (ACP), applicata a una matrice a più vie, è stata identificata la struttura latente del manifatturiero nel suo complesso, con riferimento a 13 dei primi Stati membri dell'UE, i soli di cui erano disponibili i dati completi per il periodo 1999-2004. Nello specifico, l'ACP è una tecnica esplorativa che consente di ottenere un'informazione sulla struttura interna dei dati per identificare gli aspetti più significativi (in termini di variabili latenti e punti compromesso) che influenzano l'intero fenomeno in esame. I risultati ottenuti, in termini dei principali fattori e della relativa dinamica, hanno indotto successivamente a tracciarne l'evoluzione a un livello di disaggregazione che tiene conto della diversità delle attività comprese nel manifatturiero, permettendo al tempo stesso un confronto diretto tra i singoli Paesi, riguardo alle cause e alle modalità dei processi di deindustrializzazione. Per raggiungere gli obiettivi preposti, sono stati impiegati, per ognuno degli Stati membri, i dati delle 23 divisioni (a due cifre), che la classificazione NACE prevede per le attività manifatturiere, relativamente all'intervallo 1999-2004, il solo periodo per cui l'EUROSTAT fornisce l'informazione statistica completa a questo livello di disaggregazione. Dall'analisi effettuata è emerso che le variazioni registrate possono essere spiegate soprattutto da due fattori: l'efficienza e la scala delle operazioni. Nel periodo in esame si nota che la prima è aumentata mentre la seconda si è ridotta. Si può, conseguentemente, dedurre che la deindustrializzazione si accompagni in modo prevalente all'incremento nei livelli di efficienza dei processi produttivi e, in misura minore, alla diminuzione della scala a cui vengono svolti gli stessi. Il risultato riguardante il manifatturiero europeo nel suo complesso si è riprodotto a tutti i livelli di disaggregazione considerati (territoriale e settoriale) rimarcando il carattere di generalità che contraddistingue la struttura latente identificata, seppure con differenze connesse alle specificità dei vari contesti a cui si si riferisce l'analisi.

Parole chiave *Unione Europea - Cambiamenti della struttura industriale - Deindustrializzazione - Analisi delle Componenti Principali*

1 Introduction

Since the early 1970s a great amount of research has been carried out into the process of deindustrialisation occurring in many advanced economies.

In those countries the relative importance of the industrial sector has decreased due to a decline in production, employment and investments.

According to Blackaby (1978), Jones and Lee (1985), Stanners (2001) and Gallino (2003), this fact has been caused by a loss of competitiveness within the industrial sector and could negatively affect future generations. In this context, the symptoms of deindustrialisation are recognizable from the reduction in the share of the workforce employed in industry and/or in the share of added value on GDP. In many cases it is also evident from the growing value of manufactured imports in the trade balance and in the contemporary decline in manufactured exports (Singh, 1977).

Other authors (Rowthorn, Wells, 1987; Baumol *et al.*, 1989) interpreted deindustrialisation as the natural consequence of development in a post-industrial economy, reflecting the different conditions of demand and supply arising in the long term. More specifically, according to the "stages of growth" hypothesis deindustrialisation is a long-term phenomenon explained by the progressive advantage of the tertiary sector over industrial activities, caused by the shift in domestic expenditure (Clark, 1957; Gershuny, 1978; Gemmel, 1986).

Both explanations, despite being based on different premises, employ the same instruments to deal with the subject: i.e. macro-level univariate analyses whose aim is to evaluate the effects of deindustrialisation in various economies.

The concept of deindustrialisation is, nevertheless, herete complex and an analysis carried out by means of a single proxy doesn't allow a full investigation of all the relevant issues; nor it can explain how the process takes place.

On the contrary, in this paper a multivariate analysis is used. Specifically, a *n-way* PCA is applied to the data of European countries by considering simultaneously the time, the regional and the sectorial effects.

According to this method it is possible to identify the key factors that influenced the manufacturing sector dynamics in EU15, before the enlargement occurred in 2004 when other 10 countries were admitted to take part in this supranational institution.

Specifically, I considered the manufacturing sector of each EU15 country from 1999 to 2004, the only period for which a complete information is at disposal, at the chosen disaggregation level (NACE Divisions).

The empirical analysis has been performed on the data from industrial Censuses recorded by the European Statistical Institute (EUROSTAT). Consequently, only the old member states were considered over the period 1999-2004 to understand the causes and, possibly, the main effects of deindustrialisation.

The paper is organised into five parts: section 2 gives an overview of the literature on the issue and outline the productive dynamics of whole EU15, as reflected by some economic indicators; section 3 features a brief explanation of the technique known as "Principal Component

Analysis" (PCA) applied to a *n-way* matrix; in section 4 the PCA technique is applied to the European manufacturing data; finally, concluding remarks are presented in section 5.

2 An overview of the debate

2.1 Limits of the different theories on deindustrialisation

Apart from conceptual standpoints, the difficulty in accepting explanations provided by the economists, about the source and effects of deindustrialisation, depends on several factors:

a) the scope of analysis. The arising question is: Should the research address the industrial sector as a whole or just be limited to manufacturing?

b) the level of the analysis. In this regard it should be noted that the results can be different depending on the selected disaggregation level;

c) variables used in the measurement. Even the measures used for the analysis (employment, production, value added, investments) can represent a source of errors or uncertainty in the evaluation of the process, particularly if some variables are affected by the influence of inflation[1];

d) the business relations with other economic systems at a different development level.

Regarding a) Ferguson (1988, p.169) considered that "*any of the measures taken in isolation [from the rest of the economy] (as in often done) presents but a partial view*". Thus, the analysis rehereres the simultaneous consideration of the interdependence of all sectors because "*the decline in the industrial sector occurs as market forces reallocate resources differently*"(p.172) reflecting the changes in demand and supply conditions.

Also for the remark at point b) Ferguson (p.178) stated that *"the macroeconomic viewpoint adopted by many economists analysing deindustrialisation is too crude"* and many differences inside the industrial sector are neglected. Actually, the industrial sector includes different kinds of activities that are subject to a wide variety of influences and which have dissimilar performances.

This opinion is also supported by research carried out into the changing industrial composition of employment due to "*the outsourcing to the service sector of activities previously undertaken by employees in the manufacturing sector*" (Watts, Valadkhani 2001, p. 2), with the consequent growth in the relative importance of the tertiary sector.

[1] According to Rodrik (2015, p.1), "*manufactures output at constant prices has held its own comparatively well in the advanced world, something that is typically overlooked since so much of the discussion on deindustrialization focuses on nominal rather than real values*".

Regarding c), the process of deindustrialisation is normally measured by just a fewvariables: the share of manufacturing employment (Penava, Družić, 2015; Rodrik, 2016), the share of manufacturing value added in GDP (Rowthorn, Ramaswamy, 1999; Rowthorn, Coutts 2013; Grodzicki, 2014; Palma 2014) or by setting the real GDP index against the Index of Industrial Production (IPI hereafter) (Stanners, 2001).

Owing to the lack of direct measures of deindustrialisation, the dynamics underlying the phenomenon doesn't emerge clearly. Therefore, it is impossible to draw unambiguous conclusions about its effects (particularly in the long term).

Finally, about d), it is necessary to take into account the commercial relations with other countries. In the opinion of Rowthorn and Ramaswamy (1997) deindustrialisation is also caused by external factors as foreign trade, that affects the internal structure of an economy in various ways.

Specifically, international relations can impact on manufacturing employment in countries that export goods characterised by high levels of technology and import goods from "*the new competitors, combining low labour costs with large productivity levels*", thus lowering "*international prices for manufactured products*"[2] (Boulhol, Fontagné 2006, p.9).

Concerning the external effects, Brady and Denniston (2006, p.297) consider the effects of globalization on deindustrialisation finding that "*globalization has a curvilinear, inverted U-shaped relationship with manufacturing employment".* Their results, however, are not unequivocal because of "*different varieties of capitalism, regions and historical periods*" of the countries where the phenomenon occurs.

Indeed, also being typical of western economies, a "premature deindustrialization" (Palma, 2014) has recently characterized many developing countries that "*are turning into service economies without having gone through a proper experience of industrialization*"(Rodrik, 2016, p. 2).

It is noteworthy that the evaluation proposed by Mickievicz and Zalewska (2006, p. 159) for European eastern countries does not consider in a negative way the phenomenon because "*given lower capital rehereements and possible competitive advantage of the transition economies in terms of human capital, the greatest opportunities for generating productivity increase and economic growth may be in the service sector*".

According to Penava and Družić (2015, p. 845), also now "*deindustrialisation at lower levels of income in formerly socialist (developing) countries is still mainly uncharted territory."* However, the Autors conclude (p. 851) that the process of deindustrialisation in Croatia is characterised by different factors relative to developed countries because "*deindustrialisation occurred just after Croatian independence"* and was due to *"profound structural differences"* compared with western economies.

[2] To produce the first type of goods it is necessary to employ a small amount of skilled labour while the same value of other goods rehereres a widespread use of cheap labour in capital-saving industrial processes. Such a framework of foreign trade, consequently, brings about a scaling down of employment in the manufacturing sector of the most advanced economies.

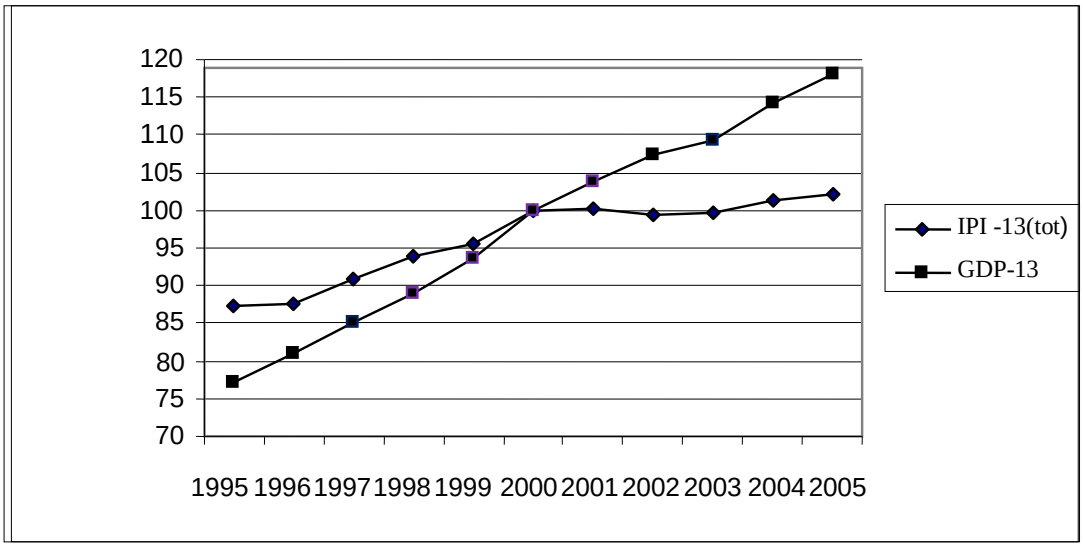
The same opinionis shared by Rodrik (2015, p.1) when asserts that "*in the developing countries trade and globalization likely played a comparatively bigger role*" against the presumptive causes of deindustrialisation in the advanced countries.

Lastly, the analysis by Palma (2014, p.19) proves that "*countries of the former Soviet Union and Eastern Europe, [. . .] experienced a process of de-industrialisation associated with a fall in income per capita that was associated with a reduction in manufacturing employment backwards: a case of 'reverse' de-industrialisation*".

2.2 A general view of the economic structure of EU15

Following the suggestion of Stanners (2001), the analysis begins with a comparison, referred to the period 1995-2005, of real GDP and IPI of 13 EU countries[3].

FIG.1. *Dynamics of GDP and IPI in EU13 (2000=100)*



Source: EUROSTAT: *Selected indicators for all activities* (1995-2005).

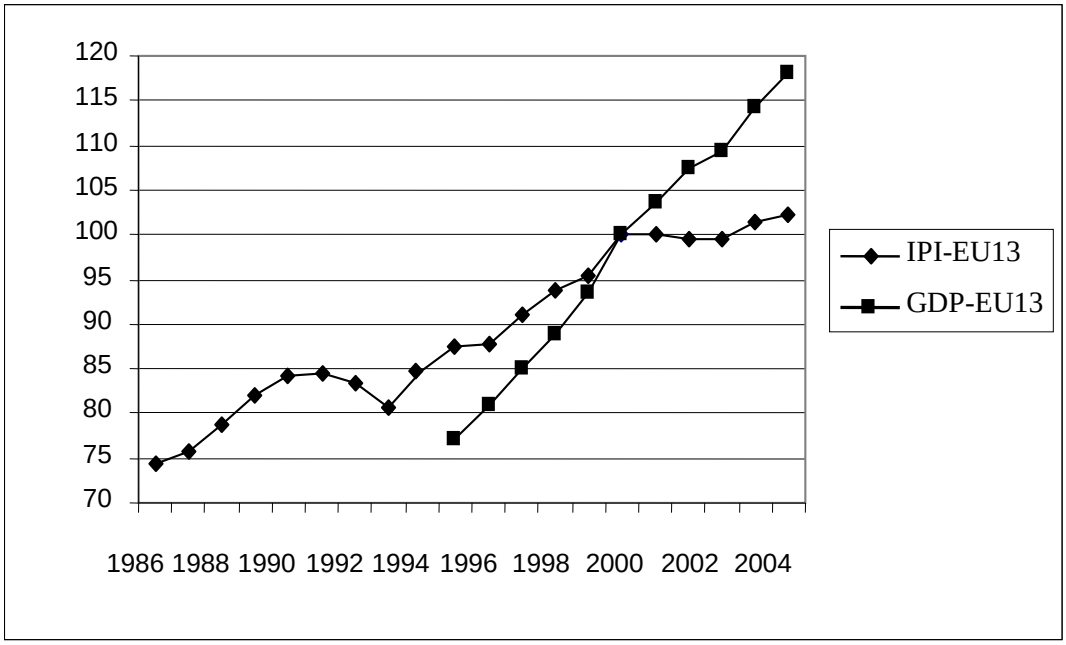
In Fig. 1, two periods can be traced out: one (1995-2000) where IPI tracks GDP and the other (2000-2005) where there is a clear divergence. Precisely, in the first five years, GDP shows an average increase of 5.35% *per annum* while in the six last years this drops at an average yearly rate of 3.38%.

[3] For this period, Eurostat data only enable a full comparison for thirteen of the EU15 Member states as information for Greece and Luxemburg is incomplete.

On the other hand, IPI in the first period grows by 2.73% *per annum* while in the second one the yearly average increase is only 0.43%. In other words, the industrial sector grows less than the total economy and in the last years it raises much more slowly.

The shortness of the period considered does not show an adequate picture of the dynamics that hasshaped the EU's economy; so the analysis has been extended to the interval 1986-2004 but only for IPI, the only variable supplied by Eurostat for this time span with regard to the 13 EU countries.

FIG.2. *Evolution of GDP and IPI in EU13 (2000=100)*



Source: Eurostat: Selected indicators for all activities (1986-2004).

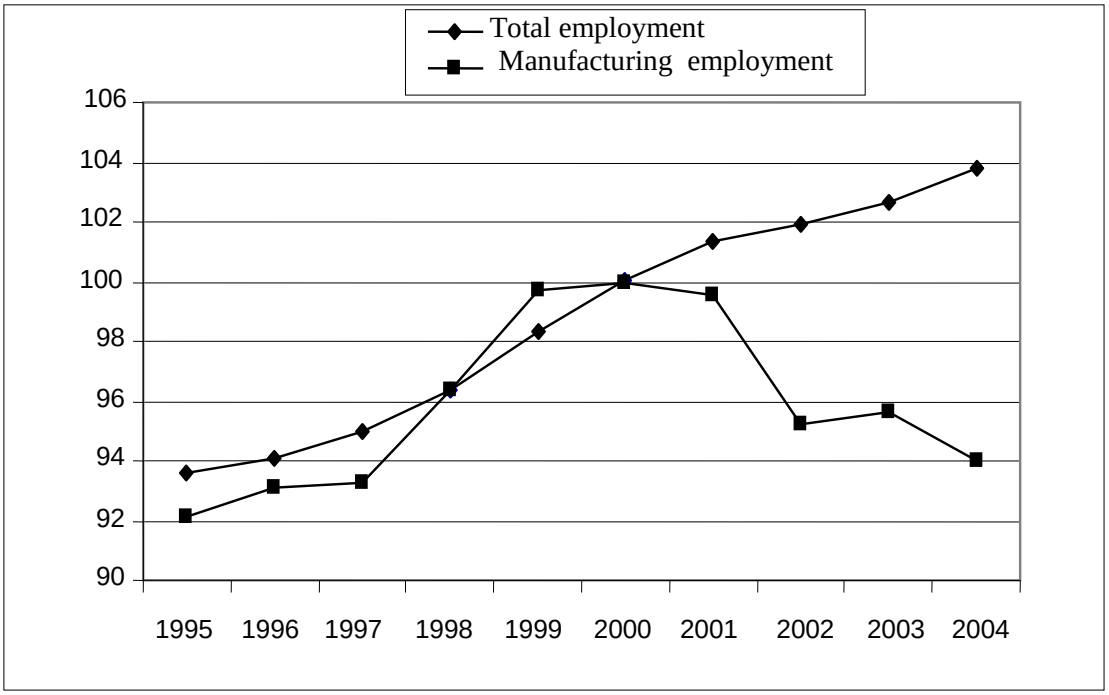
From Fig. 2 the diversity in trends between GDP and IPI emerges more clearly.

The evolution of GDP, as a measure of global production activity, strongly depends on the service sector dynamics. The dissimilar behaviour of the two indicators rehereres a clearer verification to find new evidence of a prospective process of deindustrialisation.

For this reason, manufacturing employment was compared with total employment for the period 1995-2004. Fig. 3 shows that the two lines present a very similar evolution until 2000 but, thereafter, the discrepancy between the two trends is even more noticeable than the one detected in the previous matching.

Indeed, while total employment in EU countries continues to rise at about 1% *per annum*, since 2000 the number of workers employed in the manufacturing sector declines[4] at an average yearly rate of 1.53 %.

FIG. 3. *Evolution of total and manufacturing employment in EU13 (2000=100)*



Source: EUROSTAT: Selected indicators for all activities (1995-2004).

Fig. 3 shows in a more apparent way the divergence between manufacturing employment and global employment.

Given the complexity of the process of deindustrialisation, it is useful to have an empiric tool producing a global synthesis and, at the same time, an evaluation of the alterations sustained by the productive structure(s) in the EU, at an intermediate level between macro and micro, with each matching the other.

In the opinion of Ferguson and Ferguson (1994, p. 257), "*[d]isaggregated data need to be considered if the changes occurring are to be better understood*".

For this reason, here a more complete analysis of the available information has been performed in order to consider the composition and the dynamics of European manufacturing structure, with the aim of identifying the causes of deindustrialisation.

[4] [4] It is worth noting that Rowthorn and Ramaswamy (1997) had already outlined that in the EU15 countries "*the share of manufacturing employment stood at a comparatively high level of more than 30 percent in 1970 but then fell steeply to only 20 percent by 1994* ".

Now the purpose is to find a plausible answer to reconcile the prevailing opinions on the topic by analysing the internal dynamics of the EU industrial sector by means of a simple strategy highlighting the principal features of the productive structure of the member states. More specifically, in the event that a "weakening" of manufacturing - in terms of its sectorial composition - can cause some concern for future growth, the analysis should "also" be carried inside this sector, through a comparison of the different activities constituting it.

3 The n-way Principal Component Analysis

Without an *a priori* model, the (PCA) applied to a *n-way* matrix extracts from a complex set of data the main aspects of a phenomenon, due to the time, space and structure.

The principles on which this analysis is based (Bolasco, 1999) are:

Thriftiness in the representation (by mathematical models and graphics) of a data set reduced to few meaningful dimensions;

Fundamental strength of analysis, as it is possible to highlight the data's latent structure also with data showing random measurement errors and in the absence of distributive links;

Immediate visibility by means of graphic representations as these may help researchers, whose statistic knowledge is elementary, to be autonomous in interpreting – by using scatter-plotting - the results obtained from the analysis.

The starting point for the analysis is constituted by the construction of a metrical data matrix (units per attributes) from which a similarity matrix is obtained (generally a correlation or variance and covariance matrix) among variables (or among units). The purpose is to transform connections between variables into indirect relations due to the action of few sufficiently informative PCs (Principal Components or factors) reproducing a suitable sinthesis of the original information.

Generally, the PCA "reduces the complexity" of reality and has the double aim of both simplifying the interpretative models and achieving a conceptual clarification that allows a data reduction.

A two-way orthogonal PCA, applied to correlation matrix **X'X**, may be represented through the following matrix notation:

(1) $\mathbf{X} = \mathbf{FA'} + \mathbf{E}$

where:

$\mathbf{X} = [x_{mi}]$ is the data matrix with $m = 1, 2, \ldots, n$ observations and $i = 1, 2, \ldots, p$ observed variables

$\mathbf{F} = [f_{mj}]$ with $j = 1, 2, \ldots, q$, is the matrix of latent variables normalised or PCs, with $q \ll p$;

$\mathbf{A'} = [a_{ij}] = [\ \bar{r}_{ij}]$ is columnwise orthonormal, and is the correlation coefficients estimates matrix measuring the existent similarity between each variable and each PC;

$\mathbf{E} = [e_{mi}]$ is the specific component matrix or error variables.

The application of this model has a double purpose:

(a) it expresses each variable as a linear function of the single principal components, each one with its own PC coefficient, plus an error component (e_i):

(2) $x_i = \sum_{j=1}^{q} a_{ij} f_j + e_i$

(b) it expresses the single principal components as a linear combination of all the observed variables, each one with its own loading component *(w_{ji})*:

(3) $f_j = \sum_{i=1}^{p} w_{ji} x_i$

where w_{ji} is the weight of each principal component.

The purpose is to minimize the sum of squared residuals: $\sum_{m=1}^{n}\sum_{i=1}^{p} e_{\mathrm{mi}}^2$.

In matrix algebra, this result is obtained as $|X - \mathrm{AF}'|^2$, which is Pearson's description of PCA as a technique for identifying the ordered components that can explain the maximum amount of variance in the data.

The generalization of PCA applied to a *n-way* matrix data set goes back to the mid 1960s and was introduced by Tucker (1966).

Within the Tucker tecnique it is possible to define a 4-index matrix $[\mathbf{X}_{mzti}]$ (where m = number of observations, z = number of objects, t = number of occasions and i = number of variables) from which to extract the matrix decomposed in PCs (Vandeginste *et al.*, 1998; Kiers, Mechelen 2001):

(4) $\mathbf{X} = \mathbf{X'} + \mathbf{E}$

where $\mathbf{X'}$ = S*L, $\mathbf{E}$ = error matrix representing the unexplained part of $\mathbf{X}$,

S = scores matrix, L = loadings matrix

Here $\mathbf{X'}_{(m,z,t,\ j)}$, with j = number of the new latent factors (PCs) and $j \lll i$.

In summary, by applying the PCA from the initial matrix, $\mathbf{X}_{(mtz)i}$, it can be obtained a matrix of latent variables, $\mathbf{F}=[f_{mtz,j}]$, and the reconstructed pooled matrix of latent variables, $\tilde{X}_{(\mathrm{mtz})j}$, allowing to calculate separate average values (compromise points) (Rizzi, Vichi 1995).

As the purpose, in the present analysis, is producing complex indexes that can measure changes by latent structure of manufacturing in the EU countries, cubic matrices, $\tilde{X}_{(\mathrm{mtz})j}$, are pooled together to obtain the following 2-*way* matrices:

(5) $\tilde{X}^{a_{(\bar{m})j}}$; $\tilde{X}^{b_{(\bar{z})j}}$; $\tilde{X}^{c_{(\bar{t})j}}$;

where: m = countries, z = manufacturing divisions, t = years and j = latent variables.

4 Empirical Analysis

4.1 PC Labels and Compromise points

The 4-*way* matrix PCA is applied on the following variables:

V1 - Gross operating surplus/Turnover (Gross operating rate) (%)

V2 - Labour cost per employee (Unit labour cost)

V3 - Gross value added per employee (Labour productivity)
V4 - Investment per employee
V5 - Employees in Manufacturing/Total employees (%)
V6 - Gross investment in tangible goods/Turnover (%)
V7 - Value added at factor cost/Turnover (%)
V8 – Turnover per employee in Manufacturing

The data refer to 23 manufacturing divisions (tab. 1) in each EU13[5] country in the period 1999 and 2004, the only time span for which Eurostat provides a complete information at this disaggregate level.

In order to obtain the latent dimensions it was necessary to stack the original 4-*way* matrices in one pooled matrix. By applying the PC method to the supermatrix $\mathbf{X}_{zmt,i}$, where z are 23 manufacturing divisions, m 13 geographical areas, i 8 variables and t 6 years, it is possible to obtain a similarity matrix $\mathbf{M}_{i,i}$ which is, in this case, the Bravais–Pearson's correlation coefficient matrix. From the last one, the component loadings matrix $\mathbf{A}_{i,j}$ with $j \lessdot\lessdot i$ has been extracted.

Specifically, three PCs have been drawn with eigenvalues higher than 1 and explaining about 64% of the whole variance. In Table 2, the PC loading coefficients are shown.

The first component, strongly correlated with Labour cost per employee, Gross value added per employee, Investment per employee and Turnover per employee, explains 35% of total variance and represents the "efficiency" of the sector. Indeed, this factor proves to be the most important for the analysis.

Precisely, it shows how investment intensity[6] is strictly correlated with the performance of labour, both in terms of productivity and remuneration.

The second component groups the Gross operating surplus/Turnover ratio and the Share of workforce employed in manufacturing; therefore it can be indicative of the "scale" of productive activity in the sector, explaining about 16% of total variance. On the basis of the loading coefficients, it is evident that the first variable is more influential on the identified component than the second one.

The third component, explaining 13% of total variance, aggregates the ratios: Gross investment in tangible goods/Turnover and Value added at factor cost/Turnover. This component can be considered a measure of the "vertical integration"[7].

[5] See note no. 3

[6] According to Caves and Barton (1991), gross capital expenditure has a significant positive impact on evolution of labour productivity, especially when the new eherepment installed embodies R&D and innovation in supplier industries.

[7] The vertical integration can be chosen by the industrial firms to lower the transaction costs owing to the technological interdependence among various manufacturing activities, to reduce uncertainty in the supply of

TAB. 1. *Divisions of Manufacturing*

CODE	Divisions
DA15	Manufacture of food products and beverages
DA16	Manufacture of tobacco products
DB17	Manufacture of textiles
DB18	Manufacture of wearing apparel; dressing; dyeing of fur
DC19	Tanning, dressing of leather; manufacture of luggage
DD20	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
DE21	Manufacture of pulp, paper and paper products
DE22	Publishing, printing, reproduction of recorded media
DF23	Manufacture of coke, refined petroleum products and nuclear fuel
DG24	Manufacture of chemicals and chemical products
DH25	Manufacture of rubber and plastic products
DI26	Manufacture of other non-metallic mineral products
DJ27	Manufacture of basic metals
DJ28	Manufacture of fabricated metal products, except machinery and equipment
DK29	Manufacture of machinery and equipment n.e.c.
DL30	Manufacture of office machinery and computers
DL31	Manufacture of electrical machinery and apparatus n.e.c.
DL32	Manufacture of radio, television and communication equipment and apparatus
DL33	Manufacture of medical, precision and optical instruments, watches and clocks
DM34	Manufacture of motor vehicles, trailers and semi-trailers
DM35	Manufacture of other transport equipment
DN36	Manufacture of furniture; manufacturing n.e.c.
DN37	Recycling

Source: Eurostat - Classification NACE Rev.1.1

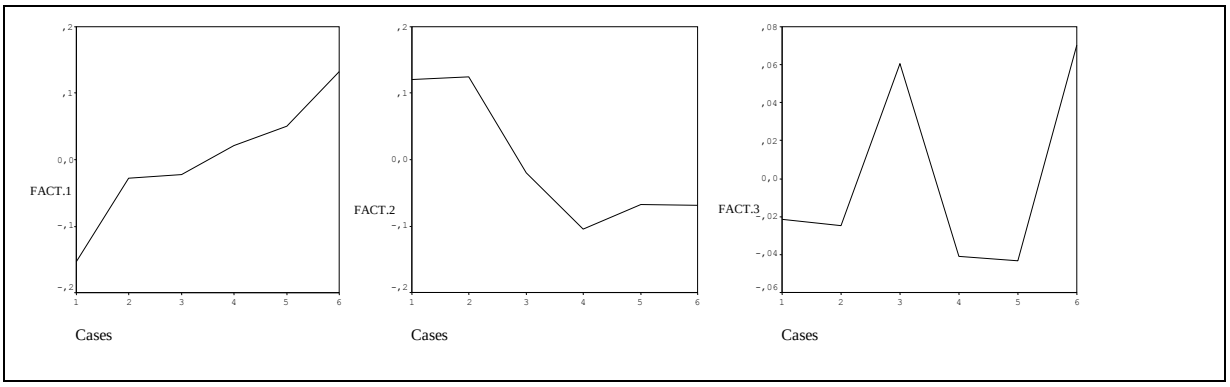
important inputs, to avoid government price controls and taxes in different stages of production process and, not least, to eliminate the possible information asymmetry between upstream and downstream producers.

TAB. 2. *Variables and Factors* (1999-2004)

Economic Indicators	Factor1	Factor 2	Factor 3
Gross operating surplus/Turnover (%)	0,154	**0,854**	-0,201
Labour cost per employee	**0,772**	-0,216	0,101
Gross value added per employee	**0,889**	0,278	-0,049
Investment per employee	**0,767**	0,144	0,0003
Employees in Manufacturing/Total employees (%)	-0,281	**0,471**	-0,133
Gross investment in tangible goods/Turnover (%)	-0,017	0,077	**0,802**
Value added at factor cost/Turnover (%)	-0,079	0,365	**0,581**
Turnover per employee in Manufacturing	**0,856**	-0,187	0,022

Source: Elaboration from Eurostat database: *Selected indicators for all Activities* (NACE *divisions*)

FIG. 4. *Trends of the factors identified by PCA inside European manufacturing in the period 1999-2004 (Fact.1= Efficiency; Fact.2=Scale; Fact.3= Vertical integration)*



In Fig. 4 the first two latent variables show a clear linear tendency while the vertical integration shows a cyclical behaviour not allowing an interpretation relevant to our analysis.

In view of the less importance that vertical integration holds as third factor, the evolution of this feature of the EU manufacturing structure has not been examined closely; consequently only the dynamics observed on the first factorial plane will be expounded.

However, the results of the present analysis do have some significance allowing to sketch out the more important characteristics and the dynamics of European industrial activity, also at a disaggregate level.

FIG.5. *The dynamics of European manufacturing (1999-2004)*

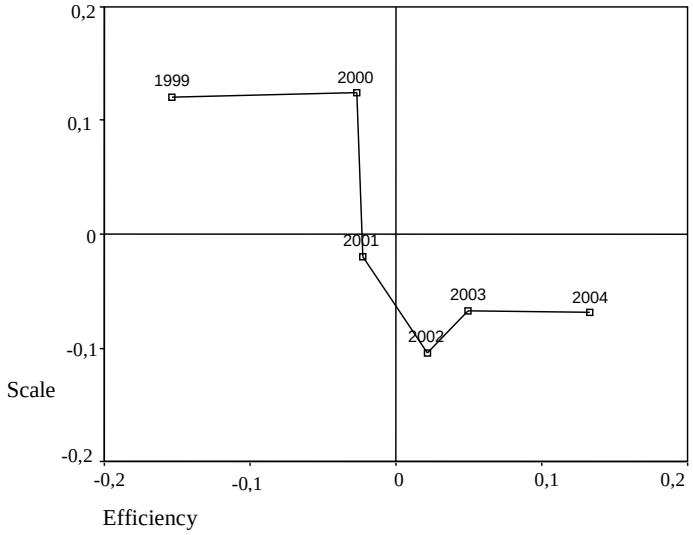


Fig. 5 reproduces the compromise points (scatterplot) related to the years considered in this part of the analysis. It gives a synthetic description of the changes that characterised the manufacturing sector in the whole EU13 in that period.

In greater detail, this branch of industrial activity experienced a downsizing between 2000 and 2002 followed by a slight recovery in 2003; at the same time, it recorded an increase in efficiency levels which, in spite of a slowdown during the intermediate years (2000-2002), seemed to regain in the end of the period.

The trend depicted in the graph appears to be consistent with the results shown by the variables previously employed in the univariate analysis (see Figg. 1, 2 and 3), but now the information is more complete with a specific reference to two factors identified by the PCA in this stage of the analysis.

In fact, the decrease in size of the manufacturing sector in the whole EU, in terms of both employees and capital invested, was counteracted by an improvement in productivity levels supported by the rationalization of productive processes.

In addition, the PCA enables to disaggregate data both in territorial terms and in NACE divisions, maintaining the possibility of comparison between the different clusters. Also, when

the compromise points are not very numerous, the same graph may be used to show all the information in one go.

Subsequently, a more detailed analysis was carried out plotting the compromise points for the 13 countries considered.

FIG. 6. *Manufacturing in the individual EU13 Countries (1999-2004)*

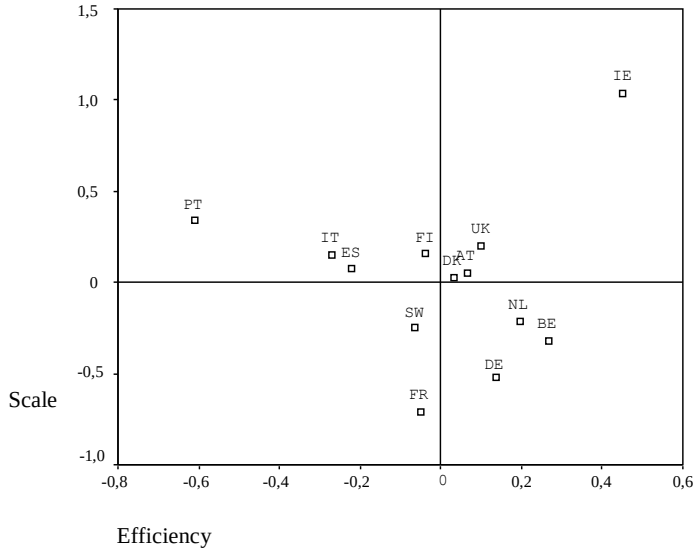


Fig. 6 shows a cohesive set constituted by Austria, Denmark, Finland and the United Kingdom. In thesecountries both manufacturing dimensions and performance levels are near the EU13 average in the period 1999-2004.

The nucleus formed by most of the member states shows dimension and efficiency levels a little beneath the average of EU manufacturing. Italy and Spain are not far away, presenting an alignment of dimensions to European mean value but reduced levels of efficiency.

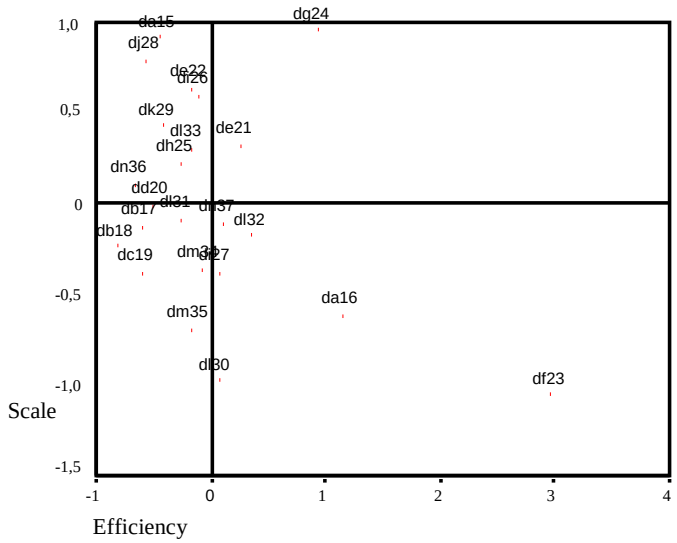
On the contrary, the Netherlands, Belgium and Germany reveal a dimensional structure that is below the European average but efficiency levels higher than average.

The position of Ireland is rather atypical[8]; its position, in the upper right part of the first quadrant, implies performance levels in manufacturing that are clearly above European averages for both factors.

Both France and Portugal are at a distance from the other countries; France because of its small manufacturing structure and Portugal because of its low efficiency levels.

[8] Grodzicki (2014, p. 109) verified for 1970 a similar unforeseen behaviour for this country since in that year "*the least developed countries of Western Europe (except for Ireland, which is an outlier in the case of manufacturing structure) specialised in low-tech activities, while manufacturing structure of more developed economies was clearly biased in favour of more advanced industries*".

FIG. 7. *Manufacturing divisions in EU13 (1999-2004)*



In Fig. 7 it is evident the cluster formed by the compromise points regarding the divisions of the manufacturing sector (NACE Rev.1.1 classification two-digit) for the whole EU countries.

Notably, a considerable number of the points are situated on the left side of the graph demonstrating that, in terms of efficiency, many partitions of manufacturing activities are below average levels.

Few divisions (those on the right side of the factorial plane) exceed the European average. In particular, the position of the Manufacture of chemicals and chemical products (division DG24) is indicative of an industrial activity associated with high dimensional values and good efficiency levels.

The other two divisions, the Manufactures of tobacco products (DA16) and Coke, refined petroleum products and nuclear fuel (DF23), show oddly a below-average dimensional structure but levels of efficiency that are higher than the mean. This is especially true in the latter sector, where efficiency levels are three points above the average.

This synthesis, based on divisions, rehereres a conceptual remark. In Fig. 7 a few divisions, some of which belong to the same subsection of manufacturing, show structural situations that are decidedly different and would be concealed if aggregated.

Two evident examples are subsections DA and DL whose compromise points should be situated in an intermediate position compared to the real localisation of the points of the constituent divisions (respectively, DA15 and DA16 and DL30, DL31, DL32, DL33), as shown in the graph.

4.2 Dynamic analysis of EU countries and manufacturing

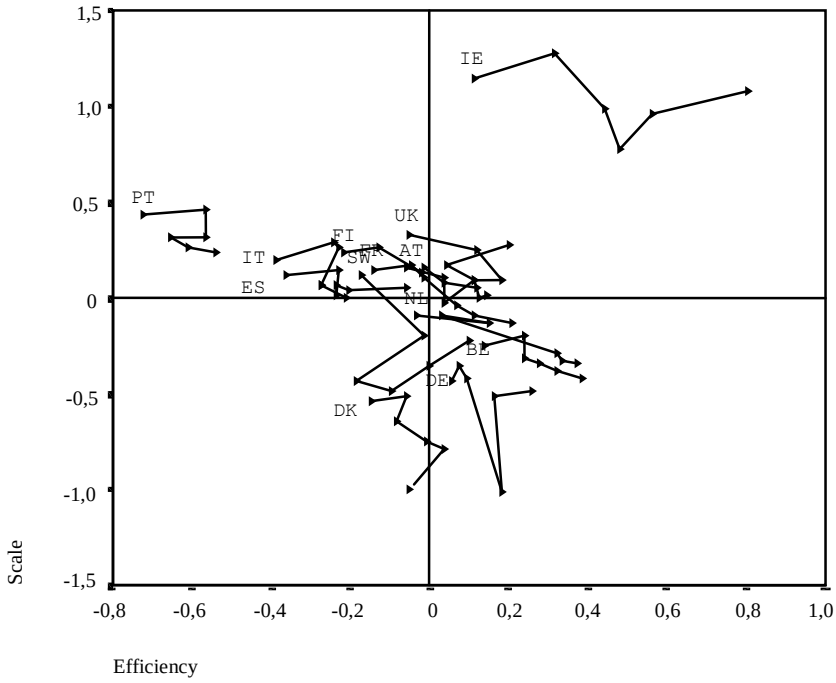
divisions

In order to analyse the dynamics of manufacturing in each country, it is very useful to explore the matrix of compromise points (Fig. 8).

The representation of compromise points is a synthesis from the reproduced matrix, $\tilde{X}_{(\mathrm{mtz})j}$, even if the raising of disaggregation on the estimates may cause some problems connected to error estimation.

Notwithstanding such a risk, it is useful to observe the evolution of manufacturing in each country during the period 1999-2004. Obviously, this information should be considered with great caution.

FIG. 8. *The evolutionary paths of manufacturing in the individual EU13 Countries (1999-2004)*



It should be noted that for nearly all the countries the pattern presented in Fig. 5 comes up once again, though with slightly different shapes, demonstrating that the dynamic analysis displays very similar territorial situations: a general reduction in the scale of industrial activity while the production processes turned out to be more efficient in most member states.

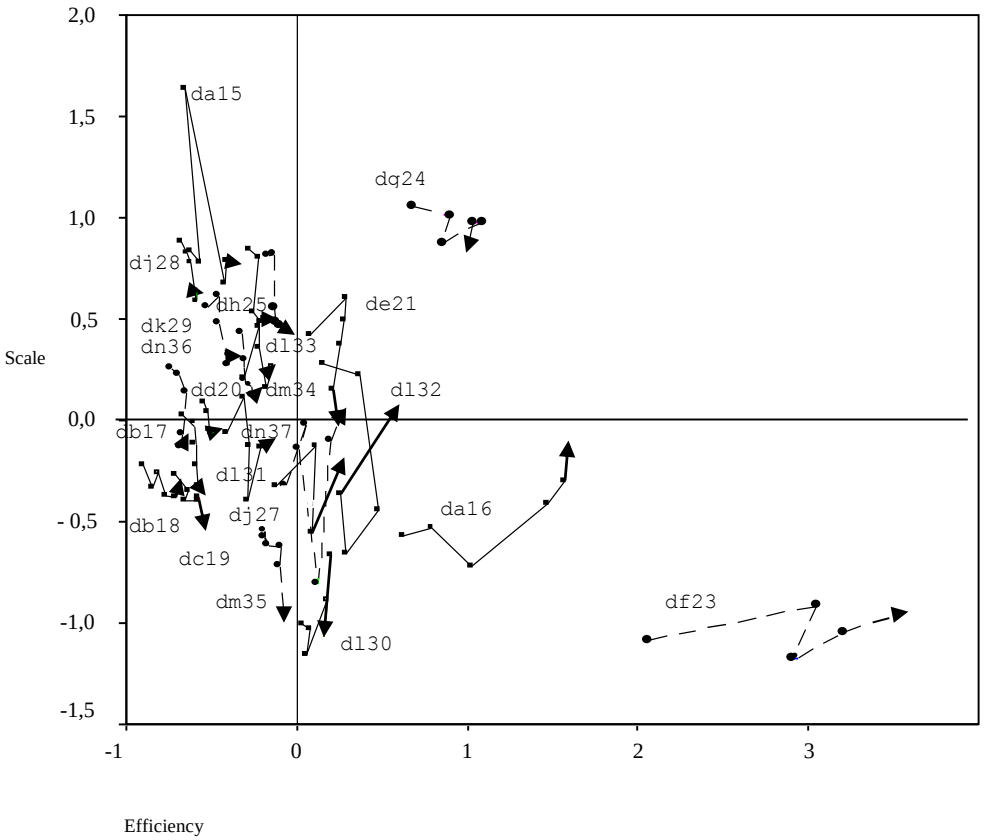
In other words, old "*EU member states became more and more homogenous in terms of economic structure – at a high level of aggregation*" (Grodzicki, 2014, p. 94).

Notwithstanding this growing homogeneity, the tracks of Ireland and Portugal are moving away from the nucleus of EU countries; France and Germany too show a similar behaviour, though to a lesser extent (see also Fig. 6).

From the above outcome, it follows that the time span does not significantly change the pattern presented in Fig. 5, namely the overall average of all countries. Therefore, the trends of the principal factors identified by PCA during the period (Fig. 4) did not show any substantial alterations in none of the manufacturing sectors of EU countries, except for Ireland, whose efficiency and scale values rose to levels much higher than the average.

What is more, the dynamics of the various manufacturing divisions in EU (Fig. 9), reveals that their trajectories are intertwined, localizing along the dimensional axis rather than on the efficiency one except, once again, for divisions DA16, DF23 and DG24 (see Fig.7).

FIG. 9. *The dynamics of manufacturing divisions in EU in 1999-2004*



For an in-depth examination, results regarding the divisions have been grouped in three macro-sectors according to the well-known Pavitt taxonomy, established on the basis of a descriptive analysis of the innovation process in different industrial activities (Figg. 10-12).

FIG. 10. *The dynamics of manufacturing divisions "Supplier dominated" (1999-2004)*

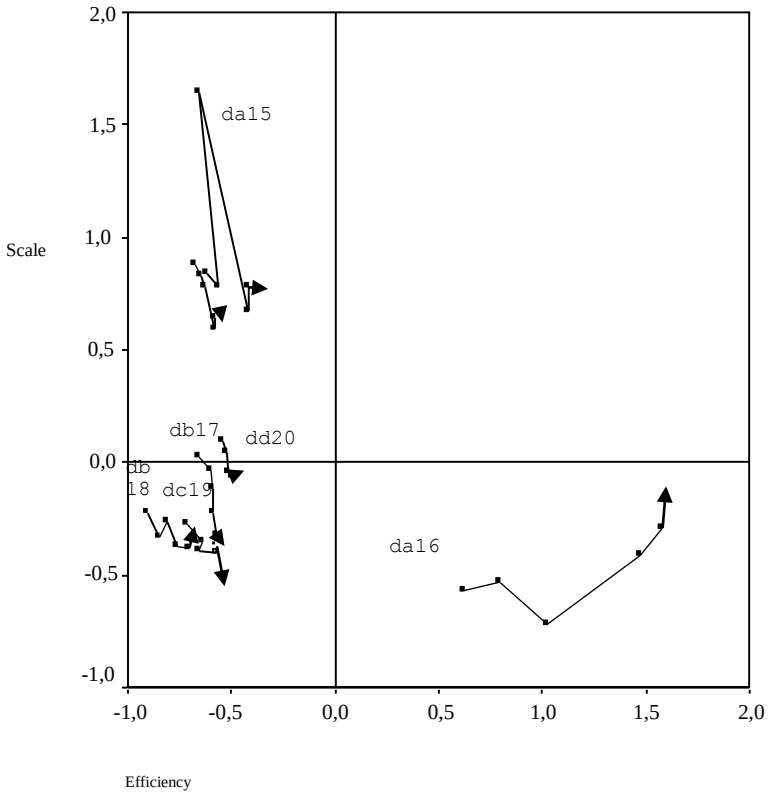


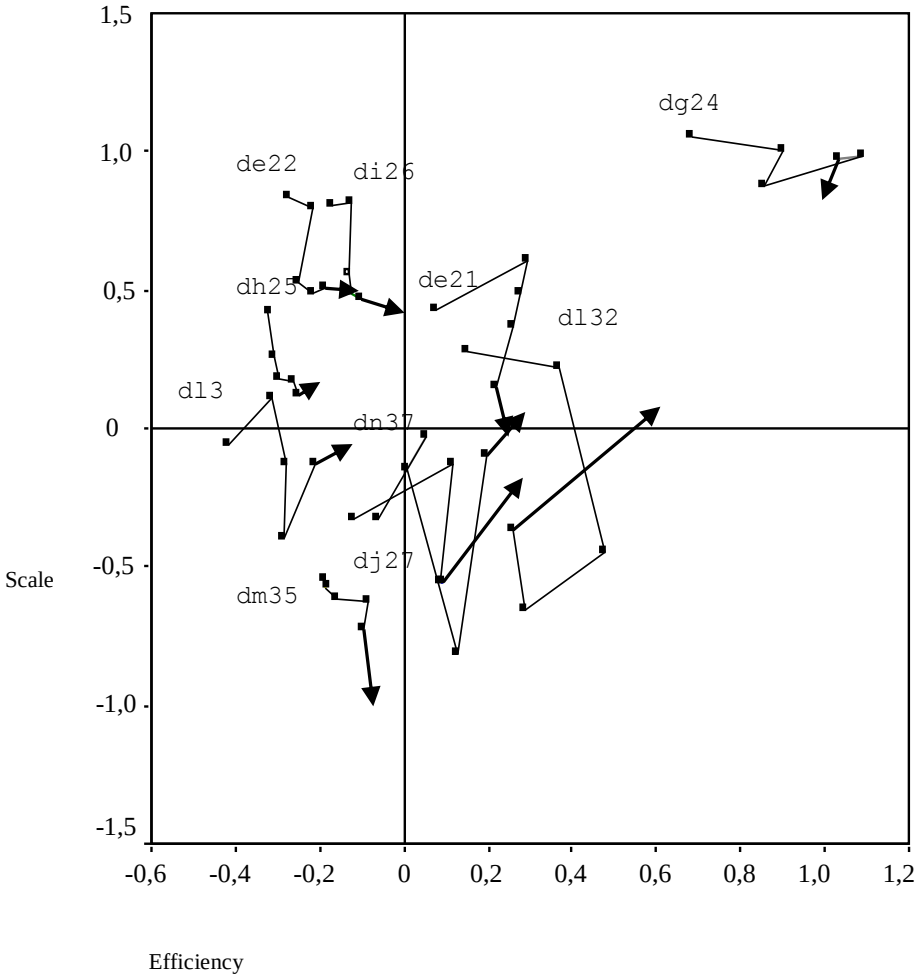
Fig. 10 shows the evolution of the divisions belonging to the category "Supplier dominated", characterized by small and medium-sized companies operating in traditional manufacturing activities and buying their technology from external sources.

The relevant trajectories are all situated on the left side of the plane thus demonstrating, for the entire period, lower efficiency levels than those for the whole EU manufacturing. Only the manufacture of tobacco products (division DA16) shows higher and ever-increasing efficiency values.

Fig. 11 shows the evolutionary paths of divisions belonging to the macro-sector, "Scale intensive". In this group, we generally find large companies producing standardized bulky products (such as steel, glass, etc.). Their technology is generally developed in-house but may also be purchased from suppliers.

The trajectories depicted in the graph show a greater dynamics in[1] terms of efficiency, in comparison to the preceding ones. Besides, the values of these divisions are near the European mean while a few (i.e. DE21, DL32 and DN37) manifest higher efficiency levels. One aspect of

FIG.11. *The dynamics of divisions "Scale intensive" (1999-2004)*

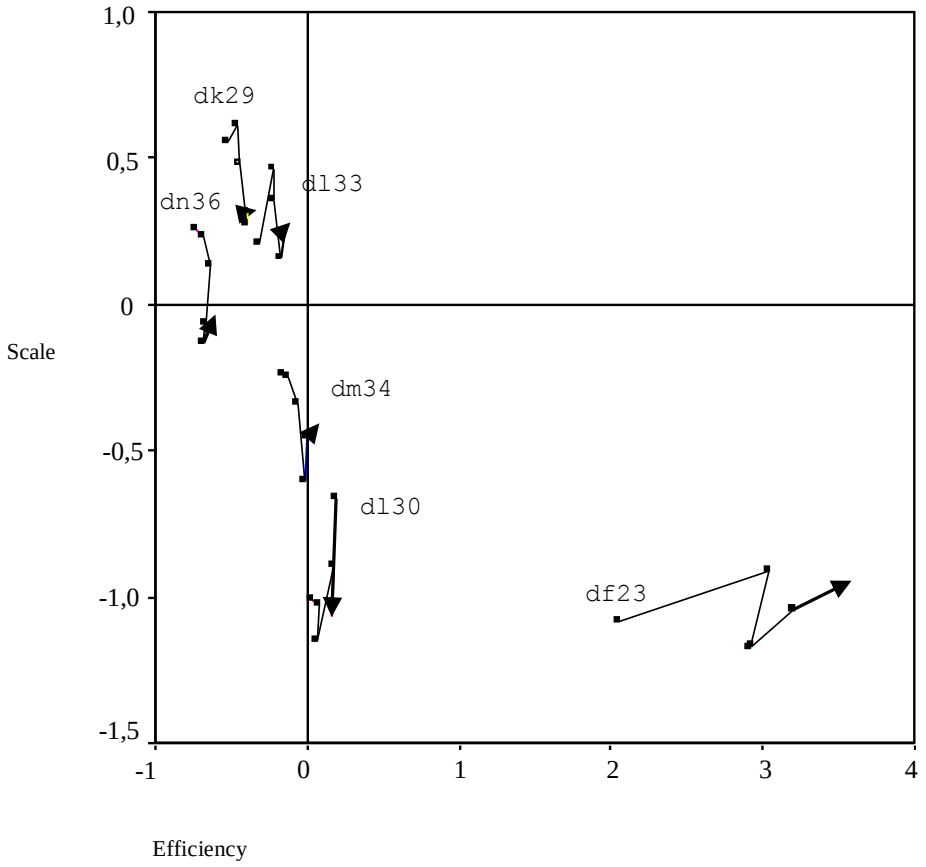


interest is the detachment of division DG24 (Manufacturing of chemicals and chemical products) for both factors.

Fig 12 shows the evolution of the divisions belonging to the "Science based" and "Specialised suppliers" macro-sectors, considered together. These sectors are generally composed of small firms. During the period under examination, downsizing of the partitions became increasingly evident as did the low mobility of the efficiency values. The only exception is the division DF23 (Manufacture of coke, refined petroleum products and nuclear fuel), whose efficiency level shows an upward trend.

5 Concluding remarks

FIG.12. *The dynamics of divisions belonging to "Science-based" and "Specialised suppliers" macro-sectors (1999-2004)*



The paper focuses on the deindustrialisation process, occurred in 13 of the first European member states, during the period 1999-2004. Data of 23 divisions (two-digit NACE classification) of manufacturing have been used. The statistical analysis has been carried out by the 4-*way* Principal Component Analysis (PCA) that, simultaneously, takes into account both the territorial and temporal dimensions of the phenomenon; further, it allows to compare data at a fitting level of disaggregation. According to the latent structure revealed by the analysis, it is easier to understand the real dynamics of deindustrialisation and its territorial and sectorial aspects.

The analysis shows a very heavy process of deindustrialisaton in EU13 countries that is strongly corroborated by the similarities of the manufacturing production processes at a fine disaggregated level.

Indeed, regardless of the industrial structure of each country, or the size of the respective economic systems, nearly all the first member states have experienced a process of deindustrialisation, both in terms of whole manufacturing and its constituent industries. This outcome underlines the high degree of cohesiveness that was typical of the European industrial sector during 1999-2004.

Specifically, the multivariate approach has highlighted the synthesis dimensions (PCs) mainly responsible for the manufacturing evolution in EU13.

In particular, the main changes recorded can be summarized as follows:

1) deindustrialisation can be explained fairly well by using just two parameters (factors): i) the efficiency of productive processes and ii) their dimensions;

2) the efficiency parameters are positive and produce the largest variance between the countries;

3) manufacturing activities are considered downsized when the percentage of the total workforce employed in manufacturing decreases and, secondly, a reduction occurs in the ratio of the Gross operating rate, (i.e. the share of capital remuneration on sales). Therefore, deindustrialisation not only concerns the employment but also the level of capitalization of the industrial system, an aspect that, until now, has not been deeply explored;

4) on a territorial level, European countries (but for a few exceptions) are not very differentiated about manufacturing structures. Moreover, considering their respective dynamics, the initial positions are not significantly far away from European mean;

5) as far as efficiency is concerned, the analysis carried out at a disaggregate level shows only three divisions whose levels are well above the European averages: chemicals, tobacco products and coke, petroleum products and nuclear fuel. It should be noted that last division, in particular, breaks away from the average global situation.

In a specific effort to understand modifications in the behaviour of each division, three clusters derived from the well-known Pavitt taxonomy were studied separately. From the trajectories describing the different trends, it is evident that divisions belonging to the macro-sector "Scale intensive" show, in general, a greater dynamism in the direction of a widespread reduction in the overall dimension of productive processes even if they continue to improve in terms of efficiency.

Within the other two groupings ("Supplier dominated", on the one hand, and "Science based" and "Specialised suppliers", on the other hand), the evolutionary process has caused a certain downsizing for the most divisions during the period, but no significant changes in efficiency levels.

In summary, using *4-way* PCA allows to monitoring different aspects of deindustrialisation, even if a remarkable portion of the variance may be ascribed to other causes which, from the available data, are unidentifiable.

References

Baumol, W., Batey Blackman, S.A., Wolff, E.N., 1989. Productivity and American Leadership: The Long View. Cambridge, Massachusetts: MIT Press.

Blackaby, F. (ed.), 1978.De-industrialisation. London, Heinemann.

Bolasco, S.,1999. Analisi multidimensionale dei dati. Carocci, Roma.

Boulhol, H., Fontagné, L., 2006. Deindustrialisation and the fear of relocations in the industry. CEPII. Working Paper, N° 2006-07, 4-32.

Brady, D., Denniston, R., 2006. Economic Globalization, Industrialization and Deindustrialization in Affluent Democracies. The University of North Carolina Press. Social Forces 85.1, 297-329.

Caves, R.E., Barton, D.R., 1991. Efficiency in U.S. Manufacturing Industries.

The MIT Press. Cambridge, Massachusetts.

Clark, C., 1957. The Conditions of Economic Progress. London: Macmillan.

EUROSTAT, 1999-2004. Selected indicators for all activities.

Ferguson, P.R., 1988. Industrial Economics: Issues and Perspectives. MacMillan Education Ltd, London.

Ferguson, P.R., Ferguson, G.J., 1994. Industrial Economics: Issues and Perspectives. 2nd Edition. MacMillan Education Ltd, London.

Gallino, L., 2003. La scomparsa dell'Italia industriale. Einaudi, Torino.

Gemmell, N., 1986. Structural Change and Economic Development. The Role of Service Sector. London, MacMillan Press.

Gershuny, I.J., 1978. After Industrial Society? The Emerging Self-Service Economy, London, MacMillan Press.

Grodzicki, M., 2014. Structural Similarities of the Economies of the European Union. Eherelibrium, Quarterly Journal of Economics and Economic Policy, vol. 9. Issue 1, 93-117.

Jones, R.M., Lee, N., 1985. Industrial and Market Structure. In: Devine, P.J., Lee, N., Jones, R.M., Tyson, W.J. (Eds.), An Introduction to Industrial Economics. London, G. Allen & Unwin Ltd., 27-70.

Kiers, H.A.L., Van Mechelen, I., 2001.Three-way component analysis: Principles and illustrative application. Psychological Methods 6, 84-110.

Mickievicz, T., Zalewska, A., 2006.De-industrialization. Rowthorn and Wells' Model Revisited. Acta Oeconomica, Vol. 56 (2), 143–166.

Palma, J.G., 2014. Industrialization, 'premature' deindustrialization and the Dutch disease. Revista NECAT, 3(5), 7–23.

Penava, M., Družić M., 2015. Croatian industrial policy in the context of

Deindustrialization. Economic Research-Ekonomska Istraživanja, 28:1, 843-852.

Rizzi, A., Vichi, M., 1995. Three-way data set analysis. In: Rizzi, A., (ed.), Some relations Between Matrices and Structures of Multidimensional Data Analysis. Giardini, Pisa. Rodrik, D., 2015. Premature deindustrializationin the developing world. In: Dani Rodrik's Weblog, Febr., 1-4

Rodrik, D., 2016.Premature deindustrialization. Journal of Economic Growth,vol. 21(1), 1-33.

Rowthorn, R., Coutts, K., 2013. Re-industrialisation – a commentary. Future of Manufacturing Project: Evidence Paper, 32, 1-28.

Rowthorn, R., Ramaswamy, R., 1997. Deindustrialization - Its Causes and Implications. Washington, IMF, Economic Issues, No.10, 1-5.

Rowthorn, R., Ramaswamy, R., 1999. Growth, Trade and Deindustrialization. IMF, Staff Papers, vol. 46, No.1, march, 3-28.

Rowthorn, R.,Wells, J., 1987. De-Industrialization and Foreign Trade. Cambridge: Cambridge University Press.

Singh, A., 1977. UK industry and the world economy: a case of de- industrialisation? Cambridge Journal of Economics, vol. 1, no. 2, 113-136.

Stanners, W., 2001. De-industrialisation II. Development and Comp Systems, 0107001, Econ WPA, 1-6.

Tucker, L.R., 1966. Some mathematical notes on three-mode factor Analysis. Psychometrika, 31 (3), 279–311.

Vandengiste, B.M.G., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., 1998. Handbook of Chemometrics and Qualimetrics: Part B. Elsevier Science, Amsterdam.

Watts, M., Valadkhani, A., 2001. The Impact of Deindustrialisation on Employment Outcomes in Australia, Japan and the USA. CofFEE (Centre of Full Employment and Eherety) Workshop: Understanding Unemployment in Australia, Japan and the USA. A Cross Country Analysis. 10-11 dec., 2-25.

UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 'in fase di assegnazione', volume I, 2017

Leisure, Social Capital and Life Turns in Deviant Youth

Fabio Massimo Lo Verde

Abstract The production of social capital in a specific area of everyday life such as leisure time and the different socio-cultural contexts it is experienced in is a very interesting research issue, especially in the light of certain specific meanings of the notion of social capital, such as Bourdieu's or, more recently, Putnam's. Nonetheless, this research issue is scarcely taken into consideration in Italy.

Albeit inexaustively, this paper intends to introduce this issue starting from a brief review on the generation of social capital in youth's leisure time contexts. In the first paragraph I problematize the notion of social capital as referred to leisure time "contexts" as well as analyze either the social capital literature dealing with the modes and experiences of leisure time, or the leisure time literature focusing on the construction of social capital in leisure time contexts. In the second paragraph I discuss some studies regarding the ways in which a particular age range - youth - produces social capital in leisure time contexts. In the third paragraph I focus on some studies regarding the issue of youth's leisure time as a potential "antisocial" time. In the fourth paragraph I introduce the discussion about the little importance given to public leisure in the service provision for youth's leisure time and the consequences determined by that in terms of social capital "erosion". The conclusion offers three metaphors for understanding the trends of leisure time and sociability

Keywords Social capital · Leisure Researches · Sociology of leisure

Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo
viale delle Scienze ed. 13, 90128
E-mail: fabio.loverde@unipa.it

Riassunto *Per quanto il tema della produzione di capitale sociale in un ambito specifico della vita quotidiana come quello del tempo libero e dei diversi modi in cui si declina nei diversi contesti socioculturali sia fra i più interessanti - oltre che da considerarsi un ambito di ricerca germinale di una specifica accezione di capitale sociale, si pensi a quella di Bourdieu o alle più recenti riflessioni di Putnam – esso risulta essere, soprattutto in Italia, scarsamente trattato. Senza pretendere a un'esaustività impossibile da realizzare in questa sede, il presente lavoro intende introdurre il tema a partire da una breve ricostruzione che ha come oggetto la generazione di social capital in contesti di leisure time giovanile. Nel primo paragrafo si problematizza il concetto di social capital in riferimento ai "contesti" di leisure time e si analizza sia la letteratura sul social capital che ha sviluppato alcune piste di ricerca interessanti a partire dalle modalità e dalle esperienze di consumo di leisure time, sia quella sul leisure time che si concentra sulla sua importanza per la costruzione di social capital. Nel secondo paragrafo si analizzano alcune ricerche riguardanti il modo in cui in una particolare fascia di età, quella giovanile, si costruisce social capital in contesti di leisure time. Nel terzo paragrafo si focalizza l'attenzione su alcuni lavori che hanno affrontato il tema del leisure time giovanile come un tempo "a rischio di antisocialità". Nel quarto paragrafo si introduce il tema inerente la scarsa importanza attribuita al public leisure nell'offerta di servizi per il leisure time giovanile e le conseguenze che ciò determina sulla "erosione" di capitale sociale. Nelle conclusioni vengono presentate tre metafore per comprendere il trend della relazione fra leisure time e sociabilità nella postmodernità.*

Parole chiave *Capitale sociale - Studi sul tempo libero - Sociologia del tempo libero*

1 Introduction

Three brief anecdotes, all coming from Sicilian common culture, clearly demonstrate how Sicilians picture their island. Two of them feature God and his relationship with Sicily, the third one concerns a cure for its problems.

The first tells about God deciding to play a bad-taste practical joke on someone. After having rotated his index finger in the air for several seconds he shouts "You!", pointing his finger and deciding that this person will be born in Sicily .

God is also the protagonist of the second anecdote. After creating Sicily as an Eden-like land, he has to compensate for such an excessive perfection - so he creates Sicilians!

In the third tale an old man finally finds the remedy to Sicilian problems after unquantifiable meditation - the airplane! In fact there are many learned Sicilians who, having decided not to leave their land permanently by plane, remain in Sicily and assume a constantly laconic expression - sad or bad-humoured, annoyed or bitter - rarely happy. That is perhaps because they mix the heavy cultural load which they feel obliged to carry - and with which they collide - with the well known (at least in Italy) tendency to use the "gattopardesque" categorisation of the supremacy of change to explain the static nature of Sicily.

In this brief note, I present a review of a quite important problem in the literature on the regions of Italy, that is the relation between leisure time and the construction of social capital.

I also introduce a description of an on-going research project about the "turning points" (Laub and Samspson 1993) in the biography of 40 Sicilian young adults (17-25 years old) who have decided to stop their deviant carrier as a consequence of a particular event (often an arrest), although in their social context it is more frequent that they continue this career. Can the young and intergenerational social capital built in leisure time be a barricade against deviant careers in Sicily too, like in other places of the world? Can this be a better solution than the airplane?

2 Leisure time and social capital. A short review

Most of the literature on the creation of social capital within leisure time contexts refers to a notion of social capital in terms of "relational goods" (Arai, Pedlar, 1997; Blackshaw, Long, 2005; Hemingway, 1999); such studies focus both on the availability of *leisure time* among people belonging to different social groups and on the relation between the amount of *leisure time*, its allocation and the construction of social capital in the community by means of the activities people do during such time (Rohe, 2004).

In a specific research dealing with the relevance of *leisure* within the literature on social capital, Glover and Hemingway (2005) – while criticizing the insufficient attention *leisure studies* have paid to the concept of social capital – show how the study of the definitions and the practices of *leisure*, combined with the patterns of social capital construction, can contribute to bring to light some important aspects of contemporary sociality.

On the other side, one of the major scholar on *leisure*, Chris Rojek, states that «Concretely, *leisure* is one of the main institutions which make it possible to accumulate social capital (2007, p. 324), since it is during this *time* and in its contexts that reciprocity dynamics can be activated, so supporting the strengthening of social ties and the improvement of community welfare (*ibidem*; see also Arai, Pedlar, 1997).

Such topic has also been widely treated by one of the best known works on social capital and its "decline" in the United States (Putnam, 2004), which shows how sport associations and *leagues* – particularly those in the field of *bowling* – are an important context for supporting social cohesion. Rather than considering *leisure time* as a *secondary* field of social life or as a variable depending on other institutional fields of society, such as work, family, etc., Putnam considers *leisure time* as a field where the basis of the voluntary tendency to form associations and the subsequent social cohesion *can be founded*. In addition, *leisure time* is important not only in terms of recreation time, but also in terms of time dedicated to those forms of social exchange which support the construction of stronger ties and norms, shared values, opinions and statements on the world; moreover, Putnam argues that the disposition toward a "lonely" consumption of free time is one of the causes of the «erosion» of social capital in the industrialized societies.

This argument is supported also by other researchers (Hamermesh, 1999, 2000; Gershuny, 2000; Reyes-Garcia *et al.*, 2008), who show that this *trend* of *lonely* consumption is raising in

the developed societies. On the other hand, even in archaic societies there seems not to be a positive relation between the lonely consumption of *leisure* activities and well-being perceived in terms of satisfaction and good life quality (Reyes-Garcia *et al*., 2008, pp. 4 and foll.). In fact, the research just mentioned proposes an estimation of the relation between well-being and the two ways of *leisure time* consumption, i.e. *lonely* and *social*, and shows how the latter in particular is able to jointly generate *social capital* and *happiness*.

After their analysis of the literature on the relation existing between social capital and *leisure time*, Glover and Hemingway (2005) argue that a notion of social capital as an extension of the individual involvement into formal and organized social structures, as well as informal associations[1] – where matters of high "civic" content are treated – can be more easily found within *leisure studies*. In other words, it is within *leisure studies* that the second meaning of the concept of social capital – i.e. the "civic" meaning – seems to prevail on the meaning of social capital as a structure of the resources, information, support etc. that can be acquired through the individual participation in a social network.

Referring to this second meaning, Hemingway (1999), for instance, suggests that the amount of social capital that can be generated within leisure time is a *function* of the *kind of leisure activities* where the individuals are involved, and of the kind of *leisure* supply existing in different societies. Obviously, this fact influences the different forms of citizenship that can be generated (*ibidem*, 154-61) and the subsequent forms of democracy and participation. Somehow this phenomenon would be related to the level of individual independence observed in the *creation* of a personal *leisure* rather than in a simple "consumption" practice, and this fact would originate a wider possibility to increase individual "capabilities", as Amartya Sen would say, and favour the creation and reproduction of social capital (see Glover, Hemingway, 2005, p. 395). As the two scholars say, «The question is not if leisure is associated with civically relevant social capital, but what kind of leisure in what kind of setting [...]» (*ibidem*). Therefore, Glover and Hemingway conclude that it seems obvious that *leisure* activities can help to increase and/or maintain social capital, but there can also be consumption practices that reduce rather than increase it, including several forms of *passive leisure* carried out especially through individual practices, as I argued elsewhere (Lo Verde, 2009; Dioguardi, Lo Verde, 2009).

In another study based on the data emerging from an English national *survey*[2] – which studied to what extent a greater consumption and variety of leisure activities could generate social capital – Warde e Tampubolon (2002, pp. 163 and foll.) show the existence of a connection between the variety of *leisure* activities and the formal participation in membership groups, in the sense that the number of leisure activities seems to increase as the involvement in associative activities increases, including, of course, those activities directed towards the management of leisure services (sports clubs, recreation clubs etc.)

As the authors argue, according to a view *à la* Putnam, the involvement in fun activities and hobbies seems to increase as *individual* social capital increases, even if the variation in

[1] Among such informal associations the two scholars include the recent network of online discussion groups.
[2] This survey is the *British Household Panel Survey* started in 1991.

the growth of such activities is positively related to education qualifications. People who are members of formal associations spend less time for each activity, since the number of the activities is greater; however, on the other hand, such people show high level of involvement, for instance, in terms of how frequently each activity is done. In addition, the growth of public and civic participation seems to be associated with the growth of the recreational activities practiced in private settings (*ibidem*, p. 166). According to these two authors, this would seem to disconfirm the theory by Hirschman (1983) that consumption and political activism are often conflicting. On the other hand, if we accept the theory that the total volume of participation in associations is strongly correlated with political activism, it would seem that whoever is politically active is simultaneously strongly involved in various forms of recreational activities. In the same study, Ward and Tampubolon analyze the relationship between friendship networks and the types of recreational activities practiced.

According to this research people who are members of caregiver associations tend to have a tendency to share different friends from themselves in terms of leisure choice, lifestyles and so on. As Ward and Tampubolon argue, «This might imply that people take some notice of the classic agony aunt recommendation that if you are lonely or sad, and haven't got many friends, then join a club». (*Ibidem*, p.169). However, people who generally have very different friends would also tend to have fewer friends. But this fact would reduce the participation in formal groups, far from the reverse situation. Relating to *leisure*, one is likely to decide to take part in formal groups *if* he/she cannot share activities with his/her friends. But it could also be plausible the reverse argument that *the associations include very different people*, in terms of age, gender, educational qualifications etc., *and this fact would consequently facilitate their mixing with other people who differ by nature*.

According to the researchers there is no connection between all these things and the presence of social capital, since the embedding into friendship-based networks *would seem to have weak effects on the participation in the activities within associations* and, consequently, on a particular dimension of social capital, the civic one. In fact, the distance among friends is more sensitive to the number of activities that one can do than it is to their frequency. In other words, if one has friends who are so much like him/her, he/she will probably prefer to do more types of activities – perhaps because he/she shares them with such friends – than doing the same activity more frequently.

3 Social capital and leisure time among young people

Our research question is how social capital among young people can be generated and, moreover, how it can be developed or eroded through their *leisure* activities.

Scholars point out that adults and policy makers usually consider *leisure time* among young people as an "empty" time, and consequently – however without ever specifying this connection – as *risky* both for themselves and the adults they live with in different communities. Equally often scholars argue that leisure time is important in the organization and life of the community.

On the one hand, there are other studies focusing on youth participation in activities that can generate social capital, civic-mindedness etc. (McFarland, Thomas, 2006; Helve, Bynner, Holland, 2009)[3] – such as activities like participation in voluntary associations. Also international organizations recognize the importance of leisure time in the development of communities (United Nations, 2004, pp. 214ss.); on the other hand, however, little research has been carried out so far about the connection between the activities done by young people in their leisure time and the construction of trust networks that are fundamental for generating *social capital*. Mainstream studies predominantly deal with the *lack* of social capital and the relation between such lack and the diffusion of deviant behaviours (Mahoney, Stattin, 2000; Deuchar, 2009) or criminal ones (Hagan, McCarthy, 1997).

As some studies show, the analysis of how social capital can be generated among young people is important since those who participate in youth groups – and thus tend to generate a specific form of social capital – are more likely to maintain this practice in adulthood too. In addition, high levels of social trust among adults are positively correlated to high levels of social trust during their teen years (Stolle, Hooghe 2004, p. 431). Briefly, according to Vesely (2006) the study of social capital among young people represents a "window on the future" of societies, in spite of the problem of the low "levels of stability" in the transition from one generation to another.

Finally there is a great amount of research on a particular kind of youth free/leisure time that can be defined as "antisocial". Such research includes, for instance, the studies dealing with the different kinds of gambling played by teenagers and young adults, or the "normalization" of drugs consumption (Parker, 1998; Griffits, 1995). In general, these studies consider *leisure time* as a "dangerous time", partially or entirely beyond the adults' control, with a high level of risk that young people may enter/stay in deviant groups, so generating *deviant* or *illegal leisure* (Rojek, 1999) rather than forms of pro-sociality *coping*.

One of these studies (Mahoney, Stattin, 2000) shows that the involvement of teenagers and young adults in "more structured" activities[4] as well as the presence of an adult as a leader (*coach*, *trainer*, etc.) were negatively correlated to antisocial behaviours; whereas "less structured" activities – such as joining youth community centres or watching television – were positively correlated to antisocial activities. In addition, adults who had taken part in less structured activities during their youth used to choose deviant individuals as their friends, have difficult emotional ties with their parents and get a lower support to their activities from operators (*ibidem*, p. 114ss.).

[3] About Italy see Bettin Lattes, 2001; Prandini, Melli, 2004, Cesareo, 2003, 2005; Donati, Colozzi, 1997; and IARD Reports (Buzzi, Cavalli e de Lillo, 1997, 2002, 2007).

[4] The expression «more structured activities» refers to those activities that need to be learned through socialization processes and that allow individuals to gain specific abilities, namely *serious leisure* activities (Stebbins, 2007; see also Caldwell, Smith, 2006, p. 402), such as the activities in the field of sport, music, art, handicraft etc.

Focusing on a "criminological" view, Caldwell and Smith (2006, p. 399) have recently classified the studies about the relation between *leisure* and youth deviant careers into four lines of research:

a) A line considers free/leisure time as *time "to be filled"*. Therefore the time that "is filled" with "pro-social" activities cannot be filled with deviant ones.

b) Another line of research looks at free/leisure time as *connected with activities done with a deviant peer group*: some activities have higher chances of generating deviant behaviours or are typical of deviant subcultures; therefore it is important to identify the genesis of such activities, to analyze the context where they were generated etc.

c) A further line analyzes *leisure time* in connection with a higher or lower level of "structured" activities: the time dedicated to those activities that are less organized, informal or with no supervision by adults is more likely to provide the context for deviant behaviours, unlike the time spent in activities that are supervised by adults or are more structured. There are opposite views especially in this case. For instance, many scholars agree that unstructured activities let people, especially teenagers, experience new roles, ideas and behaviours, hence supporting their social identity formation (Kleiber, 1999) as well as the development of an autonomous transition process to adulthood.

d) Finally, a line of research looks at the interaction between the individual and the environment as a paradigmatic element: to have self-control and share norms and conventional activities prevent from adopting deviant behaviours.

Anyhow, in spite of the unsystematic, often fragmented and sometimes contradictory nature of the studies on the contribution that leisure activities and contexts can offer to the construction of social capital and to networks of shelter pursuing precautionary or rehabilitating aims (Williams, Walker, 2006; Williams, 2009), the topics on such contribution are partially treated also with respect to the biographies of young people and teenagers who are at risk of social exclusion or who already live in marginal contexts or environments; therefore these topics are usually covered by a literature looking at the importance of institutional "networks", and no-profit organizations that benefit from the state intervention towards some social classes[5] or that live within re-educational institutions for teenagers outlaws.

Beyond the problems existing in data survey and data capture procedures on the different leisure activities (see Gershuny, 2000), the fact is that a different amount of time is spent on different activities that have a higher sociability nature and this situation can be considered as a consequence of the way free/leisure time is institutionalized within different cultures. Nowadays such way is partially function of the supply, especially the supply of places and contexts for *leisure time* provided by the public sector, which is less and less capable of "competing with"

[5] With reference to such interventions made in Italy, see Sanicola, Piscitelli, Mastropasqua (eds.), 2002.

a private sector that tends to produce *leisure* services and goods to be experienced by single individuals in private places (Lo Verde, 2009).

As a temporary conclusion, it can be argued that leisure time can possibly be, under specific conditions, an important *locus* for social capital generation within peer groups and young adults. Togetherness and peer complicity, although more in a *bonding* rather than a *bridging* way, can be a supportive and relational resource for young individuals. Is it possible "to trigger" life turns decision in deviant young adults who spend their leisure time in occasional/frequent illegal practices? And how can this leisure time be turned from illegality to the construction of social integration?

In our still on-going research we have focussed on two concepts, discontinuity and reflexivity, following Archer's typology of reflexivity (Archer, 2003). We want to know how the social relation system may contribute to a decision of discontinuity and what role reflexivity plays in this decision. We have interviewed 40 young adults aged between 17 and 25, asking them to narrate their decision of life change to see whether this decision was somehow related to some specific form of reflexivity. Moreover, we have compared these interviews with the opinions of social workers and teachers as expressed during three focus groups. We are currently working on the analysis of the data collected through such interviews and focus groups. From a first superficial analysis, there seems to be evidence that leisure time is a problematic dimension for these young adults (how and where to spend it, with whom, for how long, and so on). There also seems to be evidence that structured activities, daily engagements, the presence of an adult "coach" or "leader" or "guide", the shifts or "fractures" with the usual daily social context, may contribute to the reduction of the time spent in illegal leisure practices. Interestingly, this kind of evidence seems to bring Sicilian young adults much nearer to other global cities' young adults in the world. There is indeed room for new general hypotheses and research about leisure time, social capital and deviant young adults.

References

Arai S., Pedlar A. (1997), *Building communities through leisure: Citizen participation in a healthy communities initiative*, in «Journal of Leisure Research», vol. 29, n. 2, pp. 167-182.

Archer M. S. (2003), *Structure, Agency and The Internal Conversation*, Cambridge University Press, Cambridge.

Bettin Lattes G. (a cura di) (2001), *La politica acerba. Saggi sull'identità civica dei giovani*, Rubbettino, Soveria Mannelli (CZ).

Blackshaw T., Long, J. (2005), *What's the big idea? A critical exploration of the concept of social capital and its incorporation into leisure policy discourse*, in «Leisure studies», vol. 24, n. 3, pp. 239-258.

Buzzi C., Cavalli A. De Lillo A., (a cura di) (2002), *Giovani del nuovo secolo. Quinto rapporto* IARD *sulla condizione giovanile in Italia*, Bologna, Il Mulino.

Buzzi C., Cavalli A. De Lillo A., (a cura di) (2007), *Rapporto giovani. Sesta indagine dell'Istituto IARD sulla condizione giovanile in Italia*, Il Mulino, Bologna.

Buzzi, C., Cavalli, A. De Lillo, A, (a cura di) (1997), *Giovani verso il Duemila. Quarto rapporto* Iard *sulla condizione giovanile in Italia*, Bologna, Il Mulino.

Caldwell, L. L., Smith E. A. (2006), *Leisure as a context for youth development and delinquency prevention*, in «Australian and New Zealand Journal of Criminology», Vol. 39, n. 3, pp. 398-418.

Cesareo V. (a cura di), *I protagonisti della società civile in Italia*, Rubbettino, Soveria Mannelli (Cs) 2003.

Cesareo V. (a cura di), *Ricomporre la vita. Gli adulti giovani in Italia*, Carocci, Roma 2005.

Coleman J. S. (1990), *Foundation of Social Theory*, Harvard University Press, Cambridge MA

Deuchar R. (2009), *Gangs, Marginalised Youth and Social Capital*, Trentham Books, London

Dioguardi V., Lo Verde F. M. (2009), *Non solo "quel che resta del giorno". Un'analisi comparativa del consumo di tempo libero in Europa*, in «Studi di Sociologia», 4, pp. 345-381.

Donati P., Colozzi I. (a cura di) (1997), *Giovani e generazioni. Quando si cresce in una società eticamente neutra*, il Mulino, Bologna.

Gershuny J. (2000), *Changing Time. Work and leisure in post-industrial society,* Oxford University Press, Oxford.

Glover T. D., Hemingway, J. (2005), *Locating leisure in the social capital literature*, in «Journal of Leisure Research», vol. 37, n. 4, pp. 387-401.

Griffiths M. (1995), *Adolescent gambling*, Routledge, London.

Hagan J., McCarthy B. (1997), *Mean Streets: Youth Crime and Homelessness*, Cambridge University Press, New York.

Hamermesh, D.S. (1999), *The timing of work over time*, in «The Economic Journal» n, 109, pp. 37-66.

Helve H., Bynner J. (Eds.) (2007), *Youth and Social Capital*, Tufnell Press, London.

Hemingway J. (1999). *Leisure, social capital, and democratic citizenship*, in «Journal of Leisure Research»,vol. 31, n.2, pp. 150-165.

Hirschman A. O. (1983), *Felicità privata, felicità pubblica*, Il Mulino, Bologna.

Holland J. (2009), *Young people and social capital. Use or abuse,* in «Young», vol. 17, n. 4, pp. 331-350.

Kleiber D.A. (1999), *Leisure experience and human development: A dialectical interpretation*, Basic Books, New York.

Laub J. H., Sampson R. J. (2003), *Turning Points In The Life Course: Why Change Matters To The*

Study Of Crime, in «Criminology», Vol. 31, n.. 3, pp. 301-325.

Lo Verde F. M. (2009), *Sociologia del tempo libero*, Laterza, Roma-Bari.

Mahoney J. L, Stattin H. (2000), *Leisure activities and adolescent antisocial behaviour: The role of structure and social context*, in «Journal of Adolescence» n.23, pp. 113-127.

McFarland D. A., Thomas, R. J. (2006), *Bowling young: How youth voluntary associations influence adult political participation*, in «American Sociological Review», vol. 71, n. 3, pp. 401-425.

Parker H. (1998), *Illegal leisure: the normalization of adolescent recreational drug use*, Routledge, London.

Prandini R., Melli S. (2004), *I giovani e il capitale sociale dell'Europa futura*, Franco Angeli, Milano.

Putnam R. D. (2004), *Capitale sociale e individualismo. Crisi e rinascita della cultura civica in America*, Il Mulino, Bologna.

Reyes-García V., Godoy R., Vadez V., Ruíz-Mallén I., Huanca T., Leonard W., McDade T., Tanner S. (TAPS study Team), (2009), *The pay-offs to sociality: Do solitary and social leisure relate to happiness?* in «Human Nature. An Interdisciplinary Journal with Biosocial Perspective», Vol. 20, n. 4, pp. 431-446.

Roberts K. (1983), *Youth and Leisure*, Allen & Unwin, London.

Rohe W. M. (2004), *Building social capital through community development*, in «Journal of the American Planning Association», n. 70, pp. 143-144.

Rojek C. (1999), *Deviant leisure: The dark side of free-time activity*, in Jackson E.L., Burton T.L. (Eds.), *Leisure studies: Prospects for the twenty-first* century, Venture, State College PA, pp. 81–96.

Rojek C. (2007), *An Outline of the Action Approach to Leisure Studies*, in Page S. J., Connell J. (eds.) (2007), *Leisure Studies*, Routledge, London, 4 voll., vol. I, pp. 321-335.

Sanicola L., Piscitelli D., Mastropasqua I. (a cura di) (2002), *Metodologia di rete nella giustizia minorile*, Liguori, Napoli.

Stebbins R.A. (2007), *Serious Leisure*, Transaction Publisher, New Brunswick, New Jersey.

Stolle D., Hooghe M. (2004), *The Roots of Social Capital: Attitudinal and Network Mechanisms in the Relation between Youth and Adult Indicators of Social Capital*, in «Acta Politica», vol. 39, n. 4, pp. 422-441.

Veselý R. (2006), *Reproduction of Social Capital: How Much and What Type of Social Capital Is Transmitted from Parents to Children?* Paper prepared for Trust, Reciprocity and Social Capital The 2006 Ratio Colloquium for Young Social Scientists, Stockholm, August 25-26, 2006, pubblicato in parte in Ratio Working Papers No 105, from The Ratio Institute in http://econpapers.repec.org/paper/hhsratioi/0105.htm.

Warde A., Tampubolon G. (2002), *Social capital, networks and leisure consumption*, in «The Sociological Review», vol. 50, n.2, pp. 155-180

Williams, D J (2009), *Deviant leisure: Rethinking "the good, the bad, and the ugly"*, in «Leisure Sciences», n. 31, 207-213.

Williams, D J, Walker, G. J. (2006), *Leisure, deviant leisure, and crime: Caution: Objects may be closer than they appear*, in «Leisure/Loisir», n. 30, pp. 193-218.

UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 'in fase di assegnazione', volume I, 2017

A recap on Linear Mixed Models and their hat-matrices

Gianfranco Lovison · Mariangela Sciandra

Abstract This working paper has a twofold goal. On one hand, it provides a recap of Linear Mixed Models (LMMs): far from trying to be exhaustive, this first part of the working paper focusses on the derivation of theoretical results on estimation of LMMs that are scattered in the literature or whose mathematical derivation is sometimes missing or too quickly sketched. On the other hand, it discusses various definitions that are available in the literature for the hat-matrix of Linear Mixed Models, showing their limitations and proving their equivalence.

Keywords Linear Mixed Models · Inference · Hat matrices · Orthogonal Projectors

Riassunto *uesto working paper ha un doppio obiettivo. Da un lato, fornisce un riepilogo sui Modelli Lineari Misti (MLM): lungi dal tentare di essere esaustiva, questa prima parte del working paper si focalizza sulla derivazione di risultati teorici sulla stima dei MLM che sono sparsi in letteratura o la cui giustificazione matematica è a volte mancante o abbozzata troppo frettolosamente. Questa prima parte si articola come segue: dopo aver specificato il Modello Lineare Misto, anche nella sua utile forma "compatta" (stacked), ed aver introdotto i necessari risultati distributivi, vengono derivati formalmente gli stimatori degli effetti fissi e i predittori degli effetti casuali, sia utilizzando l'approccio marginale che quello congiunto. Dapprima, queste derivazioni vengono ottenute sotto l'assunto (poco realistico) che le matrici di varianze/covarianze sia degli*

G. Lovison: Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo,
viale delle Scienze ed. 13, 90128
E-mail: gianfranco.lovison@unipa.it
· M. Sciandra Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo,
viale delle Scienze ed. 13, 90128
E-mail: mariangela.sciandra@unipa.it

errori sia degli effetti casuali siano note; successivamente tale assunto non viene specificato, e si presentano i risultati riguardanti la stima di tali matrici di variane/covarianze, sia con il metodo ML che con quello REML. Infine, viene discussa la rappresentazione dei MLM mediante "dati aumentati", dovuta a Hodges (1998), utile per gli sviluppi successivi.

Nella seconda parte, il working paper discute varie definizioni della matrice di proiezione (hat-matrix) che sono disponibili nella letteratura sui Modelli Lineari Misti, evidenziando le loro limitazioni e dimostrandone formalmente, per la prima volta, l'equivalenza. Viene inoltre evidenziato come l'unica matrice di proiezione ortogonale, e dunque simmetrica ed idempotente, sia quella ottenibile dalla rappresentazione di Hodges, un risultato utile per ulteriori sviluppi di ricerca.

Parole chiave *Modelli Lineari Misti - Inferenza - Matrice di proiezione - Proiettori ortogonali.*

1 The Linear Mixed Models: specification

Let the data have the following structure:

$$y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij} \quad i = 1, \ldots, m; \quad j = 1, \ldots, n_i; n = \sum_{i=1}^{m} n_i \tag{1}$$

where: i is the cluster index
j is the individual (within cluster) unit index
y_{ij} is the response variable
$\boldsymbol{x}_{ij}$ is a vector of p explanatory variables (with fixed parameters)
$\boldsymbol{z}_{ij}$ is a vector of q explanatory variables (with random parameters)

For the ease of presentation, we assume in this paper $n_i = k \ \ \forall i$, so that $n = km$.
We can arrange the data according to the clustered structure as:

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{i,k} \end{bmatrix}, \boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_{i1}^T \\ \vdots \\ \boldsymbol{x}_{ij}^T \\ \vdots \\ \boldsymbol{x}_{i,k}^T \end{bmatrix}, \boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{z}_{i1}^T \\ \vdots \\ \boldsymbol{z}_{ij}^T \\ \vdots \\ \boldsymbol{z}_{i,k}^T \end{bmatrix} \qquad i = 1, \ldots, m \tag{2}$$

Suppose the data are generated by a Gaussian Linear Mixed Model (Breslow and Clayton, 1993), specified at cluster level as follows:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, m \tag{3}$$

where: $\boldsymbol{X}_i$ is a $k \times p$ matrix, $\boldsymbol{\beta}$ is a p-vector of unknown fixed parameters, $\boldsymbol{Z}_i$ is a $k \times q$ matrix and $\boldsymbol{b}_i$ is a q-vector of random parameters.

As far as the random components of the model are concerned, we assume that both the random parameters $\boldsymbol{b}_i$ and the within-cluster errors $\boldsymbol{\epsilon}_i$ are Normally distributed, with variance/covariance matrices which are full-rank unconstrained positive-definite matrices; besides, we assume that the the within-cluster errors (conditional) variance/covariance matrices are homogeneous across clusters:

$$\boldsymbol{b}_i \sim \mathcal{MVN}_q(\boldsymbol{0}_q, \boldsymbol{\Sigma}_{B_c}),\ \boldsymbol{\epsilon}_i \sim \mathcal{MVN}_k\left(\boldsymbol{0}_k, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c}\right),\ \boldsymbol{b}_i \perp\!\!\!\perp \boldsymbol{\epsilon}_i \quad i = 1, \ldots, m$$

where:

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c} = \mathcal{D}[\boldsymbol{\epsilon}_i])\ \forall i; \qquad \boldsymbol{\Sigma}_{B_c} = \mathcal{D}[\boldsymbol{b}_i])$$

are positive-definite matrices (the index "c" stands here for "at cluster level)". Notice that these assumptions imply that the conditional distribution of the response $\boldsymbol{y}_i$, given the random parameters $\boldsymbol{b}_i$, is:

$$\boldsymbol{y}_i|\boldsymbol{b}_i =\sim \mathcal{MVN}_k\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c}\right) \quad i = 1, \ldots, m$$

with:

$$\mathcal{D}[\boldsymbol{y}_i|\boldsymbol{b}_i]) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c}\ \forall i$$

For the subsequent developments, it is convenient to write model (3) in vectorised form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon} \tag{4}$$

where:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_i \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_i \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} \quad \boldsymbol{Z} = \bigoplus_{i=1}^{m} \boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{Z}_1 & \ldots & \boldsymbol{O} & \ldots & \boldsymbol{O} \\ \vdots & & \vdots & & \vdots \\ \boldsymbol{O} & \ldots & \boldsymbol{Z}_i & \ldots & \boldsymbol{O} \\ \vdots & & \vdots & & \vdots \\ \boldsymbol{O} & \ldots & \boldsymbol{O} & \ldots & \boldsymbol{Z}_m \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_i \\ \vdots \\ \boldsymbol{b}_m \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_i \\ \vdots \\ \boldsymbol{\epsilon}_m \end{bmatrix},$$

This representation is called *the stacked form* of the data and the model; here $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are n-dimensional vector, $\boldsymbol{b}$ is r-dimensional, where $r = m \times q$ is the total number of realisations of the random vector $\boldsymbol{b}_i$, $\boldsymbol{X}$ is $n \times p$, $\boldsymbol{Z}$ is $n \times r$.

The dispersion matrices for $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ (which have dimension $r \times r$ and $n \times n$ respectively), can be written in compact form as:

$$\mathcal{D}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_c}; \qquad \mathcal{D}[\boldsymbol{b}] = \boldsymbol{\Sigma}_B = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}_{B_c}$$

whence, the distributional assumptions on $\boldsymbol{\epsilon}$, $\boldsymbol{b}$ and $\boldsymbol{y}|\boldsymbol{b}$ become:

$$\boldsymbol{\epsilon} \sim \mathcal{MVN}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right), \ \boldsymbol{b} \perp\!\!\!\perp \boldsymbol{\epsilon} \tag{5}$$

$$\boldsymbol{b} \sim \mathcal{MVN}_r\left(\mathbf{0}_r, \boldsymbol{\Sigma}_B\right) \tag{6}$$

$$\boldsymbol{y}|\boldsymbol{b} \sim \mathcal{MVN}_n\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right) \tag{7}$$

2 Relevant distributions

In what follows, we shall need the following distributions:

– the joint distribution $f(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \boldsymbol{\Sigma}_B)$

The best way to derive the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{b}$ is directly from the specification of the model, which gives us the conditional distribution $f(\boldsymbol{y}|\boldsymbol{b})$ in (7) and the marginal distribution $f(\boldsymbol{b})$ in (6):

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \boldsymbol{\Sigma}_B) &= f(\boldsymbol{y}|\boldsymbol{b}; \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) f(\boldsymbol{b}; \boldsymbol{\Sigma}_B) \\ &= |2\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})\} \\ &\quad \times |2\pi\boldsymbol{\Sigma}_B|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{b})^T \boldsymbol{\Sigma}_B^{-1}(\boldsymbol{b})\} \qquad (8) \\ &= (2\pi)^{-\frac{n+r}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}|^{-1/2} |\boldsymbol{\Sigma}_B|^{-1/2} \times \\ &\quad \exp\{-\frac{1}{2}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{b}^T \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z}\boldsymbol{b} \\ &\quad + \boldsymbol{b}^T \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{b}^T \boldsymbol{\Sigma}_B^{-1} \boldsymbol{b}]\} \qquad (9) \end{aligned}$$

Focussing first on the exponential argument in (9), we see that is can be written as a quadratic form:

$$\begin{bmatrix} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T & \boldsymbol{b}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & -\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} \\ -\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & \boldsymbol{\Sigma}_B^{-1} + \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}^T \end{bmatrix} \begin{bmatrix} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \boldsymbol{b} \end{bmatrix}$$

Using standard results concerning the inverse of partitioned matrices, one gets:

$$\begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & -\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} \\ -\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & \boldsymbol{\Sigma}_B^{-1} + \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T & \boldsymbol{Z}\boldsymbol{\Sigma}_B \\ \boldsymbol{\Sigma}_B\boldsymbol{Z}^T & \boldsymbol{\Sigma}_B \end{bmatrix}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{b}}^{-1} \tag{10}$$

As for the product of determinants in (9), using a known result concerning the determinant of partitioned matrices:

$$|\boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{b}}| = \left| \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T & \boldsymbol{Z}\boldsymbol{\Sigma}_B \\ \boldsymbol{\Sigma}_B\boldsymbol{Z}^T & \boldsymbol{\Sigma}_B \end{bmatrix} \right| = |\boldsymbol{\Sigma}_B||\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T - \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T| = |\boldsymbol{\Sigma}_B||\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}| \tag{11}$$

Substituting (10) and (11) into (9), we find:

$$f(\boldsymbol{y},\boldsymbol{b};\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\Sigma_\epsilon},\boldsymbol{\Sigma}_B) = (2\pi)^{-\frac{n+r}{2}}|\boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{b}}|^{-1/2}\exp\left\{-\frac{1}{2}\left[(\boldsymbol{y}-\boldsymbol{X\beta})^T \;\; \boldsymbol{b}^T\right]\boldsymbol{\Sigma}^{-1}_{\boldsymbol{y},\boldsymbol{b}}\begin{bmatrix}(\boldsymbol{y}-\boldsymbol{X\beta})\\ \boldsymbol{b}\end{bmatrix}\right\} \tag{12}$$

i.e.:

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{b}\end{bmatrix} \sim \mathcal{MVN}_{n+r}\left(\begin{bmatrix}\boldsymbol{X\beta}\\ \boldsymbol{0}_r\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma_\epsilon}+\boldsymbol{Z\Sigma}_B\boldsymbol{Z}^T & \boldsymbol{Z\Sigma}_B\\ \boldsymbol{\Sigma}_B\boldsymbol{Z}^T & \boldsymbol{\Sigma}_B\end{bmatrix}\right) \tag{13}$$

– the marginal distribution $f(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\Sigma_\epsilon},\boldsymbol{\Sigma}_B)$

Using standard results on the Multivariate Normal distribution, we readily obtain from (13) that the marginal distribution of $\boldsymbol{y}$ is $\boldsymbol{y} \sim \mathcal{MVN}_n\left(\boldsymbol{X\beta},\Sigma_{\boldsymbol{y}}\right)$, with $\boldsymbol{\Sigma_y} = \mathcal{D}[\boldsymbol{y}] = \boldsymbol{Z\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon}$, i.e.:

$$f(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\Sigma_\epsilon},\boldsymbol{\Sigma}_B) = |2\pi\boldsymbol{\Sigma_y}|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{y}}(\boldsymbol{y}-\boldsymbol{X\beta}) \tag{14}$$

– the conditional distribution $f(\boldsymbol{b}|\boldsymbol{y})$

Again using standard results on the Multivariate Normal distribution (Stein, 1981), and in particular on the conditional distribution of a Normal sub-vector given another Normal sub-vector, we obtain from (13) that the conditional distribution of $\boldsymbol{b}|\boldsymbol{y}$ is

$$\boldsymbol{b}|\boldsymbol{y} \sim \mathcal{MVN}_r\left(\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z\Sigma}_B\boldsymbol{Z}^T+\boldsymbol{\Sigma_\epsilon})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}),\boldsymbol{\Sigma}_B-\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z\Sigma}_B\boldsymbol{Z}^T+\boldsymbol{\Sigma_\epsilon})^{-1}\boldsymbol{Z\Sigma}_B\right) \tag{15}$$

or, applying to the variance/covariance matrix formula (69) in Appendix 1 for the inverse of a Schur complement (Zhang, 2006):

$$\boldsymbol{b}|\boldsymbol{y} \sim \mathcal{MVN}_r\left(\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z\Sigma}_B\boldsymbol{Z}^T+\boldsymbol{\Sigma_\epsilon})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}),(\boldsymbol{\Sigma}_B^{-1}+\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z})^{-1}\right). \tag{16}$$

For the sake of notational brevity, we shall sometimes use the symbol

$$\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}} = (\boldsymbol{\Sigma}_B^{-1}+\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z})^{-1}.$$

3 Marginal ML estimation of $\boldsymbol{\beta}$ and Empirical Bayesian prediction of $\boldsymbol{b}$

As shown in (14), the marginal distribution of $\boldsymbol{y}$ is $\mathcal{N}(\boldsymbol{X\beta},\boldsymbol{\Sigma_y})$, with $\boldsymbol{\Sigma_y} = \mathcal{D}[\boldsymbol{y}] = \boldsymbol{Z\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon}$. Hence, the marginal likelihood is:

$$L(\boldsymbol{\beta},\boldsymbol{\Sigma_\epsilon},\boldsymbol{\Sigma}_B;\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z}) = = (2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma_y}|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{y}}(\boldsymbol{y}-\boldsymbol{X\beta})$$

and the marginal log-likelihood can therefore be written:

$$\ell(\boldsymbol{\beta},\boldsymbol{\Sigma_\epsilon},\boldsymbol{\Sigma}_B;\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z}) = -\left(\frac{n}{2}\right)\log(2\pi) - \left(\frac{1}{2}\right)\log(|\boldsymbol{\Sigma_y}|) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{y}}(\boldsymbol{y}-\boldsymbol{X\beta}) \tag{17}$$

3.1 With known $\boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B$

When $\boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B$ are known, $\boldsymbol{\Sigma_y}$ is of course also known, and estimation of $\boldsymbol{\beta}$ reduces to the solution of a weighted least squares system of equations:

$$\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{y} \tag{18}$$

or:

$$\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1}\boldsymbol{y} \tag{19}$$

which yields the (marginal) Maximum Likelihood estimator:

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1}\boldsymbol{y} \tag{20}$$

Clearly, the use of the marginal likelihood rules out the possibility of direct prediction of the random parameters $\boldsymbol{b}_i,\ i = 1, \ldots, m$, since such marginal likelihood is obtained exactly integrating over these random parameters.

The most popular approach for predicting $\boldsymbol{b}$ is therefore an *empirical Bayesian* one. The necessary ingredient for such an approach is the *posterior distribution* $f(\boldsymbol{b}|\boldsymbol{y})$, i.e. the conditional distribution of the random effects realisations given the observations $\boldsymbol{y}$. This was shown in equation (16) to be Multivariate Normal, with expected value $E[\boldsymbol{b}|\boldsymbol{y}] = \boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$. Therefore, (point) prediction of $\boldsymbol{b}$ can be carried out by estimating the posterior mode (or mean, which coincides with the mode owing to Multivariate Normality):

$$\tilde{\boldsymbol{b}} = \widehat{E[\boldsymbol{b}|\boldsymbol{y}]} = \boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{21}$$

Owing to the approach used to derive it, the predictor (21) is often called the *Empirical Bayesian predictor* and denoted in the literature with the acronym EBP (Ando, 2007).

4 Joint ML estimation

From (8) the joint likelihood of $\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B$ is:

$$L(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = L(\boldsymbol{\beta}, \boldsymbol{\Sigma_\epsilon}|\boldsymbol{b}; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})L(\boldsymbol{b}, \boldsymbol{\Sigma}_B) = |2\pi\boldsymbol{\Sigma_\epsilon}|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})$$
$$\times|2\pi\boldsymbol{\Sigma}_B|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{b})^T\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{b})$$

and the joint log-likelihood can be written:

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) &= \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma_\epsilon}|\boldsymbol{b}; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) + \ell(\boldsymbol{b}, \boldsymbol{\Sigma}_B)\\ &= -\left(\frac{n}{2}\right)\log(2\pi) - \left(\frac{1}{2}\right)\log(|\boldsymbol{\Sigma_\epsilon}|) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})\\ &\quad -\left(\frac{r}{2}\right)\log(2\pi) - \left(\frac{1}{2}\right)\log(|\boldsymbol{\Sigma}_B|) - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_B^{-1}\boldsymbol{b}\end{aligned}$$

4.1 With known $\boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B$

If $\boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B$ are known, it is sufficient to differentiate with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$ to find the score vectors:

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B)}{\partial \boldsymbol{\beta}^T} = \boldsymbol{X}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{Z}\boldsymbol{b}$$

$$\boldsymbol{u}(\boldsymbol{b}) = \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Sigma_\epsilon}, \boldsymbol{\Sigma}_B)}{\partial \boldsymbol{b}^T} = \boldsymbol{Z}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{y} - \boldsymbol{Z}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{\Sigma}_B^{-1} \boldsymbol{b}$$

Setting $\boldsymbol{u}(\boldsymbol{\beta})$ and $\boldsymbol{u}(\boldsymbol{b})$ equal to zero yields the system of Maximum (joint) Likelihood equations:

$$\begin{bmatrix} \boldsymbol{X}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{X} & \boldsymbol{X}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{X} & \boldsymbol{Z}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{y} \\ \boldsymbol{Z}^T \boldsymbol{\Sigma_\epsilon}^{-1} \boldsymbol{y} \end{bmatrix} \tag{22}$$

which are often referred to as the Henderson's *mixed model equations* in the LMM literature (Henderson et al., 1959). Solving (22) yields the ML estimators of $\boldsymbol{\beta}$ and $\boldsymbol{b}$ (with known $\boldsymbol{\Sigma_\epsilon}$ and $\boldsymbol{\Sigma}_B$):

$$\hat{\boldsymbol{\beta}}_J = \{\boldsymbol{X}^T[\boldsymbol{\Sigma_\epsilon}^{-1} - \boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}]\boldsymbol{X}\}^{-1}\boldsymbol{X}^T[\boldsymbol{\Sigma_\epsilon}^{-1} - \boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}]\boldsymbol{y} \tag{23}$$

$$\tilde{\boldsymbol{b}}_J = (\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{24}$$

Searle (1992) and Robinson (1991) showed that the predictor (24) obtained from the joint MLE method is the *Best Linear Unbiased Predictor* of $\boldsymbol{b}$; for this reason, (24) is often denoted by the acronym "BLUP" in the Linear Mixed Models literature.

4.2 Marginal MLE and EBP coincide with joint MLE and BLUP in Linear Mixed Models

A very important result, due to Searle et al. (1971), states that the MLE of $\boldsymbol{\beta}$ and the EBP predictor of $\boldsymbol{b}$ obtained in the marginal approach coincide in Linear Mixed Models with the MLE and the BLUP obtained in the joint approach.

The identity between the marginal ML estimator $\hat{\boldsymbol{\beta}}_M$ and the joint ML estimator $\hat{\boldsymbol{\beta}}_J$ can be easily proved by applying formula (69) in Appendix 1:

$$\boldsymbol{\Sigma_\epsilon}^{-1} - \boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1} = (\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1} = \boldsymbol{\Sigma_y} \tag{25}$$

whence

$$\begin{aligned}\hat{\boldsymbol{\beta}}_J &= \{\boldsymbol{X}^T[\boldsymbol{\Sigma_\epsilon}^{-1} - \boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}]\boldsymbol{X}\}^{-1}\boldsymbol{X}^T[\boldsymbol{\Sigma_\epsilon}^{-1} - \boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma_\epsilon}^{-1}]\boldsymbol{y} \\ &= [\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma_\epsilon})^{-1} \\ &= \hat{\boldsymbol{\beta}}_M\end{aligned}$$

In order to prove the identity between the EBP of $\boldsymbol{b}$, $\tilde{\boldsymbol{b}}_M = \boldsymbol{\Sigma}_B \boldsymbol{Z}^T (\boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_B \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ and the BLUP of $\boldsymbol{b}$, $\tilde{\boldsymbol{b}}_J = (\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}} \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}$ we must show that

$$\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}} \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} = \boldsymbol{\Sigma}_B \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} \tag{26}$$

Post-multiplying $\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}} \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}$ by $\boldsymbol{\Sigma}_{\boldsymbol{y}} \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$, we get:

$$\begin{aligned}
(\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} &= (\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})(\boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} \\
&= (\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} (\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{Z}^T)(\boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} \\
(\text{inserting } \boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}_B) &= (\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} (\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T)(\boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} \\
&= (\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} (\boldsymbol{Z}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1}) \boldsymbol{\Sigma}_B \boldsymbol{Z}^T (\boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} \\
&= \boldsymbol{\Sigma}_B \boldsymbol{Z}^T (\boldsymbol{Z}\boldsymbol{\Sigma}_B \boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}
\end{aligned} \tag{27}$$

Thanks to these identities, in Linear Mixed Models it is irrelevant which approach (marginal + EBP vs. joint) is used, and we can drop the subscripts M and J and write simply: $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{b}}$. The result is also crucial for what *it does not imply*: this convenient identity *does not hold* for nonlinear - non-Gaussian mixed models, for example for Generalised Linear Mixed Models (GLMMs). The lack of exchangeability between the joint and the marginal approach is one of the main difficulties in the effort to extend the methods here described for Linear Mixed Models to Generalised Linear Mixed Models.

5 With unknown $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \boldsymbol{\Sigma}_B$

In practice, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Sigma}_B$ are very rarely known, and therefore they must be estimated from the data. As well known, there are two approaches in the literature for estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Sigma}_B$: Maximum Likelihood (ML) and *Restricted* (or *Residual*) Maximum Likelihood (REML). Actually, in general the so-called *variance components*, i.e. the variances and covariances in $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Sigma}_B$, depend on a limited number, say s, of parameters, which in this section will be collected in an s- vector and denoted by $\boldsymbol{\phi}$; when needed, we stress this dependence by writing $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi})$. Notice that $\boldsymbol{\phi}$ lies in general in a restricted parametric space, which ensures admissible estimators for $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi})$, i.e.:

$$\boldsymbol{\phi} \in \boldsymbol{\Phi} \qquad \text{such that} \quad \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}), \boldsymbol{\Sigma}_B(\boldsymbol{\phi}) \quad \text{are positive definite}$$

5.1 ML estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Sigma}_B$

Assuming now $\boldsymbol{\beta}$ known, the marginal log-likelihood (17) can be written as a function of $\boldsymbol{\phi}$, given $\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}$ and $\boldsymbol{Z}$:

$$\ell(\boldsymbol{\phi}; \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = -\left(\frac{n}{2}\right) \log(2\pi) - \left(\frac{1}{2}\right) \log(|\boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})|) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{28}$$

Since $\ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})$ is a non-linear function of $\boldsymbol{\phi}$, its maximization with respect to $\boldsymbol{\phi}$ requires iterative methods. In particular, following Lindstrom and Bates (1988), we adopt either a Newton-Raphson or a Fisher scoring algorithm, which require explicit expressions for the score vector:

$$\boldsymbol{u}(\boldsymbol{\phi}) = \frac{\partial \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_j} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_s} \end{bmatrix}$$

the observed Fisher information matrix, i.e. the Hessian matrix with negative sign:

$$\mathcal{J}(\boldsymbol{\phi}) = -\frac{\partial^2 \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^T} = \left\{ -\frac{\partial^2 \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_j \partial \phi_k} \right\}$$

and the Fisher information matrix, i.e. the expected value of the observed information matrix:

$$\mathcal{I}(\boldsymbol{\phi}) = \boldsymbol{E}\left[\mathcal{J}(\boldsymbol{\phi})\right] = \boldsymbol{E}\left[-\frac{\partial^{\in} \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^{\mathcal{T}}} \right] = \left\{ \boldsymbol{E}\left[-\frac{\partial^{\in} \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_{|} \partial \phi_{\|}} \right] \right\}$$

The expressions for the score vector, the Hessian and the Fisher information matrix, which were first derived by Harville (1977), written in matrix form are quite cumbersome. It is more convenient to present them element-wise:

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_j} &= -\left(\frac{1}{2}\right) \mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \right] \\ &\quad + \left(\frac{1}{2}\right) (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad (29) \\ \frac{\partial^2 \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_j \partial \phi_k} &= -\left(\frac{1}{2}\right) \mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial^2 \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k} \right) - \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \\ &\quad + \left(\frac{1}{2}\right) (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left[\left(\frac{\partial^2 \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k} \right) - 2 \left(\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \times \\ &\quad \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad (30) \\ \boldsymbol{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_j \partial \phi_k} \right] &= \left(\frac{1}{2}\right) \mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_j} \right) \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})^{-1} \left(\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\phi})}{\partial \phi_k} \right) \right] \qquad (31) \end{aligned}$$

Once the derivatives in (29), (30) and (31) have been determined, estimation of $\boldsymbol{\phi}$, and hence estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi})$, can be carried out either by the Newton-Raphson iterative procedure:

$$\hat{\boldsymbol{\phi}}_{(s+1)} = \hat{\boldsymbol{\phi}}_{(s)} - \mathcal{J}(\hat{\boldsymbol{\phi}}_{(s)})^{-1} \boldsymbol{u}(\hat{\boldsymbol{\phi}}_{(s)}) \qquad (32)$$

or by the Fisher scoring iterative procedure:

$$\hat{\boldsymbol{\phi}}_{(s+1)} = \hat{\boldsymbol{\phi}}_{(s)} + \mathcal{I}(\hat{\boldsymbol{\phi}}_{(s)})^{-1} \boldsymbol{u}(\hat{\boldsymbol{\phi}}_{(s)}) \qquad (33)$$

5.2 REML estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Sigma}_B$

Although consistent, the MLE of variance/covariance parameters are well known to be downward biased in finite samples. This drawback is already known for the MLE $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$ of the variance of a $\mathcal{N}(\mu, \sigma^2)$ distribution, with μ unknown, estimated on an i.i.d. sample of size n. The bias comes from the denominator: dividing the sample deviance by n does not take into account the loss of one degree of freedom caused by the estimation of μ. The obvious remedy, usually taught in basic curses in Statistical Inference, is to correct for this bias, using the corrected estimator $s^2 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$, obtained dividing the sample deviance by $(n-1)$, instead of n, which incorporates the right number of degrees of freedom available for estimation of σ^2 after estimating μ, and is therefore unbiased.

It turns out that this simple correction is actually the first, simplest example of application of a more generally strategy: estimation of σ^2 by *REstricted Maximum Likelihood* (often denoted with the acronym *REML*), instead of the standard Maximum Likelihood.

Since the bias for the ML estimators comes from the need to estimate first the unknown parameters in $\boldsymbol{\mu}$, the basic idea in REML estimation is to get rid of the unknown $\boldsymbol{\mu}$ parameter when estimating variance/covariance parameters in a Multivariate Normal distribution. This can be achieved by choosing *any* matrix $\boldsymbol{K}$ such that $\boldsymbol{E}[\boldsymbol{K}^T\boldsymbol{y}] = \boldsymbol{0}$ and hence $\boldsymbol{K}^T\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{K})$. This choice effectively removes the need of estimating $\boldsymbol{\mu}$, or any parameters modelling $\boldsymbol{\mu}$, before estimating $\boldsymbol{\Sigma}_{\boldsymbol{y}}$.

The idea of using the REML approach first appeared in the statistical literature in the '50s; Patterson and Thompson (1971) presented a comprehensive treatment of REML theory applied to LMMs. It is important to notice that the estimators obtained for $\boldsymbol{\phi}$, and hence for $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi})$, are invariant to the choice of $\boldsymbol{K}$. The most typical choice for $\boldsymbol{K}$ is the OLS orthogonal projection matrix of the (marginal) residuals:

$$\boldsymbol{K} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

which clearly satisfies the property $\boldsymbol{E}[\boldsymbol{K}\boldsymbol{y}] = \boldsymbol{E}[(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y}] = \boldsymbol{E}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}] = \boldsymbol{0}$. For this reason, some authors prefer to interpret the acronym *REML* as *REsidual Maximum Likelihood*, although this seems a rather restrictive interpretation, since this choice of $\boldsymbol{K}$ matrix is just one of the (actually infinite) equivalent alternatives.

Again, the expressions for the score vector, the Hessian and the Fisher information matrix, which were first derived by Harville (1977), are more conveniently presented element-wise:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_j} &= -\left(\frac{1}{2}\right)\mathrm{tr}\left[\boldsymbol{M}\left(\frac{\partial \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_j}\right)\right] \\ &\quad + \left(\frac{1}{2}\right)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})^{-1}\left(\frac{\partial \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_j}\right)\boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \qquad (34)\\ \frac{\partial^2 \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_j \partial \phi_k} &= -\left(\frac{1}{2}\right)\mathrm{tr}\left[\boldsymbol{M}\left(\frac{\partial^2 \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k}\right) - \boldsymbol{M}\left(\frac{\partial \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_j}\right)\boldsymbol{M}\left(\frac{\partial \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_k}\right)\right] \\ &\quad + \left(\frac{1}{2}\right)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})^{-1}\left[\left(\frac{\partial^2 \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k}\right) - 2\left(\frac{\partial \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_j}\right)\boldsymbol{M}\left(\frac{\partial \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_k}\right)\right] \\ &\quad \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \qquad (35)\\ \boldsymbol{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\phi};\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})}{\partial \phi_j \partial \phi_k}\right] &= \left(\frac{1}{2}\right)\mathrm{tr}\left[\boldsymbol{M}\left(\frac{\partial \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_j}\right)\boldsymbol{M}\left(\frac{\partial \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})}{\partial \phi_k}\right)\right] \qquad (36)\end{aligned}$$

where:

$$\boldsymbol{M} = \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})^{-1} - \boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})^{-1}\boldsymbol{X}(\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{y}(\boldsymbol{\phi})^{-1}$$

Once the derivatives in (34), (35) and (36) have been determined, estimation of $\boldsymbol{\phi}$, and hence estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi})$, can be carried out either by the Newton-Raphson iterative procedure:

$$\hat{\boldsymbol{\phi}}_{(s+1)} = \hat{\boldsymbol{\phi}}_{(s)} - \mathcal{J}(\hat{\boldsymbol{\phi}}_{(s)})^{-1}\boldsymbol{u}(\hat{\boldsymbol{\phi}}_{(s)}) \qquad (37)$$

or by the Fisher scoring iterative procedure:

$$\hat{\boldsymbol{\phi}}_{(s+1)} = \hat{\boldsymbol{\phi}}_{(s)} + \mathcal{I}(\hat{\boldsymbol{\phi}}_{(s)})^{-1}\boldsymbol{u}(\hat{\boldsymbol{\phi}}_{(s)}) \qquad (38)$$

5.3 Iterative algorithm for estimating $\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \boldsymbol{\Sigma}_B$ and predicting $\boldsymbol{b}$

Summing up, estimation of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, $\boldsymbol{\Sigma}_B$ and prediction of $\boldsymbol{b}$ proceeds through an iterative process that can be schematised as follows:

- initialise $\boldsymbol{\phi}$, and with the initial value $\boldsymbol{\phi}_{(0)}$ compute $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}_{(0)})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi}_{(0)})$
- considering $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}_{(0)})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{\phi}_{(0)})$ as known, estimate $\boldsymbol{\beta}$ and $\boldsymbol{b}$ with (23) and (24), to obtain $\hat{\boldsymbol{\beta}}_{(1)}$ and $\tilde{\boldsymbol{b}}_{(1)}$
- considering $\hat{\boldsymbol{\beta}}_{(1)}$ and $\tilde{\boldsymbol{b}}_{(1)}$ as known, estimate $\hat{\boldsymbol{\phi}}_{(1)}$ using one of (32), (33), (37) or (38) (depending on whether ML or REML and Newton Raphson or Fisher Scoring is chosen), and compute $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\phi}}_{(1)})$ and $\hat{\boldsymbol{\Sigma}}_B(\hat{\boldsymbol{\phi}}_{(1)})$
- .....
- iterate the previous steps until convergence

Once the algorithm has reached convergence, we can write the ML (or REML) estimators of $\boldsymbol{\phi}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}), \boldsymbol{\Sigma}_B(\boldsymbol{\phi})$ as $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}(\hat{\boldsymbol{\phi}}), \hat{\boldsymbol{\Sigma}}_B(\hat{\boldsymbol{\phi}})$ and the ML (or REML) estimator of $\boldsymbol{\beta}$ and predictor of $\boldsymbol{b}$ as:

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{X}^T(\boldsymbol{Z}^T\hat{\boldsymbol{\Sigma}}_B\boldsymbol{Z} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\hat{\boldsymbol{\Sigma}}_B\boldsymbol{Z}^T + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{y} \tag{39}$$

$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{\Sigma}}_B\boldsymbol{Z}^T(\boldsymbol{Z}\hat{\boldsymbol{\Sigma}}_B\boldsymbol{Z}^T + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{40}$$

6 Hodges' "augmented data" representation

Hodges (1998) (see also Hodges and Sargent (2001); Vaida and Blanchard (2005)) showed that the joint MLE of $\boldsymbol{\beta}$ and the BLUP of $\boldsymbol{b}$ can be obtained as the WLS solution of a unique (general) linear model, through the construction of an *"augmented data"* response vector. This approach starts from adding to the usual LMM specification (4) the obvious identity: $\mathbf{0} = \boldsymbol{b} - \boldsymbol{b}$ and constructing the "augmented response vector:

$$\boldsymbol{y}_+ = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix}$$

Then, model (4) can be compactly written as:

$$\begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{O} & -\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{b} \end{bmatrix} \tag{41}$$

or:

$$\boldsymbol{y}_+ = \boldsymbol{T}\boldsymbol{\gamma} + \boldsymbol{\delta}$$

with:

$$\boldsymbol{\delta} \sim \mathcal{MVN}_{n+r}(\mathbf{0}_{n+r}, \Sigma_{\boldsymbol{\delta}}), \quad , \Sigma_{\boldsymbol{\delta}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_B \end{bmatrix}$$

Since $\Sigma_{\boldsymbol{\delta}}$ is a full variance/covariance matrix, the estimator of $\boldsymbol{\gamma}$ is the solution of the Weighted Least Squares system of equations:

$$\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{T}\boldsymbol{\gamma} = \boldsymbol{T}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_+ \tag{42}$$

or, using the extended expressions for $\boldsymbol{T}, \boldsymbol{y}_+, \boldsymbol{\Sigma}_{\boldsymbol{\delta}}$:

$$\begin{bmatrix} \boldsymbol{X}^T & \boldsymbol{O} \\ \boldsymbol{Z}^T & -\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_B^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{O} & -\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T & \boldsymbol{O} \\ \boldsymbol{Z}^T & -\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_B^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix}$$

whence:

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \end{bmatrix} \tag{43}$$

From (43) we see that Hodges' augmented data weighted least squares system corresponds exactly to Henderson's mixed model equations (22). As a consequence, it yields the same estimators for $\boldsymbol{\beta}$ and $\boldsymbol{b}$. To prove it, let us denote:

$$\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{T} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1} \end{bmatrix} = \boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} \\ \boldsymbol{U}_{21} & \boldsymbol{U}_{11} \end{bmatrix} \tag{44}$$

and

$$\boldsymbol{U}^{-1} = \begin{bmatrix} \boldsymbol{U}^{11} & \boldsymbol{U}^{12} \\ \boldsymbol{U}^{21} & \boldsymbol{U}^{11} \end{bmatrix} \tag{45}$$

Since the analytic expression for $\boldsymbol{U}^{-1}$ is rather cumbersome, it is given in Appendix 2. From (42) we obtain:

$$\begin{aligned}
\hat{\gamma} &= (\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_+ \qquad (46)\\
&= \boldsymbol{U}^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_+ = \begin{bmatrix} \boldsymbol{U}^{11} & \boldsymbol{U}^{12} \\ \boldsymbol{U}^{21} & \boldsymbol{U}^{11} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{U}^{11}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} + \boldsymbol{U}^{12}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \\ \boldsymbol{U}^{21}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} + \boldsymbol{U}^{22}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}^{11}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} + \boldsymbol{U}^{11}\boldsymbol{U}_{12}\boldsymbol{U}_{22}^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \\ -\boldsymbol{U}_{22}^{-1}\boldsymbol{U}_{21}\boldsymbol{U}^{11}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} + (\boldsymbol{U}_{22}^{-1} + \boldsymbol{U}_{22}^{-1}\boldsymbol{U}_{21}\boldsymbol{U}^{11}\boldsymbol{U}^{12^T}\boldsymbol{U}_{22}^{-1})\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{U}^{11}\boldsymbol{X}^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}]\boldsymbol{y} \\ \boldsymbol{U}_{22}^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} - \boldsymbol{U}_{22}^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}\boldsymbol{U}^{11}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} + \boldsymbol{U}_{22}^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{U}^{12^T}\boldsymbol{U}_{22}^{-1})\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} \end{bmatrix} \\
&= \begin{bmatrix} (\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{y} \\ \boldsymbol{U}_{22}^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}[\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{U}^{11}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y} + \boldsymbol{U}^{11}\boldsymbol{U}_{12}\boldsymbol{U}_{22}^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1})\boldsymbol{y}] \end{bmatrix} \\
&= \begin{bmatrix} (\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{y} \\ (\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}] \end{bmatrix} \\
&= \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{b}} \end{bmatrix}
\end{aligned}$$

7 The Hat-matrix of Linear Mixed Models

Unlike what happens for Linear Models, and partly for Generalised Linear Models, the literature on the hat-matrix for Linear Mixed models is rather scarce and scattered. We begin by quoting briefly a few relevant references, and then move to discuss the form of hat-matrix which appears to be the closest generalisation of the Linear Models hat-matrix in terms of mathematical and statistical properties.

7.1 A (short) literature review

Zewotir and Galpin (2007) give an expression for the hat-matrix for the residuals, rather than for the fitted values, of Linear Mixed Models. The model specification they employ is slight

different from ours, since they restrict the variance/covariance matrix of the (conditional) errors to be homoscedastic and the variance/covariance matrix of the random parameters to be also homoscedastic and defined in terms of *realative variances* $\tau_{B_j} = \dfrac{\sigma^2_{B_j}}{\sigma_\epsilon}$:

$$\mathcal{D}_{ZG}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon};ZG} = \Sigma^2_\epsilon \boldsymbol{I}_n \qquad \mathcal{D}_{ZG}[\boldsymbol{b}] = \boldsymbol{\Sigma}_{B;ZG} = \sigma^2_\epsilon \boldsymbol{\Gamma}_{\boldsymbol{B};ZG}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{B};ZG} = \boldsymbol{I}_m \otimes diag(\boldsymbol{\tau})$ and $\boldsymbol{\tau} = [\tau_{B_1}, \ldots, \tau_{B_j}, \ldots, \tau_{B_q}]^T$

With these assumptions, the marginal dispersion matrix of $\boldsymbol{y}$ is:

$$\mathcal{D}_{ZG}[\boldsymbol{y}] = \boldsymbol{\Sigma}_{\boldsymbol{y};ZG} = \sigma_\epsilon(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)$$

Zewotir and Galpin (2007) propose a hat-matrix for the conditional residuals $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\tilde{\boldsymbol{b}}$:

$$\boldsymbol{R} = (\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)^{-1} - (\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)^{-1}\boldsymbol{X}[\boldsymbol{X}^T(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)^{-1}$$

such that $\boldsymbol{r} = \boldsymbol{R}\boldsymbol{y}$.

Clearly, the corresponding hat-matrix for the fitted values $\hat{\boldsymbol{\mu}}$ is:

$$\boldsymbol{H}_{ZG} = \boldsymbol{I} - (\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)^{-1} + (\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)^{-1}\boldsymbol{X}[\boldsymbol{X}^T(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{\Gamma}_{B;ZG}\boldsymbol{Z}^T)^{-1} \tag{47}$$

Demidenko and Stukel (2005) (see also Singer et al. (2004), Nobre and Singer (2011)) move from the idea of *generalised leverage matrix* (Wei et al., 1998), which, for *any* model $\boldsymbol{y} = \boldsymbol{\mu}(\boldsymbol{\beta}) + \boldsymbol{\epsilon}$, is defined as:

$$\boldsymbol{H} = \frac{\partial \hat{\boldsymbol{\mu}}}{\partial \boldsymbol{y}^T} \tag{48}$$

i.e. as the matrix having as generic element $\{h_{ij}\}$ the rate of change of μ_i with respect to the observation y_j. Applying (48) to the Linear Mixed Model (4) we obtain:

$$\begin{aligned}
\boldsymbol{H}_{DS} = \frac{\partial(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{b}})}{\partial \boldsymbol{y}^T} &= \frac{\partial\{\boldsymbol{X}(\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{y}\}}{\partial \boldsymbol{y}^T} \\
&+ \frac{\partial\{\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}(\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{y}\}}{\partial \boldsymbol{y}^T} \\
&= \boldsymbol{X}[\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} \\
&+ \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} - \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}[\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}
\end{aligned} \tag{49}$$

Demidenko and Stukel (2005) write (49) as the sum of two generalised leverage matrices:

$$\boldsymbol{H}_{DS} = \boldsymbol{H}_{DS_1} + \boldsymbol{H}_{DS_2}$$

where:

$$\boldsymbol{H}_{DS_1} = \boldsymbol{X}[\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} \tag{50}$$

is the *generalised marginal leverage matrix* and

$$\begin{aligned}\boldsymbol{H}_{DS_2} &= \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T+\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} - \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T+\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}[\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T+\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T+\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} \\ &= \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T+\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}[\boldsymbol{I}_n - \boldsymbol{H}_{DS_1}] \end{aligned} \tag{51}$$

is the *generalised random-effects leverage matrix*.

In passing, it is worthwhile mentioning that Nobre and Singer (2011), noticing that $\boldsymbol{H}_{DS_2}$ depends on the generalised marginal leverage matrix $\boldsymbol{H}_{DS_1}$, in order to get a leverage matrix for the random effects which is not confounded with that for the fixed effects, propose to use, as an alternative to $\boldsymbol{H}_{DS_2}$:

$$\boldsymbol{H}_{NS_2} = \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T \tag{52}$$

7.2 The Hat-matrix in the Hodges' "augmented-data" representation

The main drawback of the Galpin-Zewotir's and the Demidenko-Stukel's definitions of hat-matrix for Linear Mixed Models is that they are not orthogonal projection matrices. This is readily seen by checking that they are not idempotent:

$$\boldsymbol{H}_{ZG}\boldsymbol{H}_{ZG} \neq \boldsymbol{H}_{ZG}; \qquad \boldsymbol{H}_{DS}\boldsymbol{H}_{DS} \neq \boldsymbol{H}_{DS}$$

For the context where these hat-matrices were proposed, that of influence diagnostics, this is not a serious problem, but whenever a hat-matrix is required to operate an orthogonal decomposition of the response vector into a vector of fitted values and a vector of residuals, the requirement of idempotency becomes paramount.

The Hodges's "augmented-data" representation, unlike the Galpin-Zewotir and the Demidenko-Stukel approaches, provides an orthogonal projection hat-matrix. To show this result, we start by defining the Hodges's hat-matrix. From:

$$\hat{\boldsymbol{\mu}}_+ = \boldsymbol{T}\hat{\boldsymbol{\gamma}} = \boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_+$$

we see that the (unscaled) "augmented-data" hat-matrix is:

$$\boldsymbol{H}_+ = \boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}$$

$\boldsymbol{H}_+$ is not symmetric, but if we consider the scaled version of $\boldsymbol{y}_+$ and $\hat{\boldsymbol{\mu}}_+$, $\boldsymbol{y}_+^* = \boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{y}_+$ and $\hat{\boldsymbol{\mu}}_+^* = \boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2}\hat{\boldsymbol{\mu}}_+$, we obtain:

$$\hat{\boldsymbol{\mu}}_+^* = \boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{y}_+^*$$

whence, the scaled "augmented-data" hat-matrix is seen to be:

$$\boldsymbol{H}_+^* = \boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2} \tag{53}$$

The hat-matrix $\boldsymbol{H}^*_+$ is clearly symmetric and idempotent: as such, it is the matrix which projects orthogonally the scaled "augmented" response vector $\boldsymbol{y}^*_+$ onto the space of the scaled "augmented" fitted vector $\hat{\boldsymbol{\mu}}^*_+$. In this respect, it can be considered as the closest generalisation of the usual hat-matrix of Linear Models to Linear Mixed Models.

It is useful to derive the extended form of $\boldsymbol{H}^*_+$. Using (45), we can write:

$$\boldsymbol{H}^*_+ = \boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{T}\boldsymbol{U}^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2} \tag{54}$$

$$= \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{X} & \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{Z} \\ \boldsymbol{O} & -\boldsymbol{\Sigma}_B^{-1/2} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}^{11} & \boldsymbol{U}^{12} \\ \boldsymbol{U}^{21} & \boldsymbol{U}^{11} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2} & \boldsymbol{O} \\ \boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & -\boldsymbol{\Sigma}_B^{-1} \end{bmatrix} \tag{55}$$

$$= \begin{bmatrix} \boldsymbol{H}^*_{+11} & \boldsymbol{H}^*_{+12} \\ \boldsymbol{H}^*_{+21} & \boldsymbol{H}^*_{+22} \end{bmatrix} \tag{56}$$

where:

$$\boldsymbol{H}^*_{+11} = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{X}\boldsymbol{U}^{11}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{X}\boldsymbol{U}^{12}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{Z}\boldsymbol{U}^{21}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{Z}\boldsymbol{U}^{22}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}$$
$$\boldsymbol{H}^*_{+12} = -\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{X}\boldsymbol{U}^{12}\boldsymbol{\Sigma}_B^{-1/2} - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{Z}\boldsymbol{U}^{22}\boldsymbol{\Sigma}_B^{-1/2}$$
$$\boldsymbol{H}^*_{+21} = -\boldsymbol{\Sigma}_B^{-1/2}\boldsymbol{U}^{12}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2} - \boldsymbol{\Sigma}_B^{-1/2}\boldsymbol{U}^{22}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}$$
$$\boldsymbol{H}^*_{+22} = \boldsymbol{\Sigma}_B^{-1/2}\boldsymbol{U}^{22}\boldsymbol{\Sigma}_B^{-1/2}$$

Also, it is easy to check that the scaled "augmented" fitted vector $\hat{\boldsymbol{\mu}}^*_+$ turns out to be:

$$\hat{\boldsymbol{\mu}}^*_+ = \boldsymbol{H}^*_+\boldsymbol{y}^*_+ = \begin{bmatrix} \boldsymbol{H}^*_{+11} & \boldsymbol{H}^*_{+12} \\ \boldsymbol{H}^*_{+21} & \boldsymbol{H}^*_{+22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{H}^*_{+11}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{y} \\ \boldsymbol{H}^*_{+21}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{Z}\tilde{\boldsymbol{b}} \\ -\boldsymbol{\Sigma}_B^{-1/2}\tilde{\boldsymbol{b}} \end{bmatrix} \tag{57}$$

i.e. the scaled "augmented" fitted vector $\hat{\boldsymbol{\mu}}^*_+$ has two components: the conditional (within clusters) scaled fitted values $\widehat{\boldsymbol{\mu}|\boldsymbol{b}}^* = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1/2}(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{b}})$ and the scaled predicted random parameters (with negative sign) $-\boldsymbol{\Sigma}_B^{-1/2}\tilde{\boldsymbol{b}}$.

7.3 Equivalence between $\boldsymbol{H}_{ZG}$, $\boldsymbol{H}_{DS}$ and $\boldsymbol{H}_{+11}$

It is natural to wonder whether, although apparently quite different, $\boldsymbol{H}_{ZG}$, $\boldsymbol{H}_{DS}$ and the Hodges' hat-matrix are equivalent. It is possible to prove that it is actually the case, as long as the different assumptions made for the Zewotir-Galpin hat-matrix are accounted for and only the upper-left block of the "unscaled" version of the Hogdes's hat-matrix, denoted by $\boldsymbol{H}_{+11}$, is used in the comparison.

Consider the the (unscaled) "augmented-data" Hodges' hat-matrix:

$$\boldsymbol{H}_+ = \boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1}$$

By a derivation parallel to the one used in (54), (55), (56) for the "scaled" version, we get:

$$\boldsymbol{H}_+ = \boldsymbol{T}\boldsymbol{U}^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1} \tag{58}$$

$$= \begin{bmatrix} \boldsymbol{X}\boldsymbol{U}^{11}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{X}\boldsymbol{U}^{12}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{Z}\boldsymbol{U}^{21}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{Z}\boldsymbol{U}^{22}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & -\boldsymbol{X}\boldsymbol{U}^{12}\boldsymbol{\Sigma}_B^{-1} - \boldsymbol{Z}\boldsymbol{U}^{22}\boldsymbol{\Sigma}_B^{-1} \\ -\boldsymbol{U}^{12}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{U}^{22}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} & \boldsymbol{U}^{22}\boldsymbol{\Sigma}_B^{-1/2} \end{bmatrix} \tag{59}$$

$$= \begin{bmatrix} \boldsymbol{H}_{+11} & \boldsymbol{H}_{+12} \\ \boldsymbol{H}_{+21} & \boldsymbol{H}_{+22} \end{bmatrix} \tag{60}$$

As observed by Vaida and Blanchard (2005), the upper left block $\boldsymbol{H}_{+11}$ is the hat-matrix of the (conditional) within-cluster fitted values $\widehat{\boldsymbol{\mu}|\boldsymbol{b}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{b}}$, but it is not an orthogonal projection matrix, since it is neither symmetric nor idempotent. We now set out to prove that it is this block of the "unscaled" Hodges' hat-matrix $\boldsymbol{H}_+$ which is equivalent to Zewotir-Galpin $\boldsymbol{H}_{ZG}$ and to Demidenko-Stukel $\boldsymbol{H}_{DS}$.

Result 1

Under the restrictive assumptions made by Zewotir and Galpin (2007):

$$\boldsymbol{H}_{ZG} = \boldsymbol{H}_{+11} \tag{61}$$

Proof Using (73), (75), (77) and (80) in Appendix 2, we can write $\boldsymbol{H}_{+11}$ as follows:

$$\begin{aligned}\boldsymbol{H}_{+11} = {} & \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \\ & + \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\end{aligned} \tag{62}$$

But:

$$\begin{aligned}& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \\ -\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} = \\ & \left[\boldsymbol{I}_n - \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\right]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\left[\boldsymbol{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\right]\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\end{aligned}$$

and

$$\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} = \boldsymbol{I}_n - (\boldsymbol{I}_n - \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1})$$

whence (62) can be written:

$$\boldsymbol{H}_{+11} = \boldsymbol{I}_n - (\boldsymbol{I}_n - \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}) + \left[\boldsymbol{I}_n - \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\right]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\left[\boldsymbol{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\right]\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \tag{63}$$

It is immediate to recognise that $\left[\boldsymbol{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\right] = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$. This follows from (25):

$$\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{b}|\boldsymbol{y}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}$$

whence:

$$\boldsymbol{\Sigma_\epsilon \Sigma_y^{-1}} = \boldsymbol{\Sigma_\epsilon}[\boldsymbol{\Sigma_\epsilon^{-1}} - \boldsymbol{\Sigma_\epsilon^{-1} Z \Sigma_{b|y} Z^T \Sigma_\epsilon^{-1}}] = \boldsymbol{I}_n - \boldsymbol{Z \Sigma_{b|y} Z^T \Sigma_\epsilon^{-1}} \tag{64}$$

$$\boldsymbol{\Sigma_y^{-1} \Sigma_\epsilon} = \boldsymbol{I}_n - \boldsymbol{\Sigma_\epsilon^{-1} Z \Sigma_{b|y} Z^T} \quad \text{(by symmetry)} \tag{65}$$

and (63) can be written as:

$$\boldsymbol{H}_{+11} = \boldsymbol{I}_n - \boldsymbol{\Sigma_\epsilon \Sigma_y^{-1}} + \boldsymbol{\Sigma_\epsilon \Sigma_y^{-1} X (X^T \Sigma_y^{-1} X)^{-1} X^T \Sigma_y^{-1}} \tag{66}$$

Now, recalling that under the restrictive assumptions of Zewotir and Galpin (2007) paper:

$$\boldsymbol{\Sigma_{y;ZG}} = \sigma_\epsilon(\boldsymbol{I}_n + \boldsymbol{Z \Gamma}_{B;ZG} \boldsymbol{Z}^T); \quad \boldsymbol{\Sigma_{\epsilon;ZG}} = \sigma_\epsilon^2 \boldsymbol{I}_n; \quad \boldsymbol{\Sigma_{\epsilon;ZG} \Sigma_{y;ZG}^{-1}} = (\boldsymbol{I}_n + \boldsymbol{Z \Gamma}_{B;ZG} \boldsymbol{Z}^T)^{-1}$$

formula (66) yields:

$$\begin{aligned}
\boldsymbol{H}_{+11} &= \boldsymbol{I}_n - (\boldsymbol{I}_n + \boldsymbol{Z \Gamma}_{B;ZG} \boldsymbol{Z}^T)^{-1} \\
&\quad + (\boldsymbol{I}_n + \boldsymbol{Z \Gamma}_{B;ZG} \boldsymbol{Z}^T)^{-1} \boldsymbol{X} [\boldsymbol{X}^T \frac{1}{\sigma_\epsilon^2} (\boldsymbol{I}_n + \boldsymbol{Z \Gamma}_{B;ZG} \boldsymbol{Z}^T)^{-1} \boldsymbol{X}]^{-1} \boldsymbol{X}^T \frac{1}{\sigma_\epsilon^2} (\boldsymbol{I}_n + \boldsymbol{Z \Gamma}_{B;ZG} \boldsymbol{Z}^T)^{-1} \\
&= \boldsymbol{H}_{ZG}
\end{aligned}$$

Q.E.D.

Result 2

$$\boldsymbol{H}_{DS} = \boldsymbol{H}_{+11} \tag{67}$$

Proof We can re-write (63) as:

$$\begin{aligned}\boldsymbol{H}_{+11} = &\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\left[\boldsymbol{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\right]\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \\ &+\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\left[\boldsymbol{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\right]\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\end{aligned}$$

From (65), we can substitute $\left[\boldsymbol{I}_n - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\right]\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$, whence:

$$\begin{aligned}\boldsymbol{H}_{+11} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} + \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\left[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\right] \\ &= \boldsymbol{H}_{DS_1} + \boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{I}_n - \boldsymbol{H}_{DS_1})\end{aligned}$$

Since from (26) we know that $\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} = \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$, it follows that:

$$\boldsymbol{H}_{+11} = \boldsymbol{H}_{DS_1} + \boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{I}_n - \boldsymbol{H}_{DS_1}) = \boldsymbol{H}_{DS}$$

Q.E.D.

A Appendix 1: A useful matrix algebra result

In this Appendix we recall, without proof, a useful result concerning the inverse of a Schur complement $(\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^T)^{-1}$ and of the associated form $(\boldsymbol{A} + \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^T)^{-1}$. The Schur complement, and its inverse, are very important in the derivations of theoretical results, in particular in Linear and Linear Mixed Models. The reader can find a formal proof in Henderson and Searle (1981) (equation 17).

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be nonsingular symmetric matrices of order n and m respectively, and let $\boldsymbol{C}$ be an $n \times m$ rectangular matrix. Then:

$$(\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}(\boldsymbol{B}^{-1} - \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{A}^{-1} \tag{68}$$

$$(\boldsymbol{A} + \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}(\boldsymbol{B}^{-1} + \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{A}^{-1} \tag{69}$$

B Appendix 2: The inverse of matrix U in Hodges' representation

Among various possible representations (see e.g. Seber (2008)) of the inverse of a partitioned matrix, we choose the following:

$$\begin{aligned}
\boldsymbol{U}^{11} &= (\boldsymbol{U}_{11} - \boldsymbol{U}_{12}\boldsymbol{U}_{22}^{-1}\boldsymbol{U}_{21})^{-1} \\
&= [\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}]^{-1} && (70)\\
&= \{\boldsymbol{X}^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}]\boldsymbol{X}\}^{-1} && (71)\\
&= [\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T(\boldsymbol{Z}\boldsymbol{\Sigma}_B\boldsymbol{Z}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1} && (72)\\
&= (\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} && (73)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{U}^{12} &= -\boldsymbol{U}^{11}\boldsymbol{U}_{12}\boldsymbol{U}_{22}^{-1} \\
&= -[\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} && (74)\\
&= -(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}} && (75)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{U}^{21} &= \boldsymbol{U}^{12^T} = \boldsymbol{U}_{22}^{-1}\boldsymbol{U}_{21}\boldsymbol{U}^{11} = \\
&= -(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}]^{-1} && (76)\\
&= -\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1} && (77)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{U}^{22} &= \boldsymbol{U}_{22}^{-1} + \boldsymbol{U}_{22}^{-1}\boldsymbol{U}_{21}\boldsymbol{U}^{11}\boldsymbol{U}^{12^T}\boldsymbol{U}_{22}^{-1} \\
&= (\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} \\
&\quad +(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}]^{-1} && (78)\\
&\quad \boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}_B^{-1})^{-1} && (79)\\
&= \boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}} + \boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}}\boldsymbol{Z}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{y}} && (80)
\end{aligned}$$

References

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, 443–458.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association 88*(421), 9–25.

Demidenko, E. and T. A. Stukel (2005). Influence analysis for linear mixed-effects models. *Statistics in medicine 24*(6), 893–909.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*(358), 320–338.

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. Von Krosigk (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*(2), 192–218.

Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices. *Siam Review 23*(1), 53–60.

Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(3), 497–536.

Hodges, J. S. and D. J. Sargent (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 367–379.

Lindstrom, M. J. and D. M. Bates (1988). Newtonraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association 83*(404), 1014–1022.

Nobre, J. S. and J. M. Singer (2011). Leverage analysis for linear mixed models. *Journal of Applied Statistics 38*(5), 1063–1072.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 545–554.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, 15–32.

Searle, J. R. (1992). *The rediscovery of the mind*. MIT press.

Searle, J. R., J. L. Austin, P. Strawson, H. Grice, N. Chomsky, J. J. Katz, N. Goodman, and H. Putnam (1971). *The philosophy of language*, Volume 39. Oxford University Press London.

Seber, G. A. (2008). *A matrix handbook for statisticians*, Volume 15. John Wiley & Sons.

Singer, J. M., J. S. Nobre, and H. C. Sef (2004). Regression models for pretest/posttest data in blocks. *Statistical Modelling 4*(4), 324–338.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1135–1151.

Vaida, F. and S. Blanchard (2005). Conditional akaike information for mixed-effects models. *Biometrika 92*(2), 351–370.

Wei, B.-C., Y.-Q. Hu, and W.-K. Fung (1998). Generalized leverage and its applications. *Scandinavian Journal of statistics 25*(1), 25–37.

Zewotir, T. and J. S. Galpin (2007). A unified approach on residuals, leverages and outliers in the linear mixed model. *Test 16*(1), 58–75.

Zhang, F. (2006). *The Schur complement and its applications*, Volume 4. Springer Science & Business Media.

UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 'in fase di assegnazione', volume I, 2017

Does regulatory regime matter for bank risk taking? A comparative analysis of US and Canada

Sana Mohsni · Isaac Otchere

Abstract The banking structure in the US is less concentrated, more competitive and less restrictive, whereas that of Canada is more concentrated, less competitive, and more restrictive. In the wake of the worst financial crisis in 2008, most US banks were bailed out while Canadian banks sailed through the crisis relatively unscathed. We examine the risk taking behavior of banks in the US and Canada prior to the recent financial crisis and find that Canadian banks had lower risk than their US counterparts over the study period. Further analysis shows that entry restrictions, which create concentrated banking structure, restrictions relating to capital, liquidity and activities, and strong supervisory power and discipline positively related to the z-score, suggesting that these factors constrain excessive risk taking by Canadian banks. We also decompose the z–score into its components and re-estimate our baseline regression with the view to identifying the source of the risk. We find that entry restrictions (and higher concentration) generate higher profits and lower variability of asset returns for Canadian bank whiles restrictions on activities reduce profitability and increases variability in asset return; however, the former seems to overwhelm the effect of asset restriction, given the lower risk that we observe for the Canadian banks. The less concentrated but competitive banking structure in the US is associated with higher bank risk taking.

Keywords Banks regulations · risk-taking · z-score

Isaac Otchere: Sprott School of Business, Carleton University, 1125 Colonel By Drive, Ottawa, ON, K1S 5B6, Canada;
E-mail: mailto:isaac_otchere@carleton.ca
· Sana Mohsni: Sprott School of Business,
Carleton University, 1125 Colonel By Drive, Ottawa, ON, K1S 5B6, Canada;
E-mail: sana_mohsni@carleton.ca.

Riassunto *La struttura del settore bancario negli Stati Uniti d'America è meno concentrata, più competitiva e meno restrittiva rispetto al medesimo settore in Canada. La maggior parte degli istituti creditizi statunitensi ha avuto bisogno di un piano di salvataggio sostenuto dal Governo per superare la crisi finanziaria del 2008, mentre le banche canadesi hanno superato la crisi finanziaria internazionale relativamente illesi. In questo lavoro analizziamo la tendenza all'assunzione di rischi da parte degli istituti creditizi statunitensi e canadesi precedentemente alla crisi del 2008. I risultati mostrano che le banche canadesi mostravano livelli di assunzione di rischi inferiori a quelli delle loro controparti statunitensi nel periodo indagato. Ulteriore analisi mostrano come restrizioni all'entrata nel settore, che hanno contribuito alla concentrazione elevata del settore, restrizioni in merito al capitale, alla liquidità e alle attività bancarie, assieme ad un forte poter di controllo e disciplina del settore sono correlate positivamente al valore dell'indicatore "z" (z-score), suggerendo che questi fattori limitano l'assunzione eccessiva di rischi da parte degli istituti creditizi canadesi. Al fine di individuare le singole fonti di rischio bancario, decomponiamo l'indicatore "z" e procediamo a calcolare regressioni partitamene per le singole variabili ad esso sottese. Troviamo che le restrizioni all'entrata nel settore (e alti livelli di concentrazione) contribuiscono a generare elevati tassi di profitto e ridotti gradi di variabilità del ROA degli istituti di credito, mentre le restrizioni sulle attività esercitate riducono i livelli di profitto e incrementano la variabilità del ROA degli stessi. Tuttavia, l'effetto della prima restrizione appare più significativa della seconda, dato il livello di rischi ridotto che caratterizza il sistema bancario canadese. In contrasto, la struttura meno concentrata e più competitiva del sistema bancario statunitense è associato ad un tasso di assunzione di rischi più elevato.*

Parole chiave *Regolamenti delle banche - assunzione di rischio - Score z*

1 Introduction

During the recent financial crisis, the Canadian banking system was described as resilient, healthy, and prudent; and was publicized as the soundest banking system in the world which enabled it to weather the global financial crisis well, and unlike many banks in the U.S and Europe, required no public funds injection. In fact, the *Financial Times* calls Canada's banks "the envy of the world." Paul Volcker, the former Federal Reserve Chairman has touted Canada's banks as the model for what a reformed American system should look like. Since the 2008 financial crisis, financial commentators and policy analysts have probed at the sources of the resilience of the financial system by posing and analyzing questions such as: 'Why banking crises happen in America but not in Canada?' (John Kay, Financial Times, June 4, 2014), 'How Canadian bank defied the financial crisis' (Constantine Passaris and Peter Bessey, The Daily Gleaner, March 2012) and 'Why was Canada exempt from the financial crisis?' (Renee Haltom, Econ Focus, 2013). Answers to these questions boil down primarily to the different regulatory regimes in the US and Canada.

While several factors account for such stability and resilience, but most relate to the difference in regulations. The Canadian financial system is characterized by a high degree of

concentration and consequently, a conservative and entrenched regulatory structure. Specifically, strong regulatory regime, stringent capital requirements for banks, federal supervision, strict mortgage market regulations, and a conservative appetite for risk are hallmarks of the Canadian financial system. Historically, the financial services sector has been strong in Canada. No banks collapsed in Canada during the Great Depression of the 1930s and only two small regional banks have gone out of business since 1923 (Bones, 2009). For instance, compared to the US, the banking system in Canada is more concentrated and more tightly regulated, have higher capital requirements, greater leverage restrictions, and fewer off-balance sheet activities. To minimize competing regulatory objectives, there is only one prudential (federal) regulator for both banks and insurance. Unlike the Canadian system, the US banking system is comprised of a very large number of small financial institutions and hence is much more fragmented. Unlike the US where each subsidiary of a banking conglomerate might be subject to a different regulatory authority (according to whether it was classed as an insurance company, investment bank, or commercial bank), in Canada, there is only one regulatory authority.

Another area where there is a sharp contrast between the regulatory regimes in Canada and the US is home ownership and mortgage practices. In Canada, mortgage interest is not tax deductible. The effect of these features of the mortgage market is that Canada has not seen a tax-driven distortion in the level of housing debt (Bones, 2009). It is a common practice for Canadian banks to hold mortgage on their balance sheets, which resulted in the application of high level of due diligence in the underwriting of bank loans. This has contributed to the fact that Canada did not experience the same degree of housing boom and bust that occurred in many other countries in the aftermath of the financial crisis. In contrast, there is mortgage interest deductibility for tax purposes in the U.S. (which encourages people to take on higher mortgages sometimes to fund purchases of consumer products), and longer mortgage amortization period. In addition, the implicit policy of financial access to the poor compelled banks in the US to relax their lending rules and in some cases engaged in what is now known as 'low document' mortgage lending where mortgage applicant's statement of their income level and job history were accepted without direct verification. This deviation from the traditional and prudential financial management practices in part triggered the financial crisis whose epicenter was the housing sector. This was in sharp contrast to the Canadian banks, which were not lured into the hype of risky investments that had the potential for high returns but carried excessive financial risk. Rather, they remained focused, exercising sound financial practices, holding adequate capital reserves as a buffer against financial emergencies (Passaris and Bessey, 2012).

Extant literature provides inconsistent results on the effect of regulation on risk taking. For instance, Gonzalez (2005) finds that banks in countries with stricter regulation have a lower charter value, which increases their incentive to follow risky policies. By contrast, Jin et al. (2013) find that banks that are required to comply with the FDICIA internal control requirements had lower risk taking in the pre-crisis period and are less likely to experience failure and financial trouble during the crisis period. Guyie and Lai (2003) test the presence of moral hazard in Canadian banks by analyzing the risk-taking behavior of Canadian commercial

banks following the introduction of flat rate deposit insurance in 1976. Their results show no evidence of an increase in moral hazard among Canadian commercial banks. Schawrtz and Zechner (1989) find that flat-rate deposit insurance system has resulted in cross-subsidization among Canadian commercial banks during 1980-1985.These results are not always consistent with studies that examine bank risk taking in the US market. For instance, Duan, Moreau and Sealey (1992) find that one fifth of the banks in their US sample of commercial banks succeed to transfer risk to the FDIC.

Barth et al. (2004) assess the relationship between specific regulatory and supervisory practices and banking-sector development, efficiency, and fragility. The paper examines: (i) regulatory restrictions on bank activities and the mixing of banking and commerce; (ii) regulations on domestic and foreign bank entry; (iii) regulations on capital adequacy; (iv) deposit insurance system design features; (v) supervisory power, independence, and resources; (vi) loan classification stringency, provisioning standards, and diversification guidelines; (vii) regulations fostering information disclosure and private-sector monitoring of banks; and (viii) government ownership. The results raise a cautionary flag regarding government policies that rely excessively on direct government supervision and regulation of bank activities. The findings instead suggest that policies that rely on guidelines that (1) force accurate information disclosure (2) empower private-sector corporate control of banks, and (3) foster incentives for private agents to exert corporate control work best to promote bank development, performance and stability. Saunder and Wilson (1999) investigate bank consolidation and safety-net support provision in Canada, the UK and the US over a 100-year historical period, and the impact of these policy variables on bank capital and risk-taking choices. They find among others that despite strengthened safety-net guarantees, bank capital ratios and bank asset-risk choices in the 1980s are comparable to those observed in the 1890s, while bank equity volatilities have shown approximately a 10-fold increase over this period. Kane and Wilson (2002) construct a synthetic time-series for banks safety-net capital in Canada and the U.S. and show that even in the absence of formal government guarantees, country safety nets sometimes enhanced substantially the value of each country's major banks.

Unlike the US where many banks collapsed and/or had to be bailed-out during the recent financial crisis, Canada did not experience any bank collapse. In this paper, we conduct a comparative analysis of the impact of the regulatory regimes on banks' risk taking by banks in the US and Canada to ascertain whether as a result of the more conservative and stringent regulatory regime the Canadian banks take on less risk in the pre-crisis period than the US banks. The objective of the paper is to analyze the effects of national regulation on risk taking of Canadian and US banks leading up to the financial crisis. The more conservative regulatory regime in Canada compared to the US suggests that risk taking by banks in Canada will be lower than that in the US. Although there appears to be anecdotal evidence to that effect, there has not been any systematicanalysis of whether the conservative and strict regulatory regime in Canada made them take less risk than the US banks. The results of such analysis will be relevant to policy makers around the world.

We find that Canadian banks had higher z-score and therefore lower risk than their US counterparts did over the study period. Our results based on a composite index created using World Bank survey results indicate that higher regulation and supervisory power leads to an increase in z-score (decrease in risk), however, an interaction of the country dummy and the composite regulatory index overall the Canadian regulatory index, as constructed using the World Bank survey, is less stringent than its US counterpart and produces results inconsistent with our earlier findings. The composite index could generate misleading or inconclusive results regarding the effects of regulations on bank risk taking as the different aspects of a country's regulatory system may be subsumed by responses to other survey questions that may have opposing effect. In view of this shortcoming, we use such sub indexes reflecting the different aspects of regulations. Our results show that entry restrictions which create concentrated banking structure, restrictions relating to capital, liquidity and activities, and strong supervisory power and discipline are positively related to the z-score, suggesting that these factors constrain excessive risk taking by Canadian banks. This effect is however, attenuated by the restrictions on banking activities, which lead to increase in risk. Combining these effects in a composite index could make the effects of regulations on risk taking ambiguous, which can explain the inconsistent results in the literature concerning the effects of regulations on risk taking.

We also decompose the z–score into its components and re-estimate our baseline regression with the view to identifying the source of the risk. We observe that ROA and the volatility of ROA are the channels through which the different aspects of a regulatory regime affect risk. We find that entry restrictions (and higher concentration) generate higher profits and lower variability of asset returns for Canadian bank whiles restrictions on activities reduce profitability and increase variability in asset return; however, the former seems to overwhelm the effect of asset restriction, given the lower risk that we observe for the Canadian banks. These results show that the lower risk and the stability experienced by the Canadian banks regime emanate from high profitability and lower profit variability- product of the concentrated nature of the banking industry, which seems to emphasize scale instead of competition.

Our study contributes to the literature in an important way. While the relationship between regulation and bank risk taking has been examined for other countries, the 2007-09 global financial crisis underscored the importance of financial regulation and surveillance, not only for the soundness of individual financial institutions, but also for the stability of the financial system as a whole. In the aftermath of one of the worst financial crises, there has not been any analysis of the impact of the regulatory regime type on risk taking behavior of banks in these countries with completely different regulatory regimes and experiences - where in one country the banks caused the financial crisis and in the other, the banks were resilient and sailed through the period relatively unscathed. Examining the risk taking behavior of banks that operate in a concentrated banking market with strong regulatory regime and those in a less concentrated less stringent regime with completely different outcome during the recent financial crisis adds to the regulation and risk taking literature.

The rest of the paper is organized as follows. The regulatory regimes in Canada and the United States and the hypotheses are discussed in Section 2. Section 3 deals with data and methodology. The analyses are presented in Section 4, and conclusion and policy implications are presented in Section 5.

2 Overview of Banking Regulations in the U.S. and Canada and hypothesis development

2.1 Differences in the US and Canadian Banking Systems

Some researchers argue that the reason why the Canadian banking system is resilient is more fundamental and that the Canadian banks stability and their stronger regulatory system compared to the United States are the result of divergent political systems and has deep historical roots. Calomiris and Haber in their 2014 book *Fragile by Design* argue that the structure of the U.S. political system allows popular interests to influence policy and regulations, which ultimately affect the risk taking behavior of the banks. In this section, we outline some of the more important differences between the U.S. and Canadian regulatory capital regimes, which will inform our hypothesis.

Structure:

The structure of the Canadian financial system (including the mortgage and housing markets) is different from what exists in the United States. The banking system in Canada consists of a small set of large and tightly regulated financial institutions defined by ownership, namely Canadian-owned banks, foreign banks with Canadian operations, cooperative banks, etc. These institutions of chartered banks offer financial stability in exchange for the Canadian government limiting entry to the industry. However, there are six major Canadian banks, namely Bank of Montreal, Scotiabank, TD Bank Financial Group, Royal Bank CIBC and the National Bank. These banks account for approximately 92.7% of thetotal assets and contributed almost 92% of the credit to loan markets (Office of the Superintendent of Financial Institutions Canada (2014). The concentration ratio was 92.7 % and the Herfindahl Index was 1679, which is considered to be a moderately concentrated market structure (Wu, 2015). Given the highly concentrated structure of the banking system where the six largest banks hold over 90 percent of total bank assets and since these banks perform key economic functions, the way they borrow funds, combined with the risks involved in their lending activities could in some circumstances threaten their solvency and can cause systemic problems; therefore, they need to be strictly regulated and supervised in order to ensure the stability and efficiency of the financial system.[1] Consequently, the Canadian financial system has been characterized by a strong regulatory regime.

[1] Systemic risk may arise if banks in the normal course of business take excessive risks that result in their failing, other banks may fail or be threatened with insolvency because of their connections with this failed bank. This could threaten the broader financial system and the performance of the economy.

In contrast, the US banking system is comprised of a very large number of small financial institutions and hence much more fragmented than its Canadian counterpart. According to the Conference of State Bank Supervisors (CSBS) in 2013, there were 6821 domestic and foreign institutions (CSBS, 2013). The six largest bank holding companies in U.S. account for almost 58.5% of total industry assets on June 30, 2014. The concentration ratio of largest six bank holding companies in U.S. was almost 58% and the Herfindahl Index is 705, which is considered a low level of concentration (Wu, 2015).

Supervisory responsibility

In Canada, the regulation and supervision of banks is the shared responsibility of the Department of Finance and other federal financial regulatory authorities, including the Bank of Canada, the Office of the Superintendent of Financial Institutions (OSFI) and the Canada Deposit Insurance Corporation (CDIC). However, to ensure consistency and to minimize competing regulatory objectives, the administration of the prudential regulation of Canadian financial institutions (banks and insurance companies alike) is under the jurisdiction of only one prudential regulator - The Office of the Superintendent of Financial Institutions (OSFI). Major structural reforms to banking sector regulations in the late 1980s also set the Canadian financial system apart from what exists in the US. Following the Canadian Government's 1987 deregulation bill, most of the country's large investment houses were bought by the big five commercial banks. The single regulator is empowered to regulate the whole entity. Consequently, the investment dealers have been subject to the same stringent rules as the commercial banks.

Banking is regulated at both the federal and state levels in the U.S. State-chartered banks are subject to the regulation of the state in which they are chartered. Thus, a bank's primary federal regulator could be the Federal Deposit Insurance Corporation, the Federal Reserve Board, or the Office of the Comptroller of the Currency. Thus, each subsidiary of a banking conglomerate might be subject to a different regulatory authority according to whether it was classed as an insurance company, investment bank, or commercial bank. Apart from the bank regulatory agencies, the US maintains separate securities, commodities, and insurance regulatory agencies at the federal and state level. This can create competing regulatory objectives.

Mortgage market

Nowhere does the regulatory differences manifest themselves more clearly than in mortgage finance markets. In Canada, banks cannot offer loans with less than 5 percent down payment, and the mortgage must be insured if the borrower puts less than 20 percent down. The legislation relating to mortgages requires that all high-ratio residential mortgages (currently defined as those having an initial down payment of less than 20% of the value of the property) made by banks be insured against default by either the government–owned Canada Mortgage and Housing Corporation (CMHC) or private insurers. Mortgage insurance is available, moreover, only if the household is total debt service is less than 40 percent of gross household income. These mortgage insurance providers are backed by the federal government, and are required to use conservative underwriting criteria.

Canadian banks also tend to hold on to mortgages rather than selling them to investors. Less than a third of Canadian mortgages were securitized before the financial crisis, compared to almost two-thirds of mortgages in the US. This feature of the Canadian financial system, combined with tight regulatory standards, gives Canadian banks stronger incentive to make those mortgages safe. The consequence was that prior to the recent financial crisis, not only did Canada have a much smaller housing boom than the U.S., but its mortgage delinquencies barely rose above the historical average of less than 1 percent. At its peak, 11 percent of American mortgages were more than 30 days overdue. These regulatory requirements curtailed the flourishing of the sub-prime market in Canada. Fewer than 3 percent of Canadian mortgages were classified as subprime before the crisis, compared to 15 percent in the U.S (Coyne, 2009). In addition, unlike in the United States, homeowners in Canada cannot reduce their federal or provincial taxes by the mortgage interest, as mortgage interest on residential properties is not deductible for tax purposes. The effect of these features of the mortgage market is that Canada has not seen a tax-driven distortion in the level of housing debt (Bones, 2009). This has contributed to the fact that Canada did not experience the same degree of housing boom and bust that occurred in many other countries in the aftermath of the financial crisis.

Regulatory capital requirements and leverage restrictions

While both US and Canadian banks adhere to Basel II requirement to hold minimum Tier 1 capital (defined as common shares, retained earnings and non-cumulative preferred shares to risk-adjusted assets) and Total capital ratios of 4% and 8% respectively, the OSFI requires Canadian banks to hold minimum requirements of 7% and 10% respectively. Nevertheless, Canadian banks tend to hold substantially more capital above these minimum requirements as buffers. In fact, the average capital reserves (Tier 1 capital) for Canada's Big Six banks is 9.8% (Bones, 2009), which is several percentage points above the 7% required by Canada's federal bank regulator. Another very important regulatory factor supporting the safety and soundness of the Canadian banking system is the ceiling on leverage ratio – the ratio of total assets to capital. The leverage ratio of Canadian banks is capped at no more than 20x capital. While the leverage ratios at major Canadian banks have risen steadily in recent years, the ceiling has ensured that average leverage among the major banks has remained markedly lower (an average of 18) than comparable figures for major banks in the US and UK which had an average leverage ratio of over 25x and European banks with an average ratio of over 30x prior to the financial crisis (Bones, 2009).

In summary, the Canadian financial system is characterized by a high degree of concentration, a strong regulatory regime, stringent capital requirements for banks, federal supervision, strict mortgage market regulations, and a conservative appetite for risk. The conservative and entrenched regulatory structure helped the Canadian banks in weathering the recent financial crisis relatively well. It is in light of this resilience that the World Economic Forum in its annual Global Competitiveness Report, ranks Canada banking system as the soundest in the world.[2]

[2] The U.S. came in at No. 40, Switzerland was No. 16, and Germany and Britain ranked 39 and 44 respectively.

Unlike the US where many banks collapsed and/or had to be bailed out during the recent financial crisis, there was no bank collapse in Canada. In fact, the value of the Canadian banks has actually risen in relative terms since the financial crisis. Of the 10 largest banks in North America measured by assets, four are now Canadian; a decade ago, none of the Canadian banks was in the top 10 (Coyne 2009). Financial commentators and policy analysts have probed at the sources of the resilience of the financial system by posing questions such as: 'Why banking crises happen in America but not in Canada?' (John Kay, Financial Times, June 4, 2014), 'How Canadian bank defied the financial crisis' (Constantine Passaris and Peter Bessey, The Daily Gleaner, March 2012) and 'Why was Canada exempt from the financial crisis?' (Renee Haltom, Econ Focus, 2013) could be found in the regulatory regime. The stronger supervision and stronger bank capital oversight might have been effective at preventing banks from taking excessive risks before the crisis.

2.2 Hypothesis Development

There are, at least, three channels for market regulations affect bank risk taking. First, regulations may explicitly mandate the level of leverage a bank can have. Second, entry barriers resulting from regulations may directly or indirectly determine the number of banks and the level of competition or concentration in the industry. This in turn can affect the margins and charter value of banks and therefore their risk taking incentives. Third, it explicitly restricts the operations of banks in certain segments of the market (e.g. investment banking, insurance, leverage, and the mortgage market) and this regulatory restrictions can affect the risk taking behavior of the banks. Regulations influence bank risk taking through their potential effect on bank charter value (Gonzales, 2005). Regulatory regime determines the structure of the banking system (concentrated or not) which in turn can affect the level of competition and the charter value. A regulatory regime that leads to a more concentrated banking structure and a positive effect on bank charter value will provide incentives for banks to institute conservative investment policies. On the other hand, a regulatory regime that creates less concentrated banking system with lower charter value will induce banks to take higher risk. In fact, prior studies indicate an inverse relationship between bank charter value (which itself is affected by the level of competition) and risk taking. Keeley (1990), Demsetz et al. (1996) and Anderson and Frazer (2000) show that high charter value reduces bank risk taking incentives and vice versa. Gropp and Vesala (2001) also document a negative relationship between charter value and bank risk taking for EU banks and by Konishi and Yasuda (2004) for Japanese banks. A bank with a high charter value has strong incentive to avoid high-risk choices that my trigger a drop in its charter value.

The comparative review of the banking regulations of the U.S. and Canada have over the years adopted different regulatory systems have an important influence on banks' risk taking behavior and produce a testable hypothesis about risk taking by banks in the two countries. To reiterate, the banking structures in the US can be described as less concentrated (more

competitive) and less restrictive banking in the US whereas that of Canada can be characterized as being more concentrated (less competitive) and more restrictive banking system. The differences between the U.S. and Canadian banking regulations (reflected in lower market concentration and less market power, which leads to higher level of competition among US banks and one that emphasizes economic scale and concentrated banking system in Canada) generate different expectations concerning risk taking of banks that operate in these regimes, and the differences produce a testable hypothesis about risk taking by banks in the two countries. The lower bank concentration and higher level of competition in the US can affect the charter value of the banks, which in turn can induce higher risk taking. On the other hand, banks in the highly concentrated Canadian market would command more market power, earn more revenues and as a result, they would have limited incentives to pursue more profits. This will reduce the incentives to take higher risks in Canada.

Also, Flannery (1998) and Hovakimian and Kane (2000) note that restrictions on bank activities are tools for reducing bank risk. Given the institutional features of the banking sector in the US and Canada, regulatory restrictions exist more in Canada and will affect incentives banks have to take on risk. Actually, regulatory restrictions can either reduce or increase risk. On the one hand, the greater banking freedom (as it exists) in the US can enable banks with low charter value (because of increased competition) to respond more to their high risk taking incentives and allow those banks to broaden diversification opportunities, which in turn can lead to a reduction in risk. On the other hand, the greater banking freedom, which can induce banks' entry into non-traditional areas, can increase the volatility of asset returns. Stronger regulatory restriction in Canada should affect the investment opportunity set of the banks and therefore risk-reducing effects from diversification of banks' asset portfolio will be limited. On the other hand, less banking freedom can curb banks' efforts to diversify into *non-traditional*areas, which in turn can limit the volatility of asset returns. The foregoing discussion suggests that the effects of regulation on bank risk taking can be ambiguous. A less concentrated and more competitive banking system, as exists in the US, would encourage greater risk taking (thus suggesting a negative relationship between degree of regulation and bank risk taking), but at the same time, the less restrictive nature of the regime will allow banks to diversify their asset portfolios and experience lower asset return volatility (thus suggesting a positive relationship between the degree of regulatory restriction and bank risk taking). In this study, we compare the risk taking behavior of Canadian and U.S. banks to investigate how these two factors affected the banking industry during the period leading up to the 2007-08 financial crisis.

3 DATA AND METHODOLOGY

3.1 Data

Our sample consists of the six major Canadian banks and a sample of eighteen US banks. The Canadian sample is comprised of Bank of Montreal (BMO), Canadian Imperial Bank of Commerce (CIBC), TD Bank Financial Services (TD), Royal Bank of Canada (RBC), Scotiabank,

and National Bank. This sample represents about 85 percent of bank assets and deposits in Canada. All these banks are chartered banks, i.e., commercial banks regulated by the Canadian Bank Act that run a range of activities including consumer and business loans, brokerage, investment dealing, and securitization. The US sample is comprised of 18 banks with assets between $100 billion and $800 billion and a diversified range of activities including commercial banking. [3] A list of sample banks is shown in the Appendix. Our sample period ranges from 1995 to 2008. This sample period allows us to examine the risk taking behavior of banks before the financial crisis. Bank specific data is obtained from Bankscope. Bank regulation and supervision survey data, as well as country wide macroeconomic variables from the World Bank. We employ accounting data and bank-specific data to measure risk. Our main measure of risk is the z-score, which we calculated using a five-year rolling window.

3.2 Measures of risk taking

We use both bank-specific and accounting-based measures of risk to examine the impact of regulatory regimes on risk-taking by Canadian and US banks. Our accounting-based measures are the z-score, ROA volatility, ROE volatility, the ratio of non-performing loans to gross loans, and solvency ratio. We use the z-score as our main measure of accounting-based risk and the other measures are used for robustness checks. The z-score is defined as the inverse of the probability of insolvency and is estimated as the return on assets plus the capital-to-asset ratio, divided by the standard deviation of return on assets. It measures the distance from insolvency (Roy, 1952). Following Laeven and Levine (2009), we define insolvency as a state where losses surmount equity (E<-π) (where E is equity and π is profits), ROA (=π/A) as return on assets, i.e net income divided by total assets, where A is total assets, σ(ROA) as the standard deviation of ROA, and CAR (= E/A) as the capital-asset ratio. The probability of insolvency can be expressed as prob (-ROA<CAR). If profits are normally distributed, then z equals (ROA+CAR)/σ(ROA), which is the inverse of the probability of insolvency. Thus, z indicates the number of standard deviations that a bank's ROA has to drop below its expected value before equity is depleted. A higher z-score indicates that the bank is more stable. For our analysis, we use the natural logarithm of the z-score, which is less skewed and follows the normal distribution.

We use a five-year moving window to calculate the volatility of ROA. To reduce the impact of outliers and to avoid spurious inferences due to extreme values, we winsorize the ROA series at -100% and +100%. We calculate ROE as net income divided by the book value of equity. Similar to ROA, the ROE series are winsorized at -100% and +100%, and we use a five-year moving window to calculate the volatility of ROE. The *ratio of non-performing loans* is the ratio of non-performing loans to gross loans. This ratio approximates a bank's exposure to credit risk. Barth et al. (2004) and Gonzalez (2005) use similar ratios to measure bank risk. A

[3] Our sample does not include investment banks such as Goldman Sachs, Merrill Lynch and Morgan Stanley since commercial banking does not constitute their main business activity.

higher ratio indicates a higher exposure to credit risk. *Solvency ratio* is defined as the ratio of shareholders' equity divided by total assets. A higher ratio indicates a decrease in the exposure to credit risk.

3.3 Measures of banking systems

We use several variables to capture the basic regulatory differences between the Canadian and the US banking systems as they relate to *structure*, *supervisory responsibility*, *mortgage market*, and *regulatory capital requirements and leverage restrictions*. We use the banking industry concentration ratio, which is available from the World Bank database, as a proxy for the difference in *structure* between the two countries. We use the number of regulators, which is 1 for Canada and 2 for the US as a proxy for the difference in *supervisory responsibility*. To capture differences in the *mortgage market* we use the ratio of non-performing loans over total loans. Our rationale is that more lenient mortgage requirements and higher incentives, through tax deductions for instance to increase mortgages would lead to higher delinquency rates and more non-performing loans. To proxy for differences in *regulatory capital requirements and leverage restrictions*, we use a ratio of debt multiplier as measured by long-term debt over capital.

3.4 Survey-based Measures of Financial regulation and supervision

To provide further insight into the differences between the Canadian and the US banking regulatory regime, we use data compiled by Barth, Caprio, and Levine (2001b) and updated by Barth, Caprio and Levine (2006, 2007) and currently available as part of the World Bank databases. The data represents a survey on how banks are regulated and supervised around the world, which is conducted and updated at several points in time for all countries subscribing to the Basel Accords. We use 2001, 2003, and 2007 surveys to construct indices on bank supervision and regulation for the US and Canada. First, we construct a comprehensive index (*Reg_ Index*) that evaluates the regulatory regime by assessing the stringency of the country's bank regulatory and supervisory authorities. We also construct three sub-indices that assess different aspects of the quality of bank regulation and supervision, namely, *Act_ Cap* the index of activities' restrictions, and capital and liquidity requirements; *Ent_ Mkt* the index of entry, market and ownership; and *Super* the index of supervisory power, all of which are expected to have an impact on banks' risk taking.

Most of the survey questions, which cover various aspects of regulation and supervision, require "yes" or "no" answers. Following Delis et al. (2011) and Kodongo (2016), for most questions, we assign a value of "1" for "yes" responses and a value of "0" for "no" responses and sum the values for each regulatory and supervisory aspect. Higher scores indicate higher regulatory stringency. Intuitively, more restrictive entry requirements, higher liquidity and capital requirements, higher supervisory and market powers, and higher restrictions on banks'

activities are expected to reduce banks' incentive to take on high risks. Our empirical framework allows us to test the effect of each of these aspects regulation on risk taking. The questions that are used to construct the composite index and each sub-index are shown in the Appendix.

3.5 Methodology

To understand the effects of the main features of banking system in the US and Canada (concentration, supervisory bodies and mortgage features on bank risk taking, we test our preliminary hypotheses by running the following regression:

$$Risk_{i,j,k} = \alpha_0 + \alpha_1 Con\_Ratio_{j,k} + \alpha_2 Numb\_\mathrm{Re}g_k + \alpha_3 NonPerfLoans_{i,j,k} + \alpha_4 DebtM_{i,j,k} + u_{i,j,k} \quad (1)$$

where $Risk_{i,j,k}$ is the *z-score* for bank i in year j and country k; $Con\_Ratio_{j,k}$ is the banking industry concentration ratio in year j and country k; $Numb\_\mathrm{Re}g_k$ is the number of prudential regulators of country k; $NonPerfLoans_{i,j,k}$ is the ratio of non-performing loans over gross loans for bank i in year j and country k; $DebtM_{i,j,k}$ is the ratio of long-term debt over capital. This regression would allow us to test how the general characteristics of the banking system affect banks' risk-taking.

In order to understand how specific aspects of regulation affect banks' risk-taking, we use the World Bank survey on bank regulation and supervision, to conduct a pooled cross-sectional time series regression analysis on banks' risk-taking. Using individual bank's data and the least squares estimation technique, we estimate the following model for the sample firms:

$$Risk_{i,j,k} = \alpha_0 + \alpha_1 \mathrm{Re}g_k + \alpha_2 \mathrm{Reg}\_Index_{j,k} + \alpha_3 \mathrm{Re}g_k * \mathrm{Reg}\_Index_{j,k} + \alpha_4 Size_{i,j,k}$$
$$+\alpha_5 Leverage_{i,j,k} + \alpha_6 ROA_{i,j,k} + \alpha_7 Int\_\mathrm{Rate}_{j,k} + \alpha_8 Risk\_\mathrm{Prem}_{j,k} + \alpha_9 GDPG_{j,k} + u_{i,j,k} \quad (2)$$

where $Risk_{i,j,k}$ is the *z-score* for bank i in year j and country k. $\mathrm{Re}g_k$ is country dummy variable equal to 1 for Canada (k=1) and 0 for the US (k=0). The inclusion of this variable allows us to capture differences in bank risk-taking by country. *Reg_ Index* is a composite index that summarizes regulatory restrictions and requirements in each country. The ratio helps us ascertain whether risk taking depends specifically on the country's regulatory environment. We also control for the impact of bank characteristics that might affect risk taking. These are *Size*, measured as the natural logarithm of the book value of bank assets; *Leverage*, is measured as the ratio of deposit and non-deposit liabilities to total assets; and profitability (*ROA)* is measured as return on assets. Bank stability is also affected by macroeconomic variables such as output growth, inflation, currency, real interest rates, and credit expansion (see for instance, Demirguc-Kunt and Detragiache, 1998). We control for three main macroeconomic variables, namely the equity risk premium, which is defined as the lending rate minus the Treasury bill rate, the real interest rate, and GDP growth rate. Robust standard errors are estimated using Petersen (2009) correction for firm clustering.

4 RESULTS

4.1 Preliminary results: Univariate analysis

	Canada			USA		
	Mean	Median	Standard deviation	Mean	Median	Standard deviation
Return on Assets (%)	0.639	0.684	0.129	1.154	1.131	1.722
Return on Equity (%)	13.419	13.708	2.513	14.502	15.121	5.080
Solvency Ratio (%)	4.762	4.728	0.225	8.083	8.364	3.440
Loan Loss Reserve over Gross Loans (%)	1.354	1.257	0.291	1.440	1.456	0.979
Non-Performing Loans over Gross Loans (%)	1.333	1.174	0.422	2.198	1.035	3.512
Loan Loss Provision over Interest Revenue (%)	13.285	12.467	0.341	17.268	16.479	9.321
Net Interest Margin (%)	2.066	2.049	0.102	3.341	3.527	2.010
Standard Deviation ROA (%)	0.147	0.120	0.081	0.701	0.367	1.115
Standard Deviation ROE (%)	3.157	2.469	1.698	6.085	4.136	5.574
Total Assets (10^8)	2.08	2.53	1.18	2.45	1.00	3.23
Z-Score	3.983	3.777	0.285	3.515	3.523	0.715
N	84	84	84	252	252	252

Table 1 Descriptive Statistics. This table shows descriptive statistics of the main variables over the sample period 1995-2008. Statistics are measured using annual data on the six major Canadian banks and eighteen US banks matched by size.

We estimate the mean (median) and difference in mean (median) of five risk measures for Canadian and US banks over our sample period and present the results in Table 2. The results show that the mean (median) *solvency ratio* and *non-performing loans over gross loans* are significantly lower for Canadian banks compared to US banks. We also observe that the mean (median) *standard deviation of ROE* and *standard deviation of ROA* are significantly lower for Canadian banks compared to US banks, which indicates that Canadian banks have lower volatility of returns compared to their US counterparts. The mean (median) *z-score* of Canadian banks is significantly higher than that of US banks, which indicates that Canadian banks are less risky than US banks.

	Canada		USA		Difference in Mean	Difference in Median
	Mean	Median	Mean	Median		
Solvency Ratio (%)	4.762	4.728	8.083	8.364	-3.321*** (4.074)	-3.636** [2.088]
Non-Performing Loans over Gross Loans (%)	1.333	1.174	2.198	1.035	-0.865* (0.907)	0.139*** [0.410]
Standard Deviation ROA (%)	0.147	0.120	0.701	0.367	-0.554** (-2.093)	-0.247*** [2.814]
Standard Deviation ROE (%)	3.157	2.469	6.085	4.136	-2.928** (2.002)	-1.667* [1.724]
Z-Score	3.983	3.777	3.515	3.523	0.468* (1.780)	0.254* [1.722]

Table 2 Difference in mean (median) tests. This table shows difference in mean (median) tests for Solvency ratio, non-performing loans over gross loans, standard deviation of ROA, standard deviation of ROE, and Z score over the sample period 1995-2008. Differences in mean (median) tests compare the Canadian banks mean (median) statistic to the US banks mean (median) statistic. Satterthwaite-Welch t-test statistics appear in parentheses and Wilcoxon, Mann-Whitney statistics appear in brackets. The symbols *, ** and ***indicate significance at the 0.10, 0.05 and 0.01 levels, respectively.

In summary, the preliminary analysis indicates lower capitalization of Canadian banks compared to US banks, but also lower risk as measured by *standard deviation of ROE*, *standard deviation of ROA*, and *z-score*. Since univariate tests are only suggestive and have no explanatory power, we use multivariate cross-sectional analysis to examine the difference in risk between Canadian and US banks and study the impact of the regulatory regime on bank's risk. The results of the multivariate cross-sectional analysis are presented in the next section.

4.2 Multivariate analysis

In this section, we run multivariate cross-sectional regressions using *z-score*, which is our main measure of risk, to examine the difference in risk between Canadian and US banks and how such difference relates to the country's regulatory regime. In order to avoid spurious associations and to better gauge the drivers of banks risk-taking, we conduct the analysis using different specifications.

Table 3 reports regression results of the first regression (1) which represents a preliminary test of the difference in banking system between the US and Canada and its impact on banks' risk-taking. Consistent with our expectations, the univariate regression results indicate that higher industry concentration, a lower number of regulators, and lower ratios of non-performing loans lead to lower bank risk-taking as measured by the z-score. The debt multiplier, however, does not have a significant impact on risk taking. Results of the multivariate specifications (except model (6)) are consistent with the univariate results and indicate that a regulatory

regime characterized by high concentration, a low number of regulators, and restrictions on the mortgage market activities leads to more stable and less risky banks.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
c	3.545*** (8.087)	4.545** (2.396)	3.983*** (3.375)	3.952*** (8.333)	4.458*** (8.005)	4.045*** (2.789)	5.719*** (6.954)
Con_Ratio	0.533*** (2.880)					0.145*** (2.672)	-0.668*** (-5.025)
Numb_Reg		-0.461*** (-4.894)			-0.213** (-1.961)		-0.467*** (-5.271)
NonPerfLoans			-0.077*** (-3.287)		-0.091*** (-6.447)	-0.092*** (-6.507)	-0.077*** (-6.053)
DebtM				-0.153 (-0.983)	-0.047 (-0.300)	-0021 (-0.136)	0.002 (0.224)
R squared	0.04	0.06	0.05	0.01	0.11	0.12	0.18
N	255	255	255	204	183	183	183

Table 3 Banking structure and risk taking: This table presents regression results of z-score on the banking industry concentration ratio (*Con_Ratio*), the number of prudential regulators (*Numb_Reg*), the ratio of non-performing loans over gross loans (*NonPerfLoans*), and the debt multiplier (*DebtM*). T-statistics appear in parentheses. The symbols *, ** and ***indicate significance at the 0.10, 0.05 and 0.01 levels, respectively.

To better understand how the regulatory regime affects banks' risk taking we use the World Bank survey on banking regulation and supervision to build different indices that would inform us on how different aspects of regulations might affect risk taking. We also use a dummy variable, Reg, which is equal to 1 for Canada and 0 for the US to capture the country regulatory regime effect. In the first specification, we examine whether risk-taking is affected by the country dummy variable, *Reg*. In the second regression we explicitly use a composite regulation index, *Reg_Ind*, constructed using the bank regulation and supervision survey from the World Bank. This composite index illustrates the difference in bank regulatory and supervisory power between Canada and the US based on the World Bank survey. In order to better gauge the impact of the country's regulatory regime on banks' risk, an interaction term *Reg*Reg_Ind* is introduced in regression (3). Firm characteristics and macroeconomic variables are added in regressions (4) and (5) to control for other drivers of banks' risk.

	(1)	(2)	(3)	(4)	(5)
c	3.622*** (3.651)	5.567*** (5.811)	-3.988*** (-2.572)	-2.058** (-2.283)	-2.277 (-1.616)
Reg	0.461*** (4.894)	0.215 (1.481)	3.900*** (3.062)	3.737*** (3.456)	2.406*** (2.662)
Reg_Index		-0.069** (-2.045)	1.309*** (2.854)	0.991*** (3.196)	0.574** (2.409)
*Reg*Reg_Index*			-1.388*** (-3.047)	-1.064*** (-3.450)	-0.595*** (-2.605)
Size				-0.209*** (-3.804)	-0.135*** (-4.167)
Leverage				0.455 (1.184)	0.421 (0.930)
ROA				-0.300 (-0.223)	0.421 (1.338)
Int_Rate					0.160** (2.018)
Risk_Prem					0.143*** (4.043)
GDPG					0.090** (2.166)
R squared	0.05	0.07	0.13	0.12	0.24
N	275	255	255	204	204

Table 4 Effects of the regulatory regime on banks' risk This table presents regression results of z-score on the country dummy (Reg), the regulatory composite index (Reg_Index), bank characteristics (size, leverage, ROA), and macroeconomic variables (interest rate, risk premium, and GDP growth). T-statistics appear in parentheses. The symbols *, ** and ***indicate significance at the 0.10, 0.05 and 0.01 levels, respectively.

When the *z-score* is regressed on the *Reg* dummy (regression 1), we observe that consistent with the univariate test *Reg* is positive and statistically significant, which indicates that Canadian banks are characterized by a higher z-score and therefore lower risk than their US counterparts. When *Reg_Index* is included in the regression (regression 2) *Reg* coefficient loses its significance, however *Reg_Index* coefficient is negative and statistically significant. This result seems to be puzzling but is better understood when the interaction variable, *Reg*Reg_Ind* is introduced in regression (3). Results of regression (3) indicate that the country's impact as measured by *Reg* is positive, which indicates that Canadian banks are less risky than US banks. The regulatory index, *Reg_Index,* impact is also positive, which indicates that higher regulation and supervisory power leads to an increase in z-score (decrease in risk). The interaction term, *Reg*Reg_Ind,* is negative and statistically significant, which indicates that overall the Canadian regulatory index, as constructed using the World Bank survey, is less stringent than its US counterpart. Our results remain qualitatively the same when firm specific and country-wide control variables are added in regressions (4) and (5).

4.3 Regulatory regime and risk: Index decomposition

To better gauge the relationship between banks' risk and the country's regulatory regime, we decompose the index, *Reg_ Index*, into three sub-indices that assess different aspects of the quality of banks' regulation and supervision, namely, *Act_ Cap*, an index that assesses banks' activity restrictions, and capital and liquidity requirements; *Ent_ Mkt,* an index that assesses ease of entry

into the banking market, market power and ownership; and *Super,* an index that measures supervisory power and discipline.

$$Risk_{i,j,k} = \alpha_0 + \alpha_1 \mathrm{Reg}_k + \alpha_2 Ent\_Mkt_{j,k} + \alpha_3 Act\_Cap_{j,k} + \alpha_4 Super_{j,k} + \alpha_5 \mathrm{Reg}_k * Ent\_Mkt_{j,k}$$
$$+\alpha_6 \mathrm{Reg}_k * Act\_Cap_{j,k} + \alpha_7 \mathrm{Reg}_k * Super_{j,k} + \alpha_8 Size_{i,j,k} + \alpha_9 Leverage_{i,j,k} + \alpha_{10} Int\_\mathrm{Rate}_{j,k}$$
$$+\alpha_{11} Risk\_\mathrm{Prem}_{j,k} + \alpha_{12} GDPG_{j,k} + u_{i,j,k} \quad (2)$$

Panel A of Table 5 presents descriptive statistics and difference in mean (median) tests between the US and Canada for each sub-index. John et al. (2000) illustrate that if bank regulation concentrates on bank capital ratios then it may be ineffective in controlling risk-taking if banks have high leverage ratios. Higher leverage combined with high asset risk indicates higher moral hazard. Flat deposit insurance system encourages excessive risk taking since premiums are not adjusted (see Merton, 1977). Risk transfer to the insurer occurs when banks exhibit higher risk exposures than the risk category on which the flat rate is based. The lower capital ratios for Canadian banks is consistent with Saunders and Wilson (1999) suggestion that high bank capital levels have been supplanted by increased bank consolidation and safety-net provisions. Size, as a measure of diversification (see Brewer, 1989, among others), is expected to be negatively related to risk.

Panel B of Table 5 presents the results of different specifications that examine the impact of each of these sub-indices on banks' risk taking. Regression (1) results indicate that, in general, higher market power, stringent entry regulations, and higher supervisory power do not necessarily lead to a reduction in risk as measured by z-score, whereas higher restrictions on banks' activity and capital requirements lead to a reduction in risk-taking. These somehow unexpected results can be better gauged when interaction variables are added in regressions (2), (3), and (4). The results of such interactions indicate that the regulatory characteristics of the Canadian banking sector lead to a reduction in risk taking by Canadian banks as shown by the positive and statistically significant coefficient on *Reg*Ent_ Mkt* and *Reg_ Super*. The negative coefficient on *Reg*Act_ Cap* interaction indicates that restrictions on bank activities relating to their involvement in securities, insurance, and real estate, and regulatory requirements on capital do not necessarily lead to lower bank risk. On the contrary, these restrictions lead to increase in risk, as they do not allow the banks to enjoy diversification benefits. Consistent with the caution made earlier concerning the results of the interaction of the composite regulatory index and the country dummy the strongly positive effect of the activity restriction sub-index could be driving the positive effect of the composite index in the z-score regression reported in Table 3. This might indicate that regulatory measures have little effect when capital is above

the regulatory minimum. The results remain qualitatively the same when control variables are added in regression 5-7.

Consistent with Barth et al. (2008) restricting bank activities does not reduce risk taking. Banks that are involved in a broad range of activities should find it easier to diversify their income and hence reduce their risk. Besides, fewer regulatory restrictions can increase the franchise value of banks and therefore increase incentives for more prudent behavior.

Panel A							
	Canada		USA		Difference in Mean	Difference in Median	
	Mean	Median	Mean	Median			
Ent_ Mkt	10.208	10	7.083	7	3.125*** (9.508)	3*** (8.763)	
Act_ Cap	4.333	4	8.625	9	-4.291*** (-8.737)	-5*** (7.549)	
Super	8.833	8	12.208	12	-3.375*** (-9.666)	-4*** (5.037)	
Leverage	0.240	0.265	0.291	0.290	-0.051*** (-3.094)	-0.025** (2.065)	
Panel B							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
c	7.127*** (5.730)	2.842*** (3.966)	2.615** (2.299)	10.747*** (4.570)	5.216*** (3.825)	1.237* (1.847)	7.997*** (3.544)
Reg	3.986*** (3.104)	2.460* (1.857)	7.656*** (5.777)	5.226*** (2.567)	1.581** (2.132)	4.013*** (3.237)	2.447 (1.256)
Ent_ Mkt	-0.691** (-2.336)	-1.309*** (-2.854)			-0.575* (-1.796)		
Act_ Cap	0.518*** (5.422)		0.735*** (5.522)			0.387*** (2.834)	
Super	-0.242*** (-2.576)			-0.572*** (-3.008)			-0.275* (-1.728)
*Reg*Ent_ Mkt*		0.701* (1.820)			0.401 (1.089)		
*Reg*Act_ Cap*			-0.941*** (-4.980)			-0.436*** (-2.498)	
*Reg*Super*				0.420*** (2.656)			0.230* (1.688)
size					-0.135*** (-2.803)	-0.110*** (-2.429)	-0.117*** (-2.518)

Leverage					0.421*	0.442	0.360
					(1.701)	(1.298)	(1.015)
ROA					2.018	2.051	2.615*
					(1.528)	(1.631)	(1.771)
Int_Rate					0.155***	0.170***	0.146**
					(2.688)	(3.227)	(2.504)
Risk_Prem					0.146***	0.083*	0.107**
					(3.806)	(1.895)	(2.407)
GDPG					0.089*	0.105**	0.126**
					(1.677)	(2.156)	(2.401)
R squared	0.23	0.13	0.20	0.17	0.24	0.25	0.24
N	255	255	255	255	204	204	204

Table 5 Bank regulation and supervision indices and banks' risk. This table presents regression results of z-score on the country dummy (*Reg*), the entry restrictions and market discipline variable (*Entry_Market*), the activities restrictions and capital requirements variable (*Act_Cap*), the supervisory power variable (Super), bank characteristics (size, leverage), and macroeconomic variables (interest rate, risk premium, and GDP growth). T-statistics appear in parentheses. The symbols *, ** and ***indicate significance at the 0.10, 0.05 and 0.01 levels, respectively.

4.4 Regulatory regime and risk: Sources of risk and z-score decomposition

Our multivariate regression results indicate that a country's regulatory regime has an impact on its banks risk taking as measured by the *z-score*. An improvement in the *z-score,* and thus a reduction in risk, can emanate from improvement in profitability (return on assets), a reduction in asset return volatility, and/or changes in the capital adequacy ratio. To determine the main drivers of risk we examine how the components of the z-score are affected by the regulatory regime by re-estimating some specifications of Model (3) using ROA, volatility of ROA, and CAR as dependent variables. ROA is calculated as net income divided by total assets. To reduce the impact of outliers and to avoid spurious inferences due to extreme values we winsorize the ROA series at -100% and +100%. We use a two-year moving window of quarterly ROAs to calculate the *volatility of ROA*. CAR is calculated as total equity divided by total assets. The results, reported in Table 5, indicate that ROA and the volatility of ROA are the channels through which the different aspects of regulation as measured by the three sub-indices affect risk. Specifically, the results show that the lower risk and the stability experienced by the Canadian regulatory regime emanate from increase in profitability and lower profit variability- product of the concentrated nature of the banking industry, which seems to emphasize scale instead of competition. This effect is however, attenuated by the restrictions on banking activities.

	ROA				*CAR*				*Standard deviation ROA*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
c	0.135***	0.321***	-0.060	-0.051	0.152	0.146	0.123*	-0.022	-0.021**	-0.081***	0.055***	0.015
	(3.939)	(4.031)	(-1.462)	(-0.776)	(1.668)	(0.804)	(1.738)	(-0.194)	(-2.528)	(-4.286)	(5.169)	(0.515)
Reg	0.139***	0.334***	0.062	0.054	-0.001	-0.101	-0.077	0.068	-0.041***	0.079***	-0.054***	-0.014
	(3.606)	(4.168)	(1.523)	(0.823)	(-0.177)	(-0.553)	(-1.086)	(0.605)	(-4.726)	(4.054)	(-5.113)	(-0.509)
Ent_ Mkt	-0.034***	-0.043***			-0.013***	-0.008			0.008***	0.012***		
	(-3.761)	(-3.820)			(-5.904)	(-0.301)			(3.587)	(4.568)		
Act_ Cap	0.004*		0.009*		-0.005		-0.004		-0.003***		-0.006***	
	(1.909)		(1.897)		(-1.311)		(-0.432)		(-3.998)		(-4.688)	
Super	0.007***			0.005	0.006**			0.009	0.000			-0.001
	(3.007)			(1.009)	(2.418)			(1.011)	(-0.170)			(-0.290)
*Reg*Ent_ Mkt*		0.045***				0.008				-0.012***		
		(3.973)				(0.308)				(-4.371)		
*Reg*Act_ Cap*			-0.008*				0.004				0.006***	
			(-1.700)				(0.457)				(4.775)	
*Reg*Super*				-0.005				-0.009				0.001
				(-0.938)				(-1.011)				(0.339)
R squared	0.10	0.08	0.11	0.05	0.05	0.02	0.05	0.07	0.17	0.16	0.18	0.17
N	268	268	268	268	268	268	268	268	268	268	268	268

Table 6 Bank regulation and supervision and sources of risk: z-score components. This table presents regression results of the components of the Z-score for the sample banks using ROA, CAR and Volatility of ROA as the dependent variables. Each dependent variable is regressed on the country dummy (Reg), the entry restrictions and market discipline variable (*Entry_ Market*), the activities restrictions and capital requirements variable (*Act_ Cap*), and the supervisory power variable (Super). T-statistics appear in parentheses. The symbols *, ** and ***indicate significance at the 0.10, 0.05 and 0.01 levels, respectively.

4.5 Further Analysis

4.5.1 Autocorrelation

By construction, the volatility of ROA and the volatility of ROE suffer from autocorrelation as we use a moving average to estimate the volatility. To reduce the impact of autocorrelation, we restrict our analysis to observations, which are two years apart and hence are less affected by the issue of autocorrelation. The results of this restricted regression are presented in Table 7. The results are consistent with our earlier findings and indicate that higher market power, stringent entry regulations, and higher supervisory power do not necessarily lead to a reduction in risk. The interaction effects and the control variables results are also consistent with earlier findings.

	(1)	(2)	(3)	(4)
c	4.703***	2.828	1.704	3.874***
	(4.822)	(0.752)	(1.054)	(6.515)
Reg	1.819***	3.902*	5.864***	9.925
	(3.325)	(1.803)	(4.214)	(1.344)
Ent_ Mkt	-0.164	0.126		
	(-1.111)	(1.239)		
Act_ Cap	0.491***		0.625***	
	(6.652)		(3.975)	
Super	-0.310***			-0.863***
	(-4.815)			(-5.573)
*Reg*Ent_ Mkt*		-0.363		
		(-0.587)		
*Reg*Act_ Cap*			-0.625***	
			(-3.975)	
*Reg*Super*				0.801***
				(5.591)
size		-0.058	-0.028	-0.003
		(-1.213)	(-0.628)	(-0.072)
Leverage		0.716**	0.969***	0.714**
		(2.044)	(3.089)	(2.212)
ROA		1.375	0.396	4.334*
		(0.613)	(0.176)	(1.876)
Int_ Rate		0.029	0.075	0.022
		(0.394)	(1.124)	(0.319)
Risk_ Prem		0.145***	0.036	0.017
		(2.747)	(0.748)	(0.414)
GDPG		0.153**	0.113**	0.188***
		(2.228)	(2.160)	(3.670)
R squared	0.23	0.15	0.22	0.25
N	255	182	182	182

Table 7 Autocorrelation and robustness check. This table presents regression results of z-score on the country dummy (*Reg*), the entry restrictions and market discipline variable (*Entry_ Market*), the activities restrictions and capital requirements variable (*Act_ Cap*), the supervisory power variable (Super), bank characteristics (size, leverage), and macroeconomic variables (interest rate, risk premium, and GDP growth). T-statistics appear in parentheses. The symbols *, ** and ***indicate significance at the 0.10, 0.05 and 0.01 levels, respectively.

4.5.2 Aggregate Analysis

To check the robustness of our results, we rerun some of the regressions using a country wide aggregate measure of the z-score. The results are reported in Table 8 and are consistent with the individual firms' results. They indicate that on aggregate the regulatory system has an impact on banks' risk taking behavior and that more stringent regulation on industry entry, stronger market power, and higher supervisory authority lead to lower risk taking by banks.

As a last robustness check, we use standard deviation of ROE as our measure of risk and the results remain qualitatively the same.

	Aggregate Z-score									Standard deviation ROE		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)				
c	-35.223** (-2.415)	5.615 (0.923)	-2.061 (-0.465)	13.613*** (3.808)	-3.279*** (-3.126)	10.928*** (2.904)	17.169*** (6.105)	11.284*** (4.727)	15.644*** (6.938)	0.046*** (9.691)	1.558*** (2.683)	-0.073*** (-3.447)
Reg	41.911*** (2.964)	9.599** (2.392)	3.769*** (3.628)	3.231 (0.709)	8.320*** (6.424)	5.411 (1.554)	0.703 (0.384)	2.458*** (3.208)	1.396 (0.861)	-0.020*** (-4.966)	-1.556 (-2.703) ***	-0.171*** (-13.191)
Reg_ Index	1.409*** (2.782)	0.360** (2.217)									-0.054*** (-2.607)	0.030*** (6.796)
*Reg*Reg_ Index*	-1.484*** (-2.957)	-0.382** (-2.388)									0.055 (2.680) ***	-0.017*** (-6.856)
Ent_ Mkt			-0.443** (-2.344)	-1.405*** (-2.782)			-0.251 (-1.665)					0.004 (1.608)
Act_ Cap			1.113*** (3.200)		0.817*** (6.646)			0.237*** (2.674)				
Super			-0.628*** (-3.312)			-0.587* (-1.930)			-0.152 (-1.143)			
*Reg*Ent_ Mkt*				0.797* (1.875)			0.187*** (1.812)					
*Reg*Act_ Cap*					-1.024*** (-5.349)			-0.263*** (-2.626)				
*Reg*Super*						0.435* (1.721)			0.137* (1.910)			
Size		-0.571*** (-4.187)					-0.570*** (-4.581)	-0.450*** (-4.169)	-0.464*** (-3.227)			
Leverage		-5.018*** (-2.549)					-5.011*** (-3.452)	-5.515*** (-4.590)	-6.460*** (-4.390)			

ROA		4.391* (1.970)					4.713* (1.863)	3.268 (1.028)	4.974 (0.717)			
Int_ Rate		0.091*** (3.290)					0.090*** (3.233)	0.101*** (3.788)	0.083*** (2.987)			
Risk_ Prem		0.058** (1.815)					0.057* (1.842)	0.031* (1.888)	0.051 (1.640)			
GDPG		0.036 (1.384)					0.039 (1.410)	0.045* (1.908)	0.062** (2.285)			
R Squared	0.36	0.63	0.71	0.38	0.53	0.29	0.63	0.64	0.62	0.03	0.14	0.20
N	28	28	28	28	28	28	28	28	28	275	275	275

Table 8 Aggregate analysis and robustness check. This table presents regression results of the aggregate z-score and standard deviation of ROE on the country dummy (*Reg*), the entry restrictions and market discipline variable (*Entry_ Market*), the activities restrictions and capital requirements variable (*Act_ Cap*), the supervisory power variable (Super), bank characteristics (size, leverage), and macroeconomic variables (interest rate, risk premium, and GDP growth). T-statistics appear in parentheses. The symbols *, ** and ***indicate significance at the 0.10, 0.05 and 0.01 levels, respectively.

5 Conclusion

Over the years, the U.S. and Canada have adopted different regulatory systems. The banking structures in the US can be described as less concentrated (more competitive) and less restrictive banking in the US whereas that of Canada can be characterized as being more concentrated (less competitive) and more restrictive banking system. In the wake of the worst financial crisis, the Canadian banking system was resilient and went through the crisis relatively unscathed. We examine the risk taking behavior of banks in the US and Canada prior to the recent financial crisis. We find that Canadian banks had higher z-score and therefore lower risk than their US counterparts did over the study period. Further analysis shows that entry restrictions, which create concentrated banking, structure, restrictions relating to capital, liquidity and activities, and strong supervisory power and discipline positively influence the z-score, suggesting that these factors constrain excessive risk taking by Canadian banks. We also decompose the z–score into its components and re-estimate our baseline regression with the view to identifying the source of the risk. We find that entry restrictions generate (and higher concentration generate higher profits and lower variability of asset returns for Canadian bank whiles restrictions on activities reduce profitability and increases variability in asset return; however, the former seems to overwhelm the effect of asset restriction, given the lower risk that we observe for the Canadian banks.

References

Anderson, R. C., and D. R. Fraser, 2000. Corporate control, bank risk taking, and the health of the banking industry, Journal of Banking and Finance 24, 1383-1398.

Barth, J.R., G. Caprio, and R. Levine, 2004. Bank supervision and regulation: what works best?"Journal of Financial Intermediation 13, 205-48.

Berger, A., L. Klapper, and R., Turk-Ariss, 2009. Bank competition and financial stability. Journal of Financial Services Research 35, 99-118.

Bones, A., 2009. Regulation and supervision of the Canadian financial system, Annual Meeting of the Financial Supervisory Authority of Iceland.

Boyd, J., and G., De Nicolo, 2005. The theory of bank risk-taking and competition revisited, Journal of Finance 60, 1329-1343.

Demsetz, R., S., and P. E. Strahan. 1997. Diversification, size, and risk at U.S. bank holding companies. *Journal of Money, Credit, and Banking* 29, 300-313.

Gueyie, J-.P. and V. S., Lai, 2003. Bank moral hazard and the introduction of official deposit insurance in Canada, International Review of Economics and Finance 12, 247-277.

Gonzalez Rodriguez, F., 2005, Bank regulation and risk-taking incentives: an international comparison of bank risk, Journal of Banking and Finance, 1153-1184.

Hellman, T., K., Murdock, and J., Stiglitz, 2000. Liberalization, moral hazard in banking and prudential regulation: Are capital requirements enough? The American Economic Review 90, 147-165.

Jones, A., (2009) Regulation and Supervision of the Canadian Financial System, Annual Meeting of the Financial Supervisory Authority of Iceland

Keeley, M. C., 1990. Deposit insurance, risk, and market power in banking, *American Economic Review* 80, 1183-1200

Laeven, L. and R. Levine, 2009. Bank governance, regulation and risk taking, Journal of Financial Economics 93, 259-275.

Jia, C., 2009. The effect of ownership on the prudential behavior of banks, the case of China, Journal of Banking and Finance 33, 77-87.

Mullins, H. M., 1993. Risk taking, managerial compensation and ownership structure: An empirical analysis, Working Paper (University of Oregon, Eugene, OR).

Passaris C., and P. Bessey, 2012, How Canadian banks defied the financial crisis, The Daily Gleaner, March 22.

Saunders, A., E. Strock, and N. G. Travlos, 1990, Ownership structure, deregulation and bank risk taking behavior, Journal of Finance 45, 643-654.

Wu, L., 2015, A Comparative Analysis of the Performance of Canadian and U.S. Banks from 2005 to 2013 Using The Stochastic Frontier Approach, MA Thesis, Dalhousie University

APPENDIX A: Regulatory Sub-indices

	Variable	Description and Sources
	Regulatory variables	
Act_Cap	Capital and Liquidity Requirements	This index is determined by adding 1 if the answer is yes to questions 1-6 (i.e., yes=1, no=0). The questions are: 1. Is the minimum capital-asset ratio requirement risk-weighted in line with Basel guidelines? 2. Does the ratio vary with a bank's credit risk? 3. Does the ratio vary with market risk? 4. Before minimum capital adequacy is determined, are these items deducted from the book value of capital? (a) market value of loan losses (b) unrealized securities losses (c) unrealized foreign exchange losses. 5. Are there guidelines for asset diversification? 6. Are the sources of funds to be used as capital verified by authorities?
	Activity Restrictions	This index is determined by adding 1 if the answer is yes to the following questions (i.e., yes=1, no=0): 1. Is the level of such activities restricted: (a) securities activities (b) insurance activities (c) real estate activities (d) bank ownership of nonfinancial firms 2. Are banks required to hold either liquidity reserves or any deposits at the central bank
Ent_Mkt	Entry and Ownership	This index is determined by adding 1 if the answer is yes to questions 1-3 (i.e., yes=1, no=0), and adding 1 if the answer is no for question 4 (i.e., yes=0, no=1). The questions are: 1. Is there a minimum capital entry requirement? 2. Is information on source of funds for capital required? 3. Is there a maximum percentage of capital that can be owned by single owner? 4. Can borrowed funds be used?
	Market Discipline	This index is determined by adding 1 if the answer is yes to questions 1-9 (i.e., yes=1, no=0). The questions are: 1. Does income statement contain accrued but unpaid interest/principal while loan is non-performing? 2. Are consolidated accounts covering bank and any non-bank financial subsidiaries required? 3. Are off-balance sheet items disclosed to public? 4. Must banks disclose risk management procedures to public? 5. Are directors legally liable for erroneous/misleading information? 6. Do regulations require credit ratings for commercial banks? 7. Is there an explicit deposit insurance scheme? 8. Is subordinated debt allowable (required) as part of capital? 9. Is an external audit compulsory?

Super	Supervisory Power	This index is determined by adding 1 if the answer is yes to questions 2-12 (i.e., yes=1, no=0), and adding 1 if the answer is no for question 1 (i.e., yes=0, no=1). The questions are: 1. Is there more than one supervisory body? 2. Are auditors legally required to report misconduct by managers/directors to supervisory agency? 3. Can legal action against external auditors be taken by supervisor for negligence? 4. Are off-balance sheet items disclosed to supervisors? 5. The number of onsite visits per year is at least one? 6. Is the supervisory agency head appointed by the Ministry of Finance? 7. Can the supervisory agency order directors/management to constitute provisions to cover actual/potential losses? 8. Can the supervisory agency suspend director's decision to distribute: (a) dividends (b) bonuses (c) management fees 9. Can the supervisory agency supercede bank shareholder rights and declare bank insolvent? 10. Does banking law allow supervisory agency to suspend some or all ownership rights of a problem bank? 11. Does the law establish pre-determined levels of solvency deterioration which forces automatic actions such as intervention? 12. Regarding bank restructuring & reorganization, can supervisory agency or any other govt. agency do the following: (a) supercede shareholder rights (b) remove and replace management (c) remove and replace directors

Appendix B: List of Banks

Canada	USA
Royal Bank of Canada Toronto-Dominion Bank Bank of Nova Scotia Bank of Montreal Canadian Imperial Bank of Commerce National Bank of Canada	Bank of New York Mellon Corporation Wells Fargo & Company SunTrust Banks, inc. PNC Financial Services Group, Inc. Bank of New York Company, Inc. Wachovia Corporation National City Corporation KeyCorp FleetBoston Financial Corporation State Street Corporation Fifth Third Bancorp Bank of America Corporation Washington Mutual Inc. Citigroup Inc. Capital One Financial Corporation US Bancorp BB&T Corporation Regions Financial Corporation

UNIVERSITÀ DEGLI STUDI DI PALERMO
dSEAS dipartimento scienze economiche aziendali e statistiche
department of economics business and statistics
Working Papers

ISSN 'in fase di assegnazione', volume I, 2017

Spunti per una rilettura della disciplina giuridica degli *internet service provider*

Ideas for a new reading of the law regulation of internet service providers

Fabrizio Piraino

Riassunto *Il saggio affronta la c.d. responsabilità degli internet service provider, regolata dagli artt. 12-14 dir. dir. 00/31, con specifico riguardo alla violazione del diritto d'autore. Il saggio mira a dimostrare che - a dispetto dell'opinione di gran lunga prevalente in Europa - la disciplina in esame non costituisce una fattispecie di responsabilità aquiliana, ma appronta la regolazione di una sfera di liceità d'azione a favore degli ISP. Come l'analisi della giurisprudenza della Corte di Giustizia dimostra, il rimedio primario contro le violazioni dei diritti commesse in rete è rappresentato dall'azione inibitoria, mentre il risarcimento del danno assolve al ruolo di rimedio soltanto secondario. Un tale rilievo conferma l'ipotesi che gli artt. 12-14 dir. 00/31 mirino in primo luogo a delineare una sfera di liceità a favore degli intermediari. Il saggio si chiude con una proposta di re-interpretazione delle disposizioni italiane di recepimento delle norme europee, ossia degli artt. 14-16 d. lgs. 70/2003, finalizzata ad allineare la disciplina nazionale alle previsioni europee, con specifico riguardo alle regole che governano la prestazione di hosting*

Parole chiave responsabilità ISP; Dir. 2000/31/EC; d. lgs. 70/2003 artt. 14, 15,16

Abstract The essay addresses the issue of the so called*internet service provider*'s liability under artt. 12-14 dir. 00/31, with specific regard to the violations of copyright. The study aims at demonstrating that European Law on ISP is not a law on tort, but regulates a sphere of lawful

Dipartimento di Scienze Economiche, Aziendali e Statistiche
Università degli Studi di Palermo
viale delle Scienze ed. 13, 90128
E-mail: fabrizio.piraino@unipa.it

action in favor of *internet service providers*. The analysis of the Court of Justice's case-law reveals that the primary remedy against offenses committed on the internet is an injunction, while damages are only a secondary relief. This confirms the hypothesis that artt. 12-14 dir. 00/31 draw the perimeter of the legitimate activity of the *internet service provider*. The essay ends with a re-interpretation of the Italian provisions set forth at artt. 14-16 d.l gs. 70/2003 in order to align them, with the European directive, especially with regard to the hosting performance rules.

Keywords ISP liability – tort law or not - Dir. 2000/31/EC; d. lgs. 70/2003 artt. 14, 15,16

SOMMARIO: 1. La fenomenologia delle prestazione dell'*internet service provider* e il quadro normativo: la posizione del problema – 2. La giurisprudenza della Corte di Giustizia e i rapporti tra diritto e tecnica – 3. Il quadro italiano – 4. Le incertezze sui presupposti dell'esclusione dell'hosting provider dal concorso nell'attività illecita dell'autore della violazione: l'infedele recepimento della dir. 00/31 in Italia – 5. La ridefinizione del sistema delle regole sulla liceità della condotta dell'hosting provider – 6. Segue. Il valore della segnalazione del presunto titolare del diritto violato – 7. L'abbandono della distinzione tra intermediari attivi e passivi e la "variabile tecnologica".

1. Questa è la storia estremamente curiosa di come delle disposizioni sulla delimitazione della sfera di liceità dell'azione degli *internet service provider* siano state lette, invece, come norme sulla responsabilità, alimentando così un dibattito che non sempre brilla per perspicuità e rigore. Se il civilista può offrire un contributo in questa materia[1], è auspicabile che esso si indi-

[1] La letteratura sul punto inizia a divenire cospicua: cfr., ex multis, M. Franzoni, La responsabilità del provider, in questa Rivista, 1997, p. 248 ss.; V. Zeno Zencovich, I rapporti tra responsabilità civile e responsabilità penale nelle comunicazioni in Internetriflessioni preliminari, in Dir. inf., 1999, p. 1050 ss.; F. Di Ciommo, Internet I Responsabilità civile, in Enc. giur. Treccani, VIII, Roma, 2001, p. 1 ss.; Id., Evoluzione tecnologica e regole di responsabilità civile, Napoli, 2003, passim, in part. p. 269 ss. L. Nivarra, La responsabilità degli intermediari, in questa Rivista, 2002, p. 307 ss.; G. Ponzanelli, Verso un diritto uniforme per la responsabilità degli *internet service provider*s?, in Danno e resp., 2002, p. 5 ss.; E. Tosi, Le responsabilità civili, in I problemi giuridici di internet. Dall'e-commerce all'e-business, a cura di E. Tosi, Milano, 2003, p. 516 ss.; Id., Le responsabilità civili dei prestatori di servizi della società dell'informazione, in La resp. civ., 2008, p. 197 ss.; Id., Responsabilità civile per il fatto illecito degli *internet service provider* tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca, social network e aggregatori di contenuti di terzi, in Digesto disc. priv., sez. civ., Agg., Torino, 2016, p. 688 ss.; G.M. Riccio, La responsabilità civile degli internet providers, Torino, 2002, passim, in part. 95 ss.; S. Sica, Le responsabilità civili, in Commercio elettronico e servizi della società dell'informazione, a cura di E. Tosi, Milano, 2003, p. 267 ss.; R. Bocchini, La responsabilità civile degli intermediari del commercio elettronico. Contributo allo studio dell'illecito plurisoggettivo permanente, Napoli, 2003, passim, in part. p. 123 ss.; Id., La responsabilità di Facebook per la mancata rimozione di contenuti illeciti, in Giur. it., 2017, c. 632 ss.; T. Pasquino, Servizi telematici e criteri di responsabilità, Milano, 2003, passim; A. Pierucci, La responsabilità del provider per i contenuti illeciti della Rete, in Riv. crit. dir. priv., 2003, p. 143 ss.; M. Gambini, La responsabilità civile dell'*internet service provider*, Napoli, 2006, passim, in part. p. 227 ss.; G. Facci, La responsabilità dei providers, in Commercio elettronico, a cura di C. Rossello-G. Finocchiaro-E. Tosi, in Tratt. dir. priv., diretto da M. Bessone, Torino, 2007, p. 233

rizzi verso il chiarimento dei termini del problema e verso la sua riformulazione secondo moduli più rispettosi delle categorie e dei dispositivi tecnici che governano la responsabilità civile. La c.d. responsabilità dell'*internet service provider*[2] rappresenta un capitolo, ancorché assai significativo, del più ampio tema della responsabilità ai tempi di Internet[3], all'interno del quale si è proposto di distinguere gli illeciti di Internet, gli illeciti contro Internet, e gli illeciti per mezzo di Internet[4]. L'opinione di gran lunga prevalente ravvisa nella disciplina giuridica degli *internet service provider* un complesso di regole di responsabilità volte ad escludere l'applicazione a questi ultimi di un regime troppo rigoroso, quale potrebbe essere quello di responsabilità oggettiva

ss.; M. De Cata, La responsabilità civile dell'*internet service provider*, Milano, 2010, passim, in part. p. 157 ss.; V. Franceschelli, Sul controllo preventivo del contenuto dei video immessi in rete e i provider. A proposito del caso Google/Vividown, in Riv. dir. ind., 2010, p. 347 ss.; A. Mantelero, La responsabilità degli intermediari di rete nella giurisprudenza italiana alla luce del modello statunitense e di quello comunitario, in Contratto impr. /Europa, 2010, p. 529 ss.; E. Falletti, Internet e diritto d'autore, in Digesto disc. priv., sez. civ., Agg. V, Torino, 2010, p. 797 ss.; P. Sammarco, La posizione dell'intermediario tra l'estraneità ai contenuti trasmessi e l'effettiva conoscenza dell'illecito: un'analisi comparata tra Spagna, Francia e regolamentazione comunitaria, in Dir. inf., 2011, p. 285 ss.; Al. di Majo, La responsabilità del provider tra prevenzione e rimozione, in Corr. giur., 2012, p. 551 ss.; R. D'Arrigo, Recenti sviluppi in tema di responsabilità degli *internet service provider*s, Milano, 2012, passim, in part. p. 18 ss.; A. Montanari, Contratto di AdWords e profili di responsabilità. Osservazioni a margine di Corte di giustizia 23 marzo 2010, cause riunite da C-236/08 a C-238/08, in Dir. comm. int., 2011, p. 524 ss.; Id., Prime impressioni sul caso SABAM c. Netlog NV: gli *internet service provider* e la tutela del diritto d'autore online, in Dir. comm. int., 2012, p. 1082 ss.; M. Ricolfi, Contraffazione di marchio e responsabilità degli *internet service provider*s, in Dir. ind., 2013, p. 237 ss.; G. Giannone Codiglione, Opere dell'ingegno e modelli di tutela. Regole proprietarie e soluzioni convenzionali, Torino, 2017, p. 185 ss.

[2] Per un'analisi comparata del tema cfr. Aa.Vv., *Secondary Liability of Internet Service Providers*, a cura di G.B. Dinwoodie, Cham, 2017, *passim*.

[3] Al tema ha dedicato uno studio apposito Di Ciommo, *Evoluzione tecnologica e regole di responsabilità civile*, cit., p. 165 ss.

[4] Cfr. S. Magni-S.M. Spolidoro, La responsabilità degli operatori in Internet: profili interni e internazionali, in Dir. inf., 1997, p. 61 ss.; Zeno Zencovich, I rapporti tra responsabilità civile e responsabilità penale nelle comunicazioni in Internetriflessioni preliminari, cit., p. 1053 ss.; De Cata, La responsabilità civile dell'Internet service provider, cit., p. 29 ss. Si è soliti includere tra gli illeciti di Internet quelli commessi dai soggetti che a vario titolo consentono all'accesso alla rete, ne definiscono i protocolli, attribuiscono gli indirizzi IP e tra di essi vanno inclusi anche quelli perpetrati dai providers nei confronti degli utilizzatori, quali i comportamenti discriminatori o la fissazione di tariffe eccessive o sottocosto. Non di rado questi illeciti rientrano nel diritto antitrust o nell'ambito delle fattispecie di concorrenza sleale. Si designano, invece, illeciti contro Internet quelle condotte che danneggiano la rete e i suoi operatori e vi rientrano: le violazioni attuate mediante la propagazione di virus, di worms, di spywares; il "bombardamento" di un server per determinarne la paralisi; gli atti di pirateria informaticahacking finalizzati a compiere accessi illeciti, alterazioni e distruzione di dati custoditi o anche soltanto veicolati nelle rete; violazione delle norme sulle firme elettroniche etc. In latri termini, si tratta di illeciti che si compiono e si strutturano interamente all'interno della rete e anzi questo profilo integra un loro elemento costitutivo. Da questi ultimi vanno, quindi, distinti gli illeciti per mezzo di Internet, i quali sono comportamenti caratterizzati dalla violazioni di altrui prerogative protette e attuabili anche al di fuori di Internet, di cui quest'ultimo diviene, quindi, uno strumento che ne facilità il compimento o ne incrementa il tasso di verificazione. Vanno inclusi in questa categoria le violazioni dei diritti della persona, del diritto d'autore, del diritto industriale, delle norme a tutela della concorrenza.

per attività d'impresa oppure quello para-oggettivo della responsabilità per attività pericolose[5], così da scongiurare il rischio di disincentivare gli operatori economici dal prestare i servizi di intermediazione, assolutamente essenziali per il funzionamento della rete, oppure di sospingerli a investire in mercati diversi da quello europeo perché più allettanti anche grazie al trattamento giuridico meno severo[6]. Non deve destare stupore, quindi, se una delle piegature più frequenti del discorso sulla responsabilità degli intermediari sia quella dell'analisi economica del diritto: il criterio degli incentivi e dei disincentivi e la logica dell'underdeterrence e dell'overdeterrence campeggiano in molte analisi, specie nordamericane[7], e una delle osservazioni più frequenti si appunta sulla concentrazione del mercato in un oligopolio destinata a essere prodotta da un regime troppo severo di responsabilità, che inevitabilmente porterebbe a espellere dal mercato i providers meno solidi dal punto di vista economico, come tale non in grado di sopportare i costi di un sistema capillare di prevenzione degli illeciti oppure di un allargamento del fronte dei risarcimenti dovuti[8]. I riflessi della creazione di un oligopolio si proietterebbero in varie direzioni: non soltanto in quella della qualità e della varietà dell'offerta, ma anche in quella della riduzione del pluralismo in rete sotto il profilo tanto della diminuzione degli spazi di esercizio della libertà di espressione del pensiero, di comunicazione e di informazione, quanto del rischio di censura collegato all'intensificazione delle operazioni di filtraggio dei contenuti[9]. Il timore fortemente avvertito è che esaltando la figura dei provider in sede di regolazione della rete si finisca inevitabilmente per approdare a un regime di responsabilità troppo severo: o perché fondato sulla colpa ma ancorato a doveri gravosi di setacciamento delle informazioni e dei contenuti trasmessi o memorizzati o perché incentrato sul meccanismo tipologico che imputa una determinata categoria di danni - in questo caso: quelli cagionati servendosi delle infrastrutture della rete - agli intermediari a titolo di responsabilità oggettiva in ragione della circostanza che essi occupano la posizione migliore per ridurre il tasso di verificazione di quel tipo di danni, investendo in prevenzione. In entrambi gli scenari è concreto il rischio di far gravare sugli intermediari una posizione di garanzia rispetto alle violazioni dei diritti patrimoniali e non patrimoniali compiute nella rete, se non addirittura di dare vita a un regime degli

[5] Così Nivarra, *La responsabilità degli intermediari*, cit., p. 312.

[6] Cfr., *ex multis*, R. Bocchini, *La responsabilità extracontrattuale del provider*, in *Manuale di diritto dell'informatica*, a cura di D. Valentino, Napoli, 2016, p. 540 ss.; Id., *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 634; Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., par. 4; Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 238; D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p.18 ss., in part. 20; Gambini, *La responsabilità civile dell'Internet* service provider, cit., p. 33; M. Tescaro, *La responsabilità dell'internet provider nel d.lgs. n. 70/2003*, in *La resp. civ.*, 2010, p. 167.

[7] M. Schruers, *The History and Economics of ISP Liability for Third party Content*, in *88 Virginia L.Rew.*2002, p. 205 ss..

[8] M. Bellia-G.A.M. Bellomo-M. Mazzoncini, *La responsabilità civile dell'*Internet Service Provider *per violazioni del diritto d'autore*, in *Dir. ind.*, 2012, pp. 352-354.

[9] Cfr. ancora Bellia-Bellomo-Mazzoncini, *La responsabilità civile dell'*Internet Service Provider *per violazioni del diritto d'autore*, cit., p. 353.

intermediari che ben si potrebbe battezzare "responsabilità giuridica dei sicofanti"[10]. Alcuni di questi timori sembrano francamente eccessivi, specie l'affermazione, divenuta una sorta di refrain, secondo cui alla base dell'introduzione di una disciplina specifica degli intermediari si agita l'esigenza di evitare l'applicazione a questi ultimi del regime rigoroso della responsabilità oggettiva d'impresa[11]: eccessivi per lo meno se si vuole rimanere fedeli alle coordinate dei sistemi giuridici di civil law, e certamente di quello italiano, che, nonostante il superamento del dogma jheringhiano ohne Schuld keine Haftung[12], esigono ciononostante una previsione normativa per l'instaurazione di una regola di responsabilità oggettiva, in conseguenza della natura tipologica di questa forma di imputazione dei danni, al limite estesa per via analogica a casi diversi fatti emergere dall'evoluzione dei rapporti sociali[13].

Il punto di vista offerto dalla riflessione sulla posizione giuridica degli intermediari nel diritto privato europeo porta un numero sempre crescente di osservatori ad allargare il fuoco al tema, ben più impegnativo, dell'opportunità di modernizzare il diritto d'autore, sul modello delle recenti esperienze canadese e inglese, al fine di allentare la logica proprietaria per venire incontro – con un atteggiamento giusrealistico, se non addirittura smaccatamente sociologico – al comune sentire degli utenti di internet sempre più insofferente alle bardature prodotte dall'impostazione tradizionale del diritto d'autore, per intenderci quella che si è sviluppata in

[10] È diffuso in letteratura il *caveat* sull'inopportunità di trasformare i *provider* nei guardiani della rete o in giudici dei conflitti che ivi sorgono. In maniera del tutto condivisibile Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 240 ritiene che scongiurare un tale approdo sia proprio la *ratio* principale della disciplina europea degli *Internet service provider*: «La disciplina europea si basa su di una scelta fondamentale: se l'ambiente *online* offre occasioni senza precedenti per la violazione di diritti privati e, fra di questi, dei diritti di proprietà intellettuale e di marchio, l'onere di operare come "poliziotti della rete" deve far capo ai titolari dei diritti violati e non può essere ribaltato sui *provider*».

[11] Cfr., per tutti, Trib. Roma, sez. spec. proprietà ind. e int., ord., 11 luglio 2011, in *Riv. dir. ind.*, 2012, II, p. 37 ss., in part. 41, il quale ritiene che «la limitazione di responsabilità introdotta a beneficio degli ISP è principalmente volta ad evitare l'introduzione di una nuova ipotesi di responsabilità oggettiva non legislativamente tipizzata o quantomeno l'ipotesi di una compartecipazione dei *providers* ai contenuti illeciti veicolati da tersi utilizzando il servizio di connettività da essi fornito».

[12] R. von Jhering, *Das Schuldmoment im römischen Privatrech*1867, in *Vermischte Schriften juristischen Inhalts*, Leipzig, 1879, p. 155 ss., in part. 163.

[13] C. Castronovo, *La nuova responsabilità civile*[3], Milano, 2006, p. 275 ss., in part. p. 347 ss., il quale osserva che la contrarietà all'estensione analogica delle fattispecie legislative di responsabilità oggettiva è alimentata dal pregiudizio che considera quest'ultima eccezionale rispetto alla responsabilità per colpa; mentre «la responsabilità oggettiva costituisce un modello alternativo di responsabilità, in quanto non si limita ad essere responsabilità nonostante l'assenza di colpa ma è responsabilità in base a un criterio di imputazione diverso, come in particolare dimostra la vicenda storica della responsabilità extracontrattuale nella quale solo in un contesto tutto impregnato dell'assioma "senza colpa nessuna responsabilità" la responsabilità oggettiva ha potuto essere pensata come eccezione alla colpa».

ambiente analogico[14]. Il tema è enorme e andrebbe affrontato con maggiore serietà, il che non è possibile in questa sede.

A sostegno della veduta della disciplina degli intermediari in termini di responsabilità e, per di più, di un regime di favore, si fa anche notare che la dottrina è solita evidenziare[15] che all'emergere di un nuovo scenario economico, qual è certamente il commercio elettronico, e al formarsi di nuovi mercati si avverte l'opportunità di privilegiare il criterio soggettivo di imputazione della colpa piuttosto che quello più rigoroso di natura oggettiva, particolarmente diffuso nel campo della responsabilità d'impresa, e ciò in quanto, instaurando un regime più mite, l'imputazione colposa finisce per favorire l'iniziativa economica nella fase di decollo dell'attività nei mercati nascenti[16]. Una delle funzioni principali della disciplina europea sugli intermediari di internet viene individuata proprio nel ritagliare un'area di irresponsabilità oltre la quale tornano ad operare le regole della responsabilità civile per colpa, giacché qui l'alternativa non è tra irresponsabilità e responsabilità per colpa, dato che nessuno dubita che una qualche forma di rispondenza debba gravare sui provider, ma tra responsabilità per colpa e responsabilità oggettiva[17]. E, d'altro canto, anche la dottrina penalistica che si è occupata del tema ha riconosciuto che la disciplina europea degli intermediari non delinea una fattispecie speciale di responsabilità, ma si preoccupa di limitare i presupposti di una responsabilità fondata sulle norme generali[18].

La valutazione della normativa sugli *internet service provider* come un segmento significativo della disciplina del mercato è senza alcun dubbio corretta e altrettanto si può dire del rilievo secondo cui uno degli obiettivi del legislatore europeo sia quello di affrancare gli intermediari dall'applicazione, magari per via analogica, di regimi di responsabilità particolarmente severi come quello del produttore per i danni provocati dai prodotti difettosi. E, tuttavia, la fondatezza di queste considerazioni non impone affatto di considerare la disciplina degli intermediari in internet come un particolare regime di responsabilità o, per meglio dire, come uno

[14] Cfr. A. Bertoni-M.L. Montagnani, *La modernizzazione del diritto d'autore e il ruolo degli intermediari internet quali propulsori delle attività creative in rete*, in *Dir. inf.*, 2015, p. 111 ss.; ma già Id., *Il ruolo degli intermediari Internet tra tutela del diritto d'autore e valorizzazione della creatività in rete*, in *Giur. comm.*, 2013, I, p. 537 ss.

[15] G. Ponzanelli, *La responsabilità civile. Profili di diritto comparato*, Bologna, 1992, p. 54.

[16] Bocchini, *La responsabilità civile degli intermediari del commercio elettronico*, cit., p. 9 ss.; Id., *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 636.

[17] Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 636.

[18] S. Seminara, *Internetdiritto penale*, in *Enc. dir.*, Annali VII, Milano, 2014, p. 592: «l'assenza di specifici precetti induce a ritenere che esse [le deroghe alla responsabilità previste dagli artt. 14 e ss. d.lgs. 9 aprile 2003, n. 70 in attuazione degli art. 12 e ss. dir. 00/31: *n.d.a.*] si allineano alle altre norme dell'ordinamento giuridico: lungi dall'introdurre nuovi obblighi o nuove forme di responsabilità, le disposizioni in esame mirano semplicemente a limitare i presupposti di una responsabilità fondata sulle norme "comuni" del diritto civile e penale, fungendo così da filtro selettivo sul piano della tipicità di queste ultime». Da tale premessa sistematica, S. ricava due conseguenze: *a* le disposizioni del d.lgs. 70/2003 non delineano alcun autonoma fattispecie incriminatrice a carico degli intermediari; *b* l'eventuale inapplicabilità delle deroghe del d.lgs. 70/2003 non determina di per sé la responsabilità degli intermediari, bensì la riespansione dei principî comuni, il che non esclude, quindi, che la responsabilità possa essere comunque esclusa per altra via.

speciale regime di esclusione della responsabilità. Come si tenterà di dimostrare nel prosieguo, la prospettiva della responsabilità è angusta perché non offre l'orizzonte di senso migliore per comprendere il trattamento giuridico degli *internet service provider*, favorendo anzi equivoci e distorsioni. Ciò non equivale ad affermare che le norme sugli intermediari non abbiano nulla a che vedere con la responsabilità: ce l'hanno eccome, ma in via secondaria, come conseguenza di una disciplina che mira in via principale a tracciare i contorni della sfera di liceità di azione degli intermediari piuttosto che a rimuovere pregiudizi. Non è, quindi, l'imputazione dei danni collegati all'attività dannosa posta in essere dagli utenti della rete la chiave di lettura migliore per costruire un coerente sistema di norme finalizzate a regolare l'attività degli *internet service provider*.

Va riconosciuto che un contributo notevole a intorbidire le acque intorno alla posizione degli *internet service provider* per violazione dei diritti in rete, specie dei diritti di proprietà intellettuale, proviene innanzitutto dal quadro normativo di riferimento. Com'è noto, la materia è stata oggetto di una disciplina europea di armonizzazione che ha preso le mosse, innanzitutto, dalla dir. 00/31/CE del 8 giugno 2000 su taluni aspetti giuridici dei servizi della società dell'informazione, in particolare il commercio elettronico, nel mercato interno, meglio nota come direttiva sul commercio elettronico[19]. La direttiva ha un contenuto variegato che abbraccia la disciplina dei contratti conclusi per via elettronica e le regole da applicare ai prestatori di servizi che operano come intermediari in Internet, il che rende assai bene la centralità riconosciuta nell'attuale economia dell'informazione agli *internet service provider* e la delicatezza del problema del bilanciamento tra l'interesse generale al potenziamento della rete e delle relazioni in essa intrecciate, preservando quanto più possibile le caratteristiche originarie di Internet come spazio di libertà di espressione, di comunicazione e di intrapresa economica[20], con l'interesse ad assicurare la salvaguardia dei diritti patrimoniali e no che online vedono acuite le possibilità di lesione. La delicatezza del contemperamento è accentuata dalla natura fondamentale degli interessi contrapposti, perché tali sono sia la libertà di espressione e di informazioneart. 11 Carta dei diritti fondamentali, sia la libertà di impresaart. 16 Carta dei diritti fondamentali, sia il diritto di proprietà intellettualeart. 17, comma 2, Carta dei diritti fondamentali, sia ancora il diritto alla riservatezzaart. 7 Carta dei diritti fondamentali e quello alla protezione dei dati personaliart. 8 Carta dei diritti fondamentali. Senza dimenticare che nell'inquadramento giuridico di Internet nell'ordinamento giuridico italiano assume centralità la libertà di comunicazione: sia quella tra soggetti determinati riconosciuta dall'art. 15 Cost., sia quella rivolta a soggetti indeterminati e fungibili sancita dall'art. 21 Cost.[21].

[19] Cfr. U. Draetta, *Internet e commercio elettronico*, Milano, 2001, *passim*.

[20] Sull'inquadramento costituzionale di Internet cfr. F. Donati, *Internetdiritto costituzionale*, in *Enc. dir.*, Annali VII, Milano, 2014, p. 532 ss.; P. Costanzo, *Internetdiritto pubblico*, in *Digesto disc. pubbl.*, Agg. I, Torino, 2000, p. 347 ss.; M. Orofino, *L'inquadramento costituzionale del web 2.0: da nuovo mezzo di per la libertà di espressione a presupposto per l'esercizio di una pluralità di diritti costituzionali*, in AAVV., *Da Internet ai Social Network*. Il diritto di ricevere e comunicare informazioni e idee, coord. da D. Diverio-M. Orofino, Santarcangelo di Romagna, 2013, p. 33 ss.

[21] Donati, *Internetdiritto costituzionale*, cit., p. 532 ss.

Per quanto concerne la posizione giuridica degli intermediari, ne va innanzitutto sottolineato il ruolo cruciale nel quadro del funzionamento della rete in quanto essi ne supportano le necessarie tecnologie, assolvendo a una funzione che, dal punto di vista sia tecnologico sia economico, risulta cruciale nello sviluppo delle relazioni online, e in particolare del commercio elettronico[22]. Tale ruolo di snodo imprescindibile nel sistema delle comunicazioni in rete influisce sull'inquadramento giuridico della loro posizione, specie nei confronti dei terzi lesi o pregiudicati nelle proprie posizioni sostanziali dall'azione illecita o abusiva di altri utenti della rete, tanto più alla luce della riconosciuta funzione di utilità sociale svolta da Internet[23], legata non più soltanto all'esplicazione della libertà fondamentale di comunicazione ma anche all'esercizio di molti altri diritti fondamentali[24], quali, ad es., i diritti politici o il diritto alla salute[25]. Non può, dunque, stupire che, nell'arduo tentativo di realizzare un equilibrio acconcio di tutti gli interessi coinvolti dalla diffusione di internet e dall'espansione della attività sociali e economiche che in tale ambiente sorgono e si sviluppano pressoché integralmente, la disciplina riservata agli intermediari venga giudicata sostanzialmente premiale[26].

A tal riguardo, il legislatore europeo ravvisa nella difformità dei diritti nazionali sul punto un fattore che ostacola il buon funzionamento del mercato interno, con particolare riguardo allo sviluppo dei servizi della società dell'informazione di carattere transnazionale e alla distorsione della concorrenza. In questo quadro, rientra anche l'obiettivo di contrastare le attività illecite svolte in rete mediante la predisposizione di «sistemi rapidi e affidabili idonei a rimuovere le informazioni illecite e a disabilitare l'accesso alle medesime»[27]. In difformità rispetto a premesse così tanto chiare sulla ratio dell'intervento normativo, la dir. 00/31 imbocca la via dell'inquadramento delle prestazioni degli *internet service provider* all'interno della logica della responsabilità. Ed è proprio questa scelta, verosimilmente influenzata dall'approccio al tema del Digital Millennium Copyright Act statunitense del 1998DMCA[28] – com'è noto incentrato sul sistema delle exemptions del § 512b2B, autentico architrave del sistema Mechanics of Sa-

[22] Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., par. 3.

[23] Cfr. G. Pascucci, *Internetdiritto civile*, in *Digesto disc. priv., sez. civ.*, Agg. I, Torino, 2000, p. 225 ss.; F. Di Ciommo, *Internet e crisi del diritto privato: tra globalizzazione, dematerializzazione e anonimato virtuale*, in *Riv. crit. dir. priv.*, 2003, p. 117 ss. Per una panoramica sulle principali questioni civilistiche sollevate da Internet cfr. Aa.Vv., *Internet e diritto civile*, a cura di C. Perlingieri-L. Ruggeri, Napoli, 2015, *passim* e con riferimento ai problemi di responsabilità p. 309 ss.

[24] Sul punto cfr. M. Betzu, *Regolare internet.* Le libertà fondamentali di informazione e di comunicazione nell'era digitale, Torino, 2012, *passim*, in part. p. 77 ss. e, per quanto concerne i problemi di tutela della proprietà intellettuale, p. 157 ss.

[25] Si allude al fenomeno della telemedicina.

[26] Nivarra, *La responsabilità degli intermediari*, cit., p. 311, p. 314.

[27] Considerando n. 40.

[28] Sul punto cfr. R. Petruso, *Responsabilità degli intermediari di internet e nuovi obblighi di conformazione:* robo-takedown, policy of termination, notice and take steps, in *Europa dir. priv.*, 2017, p. 451 ss.; ma v. già Id., *Fatto illecito degli intermediari tecnici della rete e diritto d'autore: un'indagine di diritto comparato*, *ivi*, 2012, p. 1175 ss.; Id., *La responsabilità degli* e-providers *nella prospettiva comparatistica*, *ivi*, 2011, p. 1107 ss.

fe Harbor[29] – che sta all'origine delle distorsioni di prospettiva che in larga misura affliggono l'impostazione del tema. La sezione 4 della dir. 00/31 è stata, infatti, intitolata alla responsabilità dei prestatori intermediari e i successivi artt. 12, 13 e 14 fissano le condizioni sotto le quali questi ultimi non possono essere chiamati a rispondere dell'eventuale illiceità delle informazioni trasmesse. Tali condizioni mutano a seconda della tipologia del servizio offerto: accesso alla rete e mero trasporto delle informazionimere conduit; memorizzazione automatica, intermedia e temporanea delle informazionicaching[30]; memorizzazione duratura delle informazioni fornite dal destinatario del serviziohosting. La sezione si chiude con l'esclusione da parte dell'art. 15 tanto di un obbligo generale di sorveglianza sulle informazioni trasmesse o memorizzate in capo all'intermediario quanto di un obbligo generale di quest'ultimo di ricercare attivamente fatti o circostanze che indichino la presenza di attività illecite. L'impostazione della dir. 00/31 è chiarita dal considerando 42, il quale ravvisa nelle previsioni degli artt. 12-14 altrettante "deroghe alla responsabilità" che si giustificano in presenza di attività degli intermediari circoscritte al compimento di un mero processo tecnico, automatico e passivo, che può consistere ora nell'attivazione o nella fornitura di accesso a una rete di comunicazione sulla quale sono trasmesse o temporaneamente memorizzate le informazioni trattate in maniera illecitamere conduit; ora nella memorizzazione temporanea di tali informazionicaching, anche laddove siano state compiute nel corso della trasmissione manipolazioni di carattere tecnico ove esse non alterino l'integrità dell'informazione; ora nella memorizzazione duraturahosting. La categoria fondamentale su cui si impernia la normativa europea è, dunque, quella dell'esclusione della responsabilità e l'impostazione della dir. 00/31 sembra ricalcare – lo si è anticipato - il concetto di Safe Harbor su cui ruota il § 512 DMCA. Come si avrà modo di chiarire nel prosieguo, questa prospettiva si presenta sotto più profili fallace, non foss'altro per la ragione che, se così fosse, allora ne dovrebbe conseguire che, in assenza delle condizioni di cui agli artt. 12-14 dir. 00/31, si dovrebbe dispiegare la responsabilità degli intermediari, il che non è affatto detto[31]. La disciplina europea non sembra ricalcare gli stilemi tipici degli elementi negativi della responsabilità civile, propri delle c.d. scriminanti o cause di giustificazione[32]. In altri termini, qui

[29] Per una valutazione critica della disciplina statunitense cfr. A. Hassanabadi, *Viacom v. YouTube* – All Eyes Blind. The Limits of the DMCA in a Web 2.0 World, in *26 Berkeley Tech. L.J.*2011, p. 405 ss., in part. p. 412 ss.

[30] È il tipo di prestazione offerta dall'intermediario quando consente ai clienti di costituire di *mailing list* e *newsgroup*: cfr. D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p. 26 ss., il quale reputa che i tre caratteri della memorizzazione delle informazioni, ossia di tipo automatico, intermedio e temporaneo, debbano concorrere. La prestazione di *caching* presuppone, quindi, che la memorizzazione delle informazioni avvenga senza l'intervento dell'operatore, sia funzionale esclusivamente al successivo inoltro ad altri destinatari a loro richiesta, rivesta carattere transitorio perché limitata al tempo tecnicamente necessario a tale inoltro. Nivarra, *La responsabilità degli intermediari*, cit., pp. 309-310 sottolinea come la prestazione di *caching* sia un utile espediente tecnico grazie al quale è possibile trasmettere il materiale richiesto senza dover ripassare dalla fonte originaria.

[31] Nello stesso senso v. Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 239.

[32] Sul punto cfr. Castronovo, *La nuova responsabilità civile*, cit., p. 17 ss. Mi sono occupato del tema in F. Piraino, *«Ingiustizia del danno» e antigiuridicità*, in *Europa dir. priv.*, 2005, p. 703 ss. Con specifico riferimento

non sembrano delineate fattispecie caratterizzate dalla sussistenza di tutti gli elementi della responsabilità civile ma nella quali essa non ricorre a causa della presenza di una circostanza di contesto che rende iniqua l'imputazione della responsabilità. Gli artt. 12 e ss. dir. 00/31 sembrano piuttosto tracciare delle sfere di liceità di azione a favore degli *internet service provider* che escludono in radice la possibilità di contestarne le relative condotte e di considerarli in qualunque modo compartecipi dell'illecito perpetrato in rete dagli utenti che si avvalgono dei loro servizi di intermediazione. Deve essere chiaro sin da subito infatti che, salva l'ipotesi del fatto illecito del content provider, al quale si applicano le comuni regole di responsabilità[33] per fatto proprio tipiche dell'autoria[34], la gran parte della casistica europea e nazionale riguarda la possibilità di prefigurare il concorso dell'intermediario nell'illecito commesso dall'utente. Ed è proprio questa la prospettiva adottata dagli artt. 12 e ss. dir. 00/31, ma il quadro normativo non sarebbe completo senza le norme europee sulla protezione del diritto d'autore e dei diritti di proprietà intellettuale, come d'altro canto chiarito dal considerando 50 della dir. 00/31. L'art. 3 dir. 01/29/CE, sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione, riconosce all'autore il diritto esclusivo di autorizzare o vietare qualsiasi comunicazione, su filo o senza filo, delle proprie opere, compresa la messa a disposizione del pubblico in maniera tale che ciascuno possa avervi accesso dal luogo e nel momento individualmente prescelti. L'art. 8, par. 1, dir. 01/29 impone agli Stati membri di prevedere adeguate sanzioni e mezzi di ricorso contro le violazioni dei diritti d'autore e, in particolare, di assicurare che le prime risultino efficaci, proporzionate e dissuasive; mentre la medesima disposizione, al par. 3, accorda agli autori il diritto di domandare un provvedimento inibitorio nei confronti degli intermediari i cui servizi siano utilizzati da terzi per violare un diritto d'autore o diritti connessi. E, infatti, il considerando 59 riconosce che in ambito digitale i servizi degli intermediari sono sempre più utilizzati da terzi per il compimento di attività illecite e che, in molti casi, proprio gli intermediari si rivelano i soggetti più idonei a porre fine agli illeciti. L'art. 3 dir. 04/48/CE, sul rispetto dei diritti di proprietà intellettuale, affida agli Stati membri l'individuazione di misure, procedure e mezzi di ricorso a tutela dei diritti di proprietà intellettuale leali, equi, non inutilmente complessi o costosi e tali da non comportare termini irragionevoli né ritardi ingiustificati, precisando altresì che essi debbano essere anche effettivi, proporzionati e dissuasivi, nonché applicati in maniera tale da scongiurare la creazione di ostacoli al commercio legittimo e da prevedere salvaguardie contro gli abusi. Inoltre l'art. 9 dir. 04/48 accorda al titolare del diritto di proprietà intellettuale, tanto nei confronti dell'autore

alla scriminante dello stato di necessità cfr. di recente L. Nonne, *Contributo ad una rilettura dell'art. 2045 c.c. Fattispecie e disciplina*, in *Giust. civ.*, 2017, p. 435 ss.; e anche Id., *Profili critici dello stato di necessità nel diritto privato*, in *Riv. dir. civ.*, 2017, p. 582 ss.

[33] In tal senso v. di recente Trib. Roma, sez. spec. in materia di impresa, 5 ottobre 2016, in *Dir. ind.*, 2017, p. 61 ss., con commento di R. Panetta, *La responsabilità civile degli* internet service provider *e la tutela del diritto d'autore.*

[34] Cfr. Nivarra, *La responsabilità degli intermediari*, cit., p. 307; Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., par. 5.

della violazione quanto nei confronti dell'intermediario i cui servizi siano utilizzati da terzi per compiere violazioni, un'ingiunzione interlocutoria finalizzata a prevenire la lesione imminente dei diritti di proprietà intellettuale o a vietarne la reiterazione, anche grazie a un presidio di pene pecuniarie per ogni successiva violazione. Infine l'art. 11 impone agli Stati membri di prevedere che l'autorità giudiziaria possa emettere nei confronti dell'autore della violazione della proprietà intellettuale un'ingiunzione diretta a vietare la prosecuzione dell'attività illecita e che tale ingiunzione possa essere indirizzata anche nei confronti degli intermediari i cui servizi sono utilizzati da terzi per violare un diritto di proprietà intellettuale.

Ancorché la protezione del diritto d'autore abbia conquistato la ribalta, non si deve dimenticare che in rete sono immesse, vengono ospitate e circolano informazioni di ogni genere, alcune delle quale espongono a pregiudizio profili della persona[35], e, pertanto, il quadro normativo sull'attività degli intermediari andrebbe completato con le disposizioni a tutela della persona con riferimento al trattamento dei dati personali contenute nelle dir. 95/46 CE e 02/58/CE, e ora ridefinite dal reg. 2016/679 del 27 aprile 2016, regolamento generale sulla protezione dei dati personali[36]. Non andrebbe trascurato, infatti, che l'imposizione agli *internet service provider* di un dovere generale e generalizzato di sorveglianza per prevenire o per contrastare le condotte illecite degli utenti finirebbe non soltanto per comprimere la libertà di informazione e di comunicazione dei fruitori di internet, ma anche per violare le regole poste a presidio delle informazioni di carattere personale su questi ultimi.

Dall'assetto normativo attuale emerge con una certa chiarezza che l'obiettivo del legislatore non sia tanto quello di delineare, sebbene in negativo, un complesso di regole di responsabilità civile per gli *internet service provider*, adottando la prospettiva, a dire il vero piuttosto insolita, della formalizzazione di presupposti dell'esonero dalla responsabilità, quanto piuttosto quello di fissare con certezza quella sfera di liceità di azione, nel cui ambito i servizi effettuati non espongono gli *internet service provider* ad alcun appunto di antigiuridicità rispetto agli eventuali illeciti commessi dagli utenti in termini sia di pura illiceità, che legittima il ricorso a misure

[35] Nivarra, *La responsabilità degli intermediari*, cit., pp. 307-308.

[36] Per un commento a prima lettura del Reg. 2016/679 cfr. F. Pizzetti, Privacy *e il diritto europeo alla protezione dei dati personali.* Dalla Direttiva 95/46 al nuovo Regolamento europeo, Torino, 2016, *passim*, in part. p. 147 ss.; C. Bistolfi-L. Bolognini-E. Pelino, *Il Regolamento Privacy europeo.* Commentario alla nuova disciplina sulla protezione dei dati personali, Milano, 2016, *passim*; Aa.Vv., *La nuova disciplina europea della* privacy, a cura di S. Sica-V. D'Antonio-G.M. Riccio, Padova, 2016, *passim*; ma anche G. Finocchiaro, *Introduzione al Regolamento europeo sulla protezione die dati personali*, in *Nuove leggi civ. comm.*, 2017, p. 1 ss.; A. Mantelero, *Responsabilità e rischio nel Reg. UE 2016/679*, in *Nuove leggi civ. comm.*, 2017, p. 144 ss.; M. Granieri, *Il trattamento di categorie particolari di dati personali nel Reg. UE 2016/679*, *ivi*, 2017, p. 165 ss.; A. Thiene, *Segretezza e riappropriazione di informazioni di carattere personale: riserbo e oblio nel nuovo Regolamento europeo*, *ivi*, 2017, p. 410 ss.; F. Piraino, *Il Regolamento generale sulla protezione dei dati personali e i diritti dell'interessato*, *ivi*, 2017, p. 369 ss.; M.G. Stanzione, *Il regolamento europeo sulla* privacy*: origini e ambito di applicazione*, in *Eur. e dir. priv.*, 2016, p. 1249 ss.; A. Ricci, *Sulla «funzione sociale» del diritto alla protezione dei dati personali*, in *Contr. e impr.*, 2017, 586 ss. Sul problema cruciale della protezione dell'individuo contro i *big data* cfr. A. Soro, Big data *e* privacy*. La nuova geografia dei poteri*, in *Osservatorio dir. civ. e comm.*, 2017; G. Palazzolo, *La banca dati e le sue implicazioni civilistiche in tema di cessione e deposito alla luce del reg.UE n. 2016/679*, in *Contr. impr.*, 2017, p. 613 ss.

inibitorie, sia di responsabilità, che fonda l'azione di risarcimento del danno[37]. D'altro canto, gli artt. 12 e ss. dir. 00/31 concorrono a comporre un tassello fondamentale della disciplina del mercato, quel tassello rappresentato dalla regolazione del mercato digitale, e tale direzione funzionale della direttiva sul commercio elettronico è dichiarata in forma programmatica nel considerando 40. Come in ogni intervento normativo che miri in via diretta e principale alla regolazione del mercato, anche qui si è innanzitutto perseguito l'obiettivo della certezza delle coordinate di comportamento che gli operatori devono o possono adottare, delineando regole di condotta relative all'attività negoziale dei professionisti che forniscono servizi della società dell'informazione e fissando un perimetro d'azione al cui interno gli intermediari tra i professionisti e i destinatari dei loro servizi possono liberamente operare senza essere coinvolti nelle conseguenze delle attività o delle comunicazioni effettuate dagli uni e dagli altri. Sia chiaro – e lo si è anticipato – il mercato si può regolare, e in effetti lo si regola, anche mediante norme sulla responsabilità, ma questa strategia predilige le valutazioni ex post, lasciando un campo di libertà di azione agli agenti ma esponendoli alle conseguenze dannose delle loro condotte. La disciplina europea degli *internet service provider* ha, però, prediletto un'altra via: quella della fissazione di coordinate d'azione che forniscono agli operatori del mercato prescrizioni di condotta ex ante e che, ovviamente, se violate, innescano meccanismi rimediali, nei quali è incluso il risarcimento del danno conseguente all'accertamento della responsabilità.

L'elemento che caratterizza le diverse sfere di immunità[38] dalle conseguenze giuridiche delle attività svolte dagli utenti di internet mediante i servizi e le infrastrutture offerte dagli intermediari e che le accomuna consiste nel carattere di mero collegamento dell'attività svolta da questi ultimi. L'intermediario che si limiti a fornire l'accesso alla rete, il servizio di trasporto, quello di memorizzazione automatica, intermedia e temporanea e finanche quello di memorizzazione duratura di informazioni, fornite però dal destinatario e rispetto alle quali l'intermediario resta estraneo, pone in essere una gamma di servizi caratterizzati da neutralità. In altri termini, in ipotesi siffatte l'intermediario opera come un diffusore passivo di contenuti informativi immessi o predisposti dagli utenti e ciò giustifica sia il regime di immunità sia l'esclusione dell'obbligo generale e generalizzato di sorveglianza sulle informazioni veicolate. L'assenza di qualsiasi interferenza nelle informazioni immesse dagli utenti nonché di significative manipola-

[37] Sebbene ravvisi nella disciplina degli artt. 12 ss. un complesso di regole di responsabilità, in linea, d'altronde, con l'opinione dominante, Montanari, *Prime impressioni sul caso SABAM c. Netlog NV: gli* Internet Service Provider *e la tutela del diritto d'autore* online, cit., pp. 1085-1086 correttamente evidenza che la normativa sugli intermediari si incentra sulla definizione dell'attività che il soggetto esercita piuttosto che sulla definizione del soggetto in sé.

[38] Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 239 precisa che si tratta comunque di sfere di immunità temporanee e, per di più, parziali e «tuttavia val probabilmente la pena impiegare lo stesso il termine un po' enfatico di "immunità" perché esso dopo tutto ben descrive il fine di aprire spazi di libertà a vantaggio degli operatori di rete».

zioni[39] rappresenta uno dei presupposti dell'architettura aperta e decentralizzata di Internet e solleva la questione fondamentale ed enorme della c.d. neutralità della rete[40].

Ovviamente – e lo si è anticipato – non tutte le prestazioni degli *internet service provider* rientrano nella categoria della diffusione passiva di informazioni fornite dagli utenti, poiché svariati servizi erogati nella rete presuppongono la comunicazione di informazioni e di contenuti predisposti direttamente dall'intermediariocontent provider. Vi sono poi servizi di intermediazione di incerta collocazione quali i motori di ricerca[41], i social network[42] e gli aggregatori di contenuti di terzi, il cui regime giuridico dipende dal contenuto concreto dell'attività svolta dall'intermediario. E proprio a tal fine si è elaborata la coppia ISP attivi e ISP passivi, la quale dovrebbe fornire alla giurisprudenza dei criteri di inquadramento di queste figure che corrono sul crinale[43].

I termini del problema che si è soliti indicare con l'espressione "responsabilità degli *internet service provider*" possono essere così sintetizzati: il legislatore europeo ha sì introdotto una disciplina di armonizzazione dei diritti nazionali relativa agli intermediari, ma si è limitato a prendere in considerazione le prestazioni di servizio di natura squisitamente neutrale, ossia quelle nelle quali l'intermediario si limita a fungere da diffusore passivo di contenuti altruimere conduit, caching, hosting. Escono, quindi, dal fuoco della disciplina europea, in maniera indiscutibile, le prestazioni dei content provider; mentre, in misura ben più problematica, quelle di motori di ricerca, social network e aggregatori di contenuti di terzi. La parte più significativa della giurisprudenza europea e nazionale sin qui formatasi in materia si è affaticata proprio sul terreno della riconduzione o meno di tali ultime prestazioni alla zona franca di immunità dalla compartecipazione agli illeciti commessi dagli utenti per i tramite dei servizi infrastrutturali forniti dal provider-diffusore passivo tracciata dagli artt. 12 ss. dir. 00/31. E la casistica svela che le violazioni più significative sulle quali è insorto contenzioso finalizzato a coinvolgere gli intermediari nell'attività illecita di coloro che a tal fine si sono serviti dei loro servizi riguardano i diritti di proprietà intellettuale e i segni distintivi. Ci si può legittimamente domandare quali

[39] In dottrinaD'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p. 22 si fa notare, infatti, che anche il *provider* c.d. intermedio svolge comunque un qual certo ruolo attivo nella gestione e nello smistamento delle comunicazioni in transito, senza che ciò comporti, però, una compartecipazione nella definizione dei relativi contenuti. Addirittura l'art. 13 dir. 00/31 prevede che nella prestazione di *caching* l'intermediario possa aggiornare le informazioni temporaneamente memorizzate nel modo ampiamente riconosciuto e utilizzato dalle imprese del settore senza essere esposto a contestazioni sull'illiceità della propria condotta.

[40] Donati, *Internetdiritto costituzionale*, cit., p. 535 ss.; nonché, in una prospettiva più specifica, R. Bocchini, *La centralità della qualità del servizio nel dibattito in tema di* network neutrality.

[41] Sul punto v. ora G. Giannone Codiglione, *I motori di ricerca*, in questa *Rivista*, 2017, in corso di pubblicazione, letto in anteprima grazie alla cortesia dell'autore.

[42] In tema, in termini più generali, cfr. C. Perlingieri, *Profili civilistici dei* social networks, Napoli, 2014, *passim*, in part. p. 66 ss. e, con specifico riguardo alla questione della responsabilità civile, cfr. G.M. Riccio, *Social networks e responsabilità civile*, in *Dir. inf.*, 2010, p. 859 ss.

[43] Cfr. Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., par. 3.

siano le ragioni che inducono specie i titolari di diritti di privativa su entità immateriali ad agire nei confronti degli intermediari in presenza di illeciti compiuti da utenti dei loro servizi. In primo luogo, va tenuto in considerazione il nodo dell'anonimato, che non di rado rende estremamente arduo individuare l'autore materiale della violazione del diritto di esclusiva sicché diviene giocoforza agire nei confronti dell'intermediario[44]: il problema si pone principalmente nel contesto delle imputazioni penali, ma non è certo estraneo ai giudizi civili[45]. In secondo luogo, non si possono trascurare le ragioni di strategia processuale che inducono ad agire non solo nei confronti dell'autore della violazione ma anche nei confronti dell'intermediario in quanto soggetto per lo più maggiormente solvibile, la deepest pocket, con l'obiettivo di assicurarsi i maggiori margini di tutela. Infine, non va sottovalutato l'atteggiamento recalcitrante che talvolta gli *internet service provider* oppongono alle richieste di cancellazione o di disabilitazione all'accesso delle informazioni che violano il contenuto dell'esclusiva del titolare del diritto di proprietà intellettuale oppure i diritti inviolabili della persona, il che sposta il contenzioso nei loro confronti: ora per ottenere il provvedimento inibitorio richiesto e rifiutato ora per domandare il risarcimento dei danni provocati dal perdurare della violazione ascrivibile all'inezia dell'intermediario.

2. Un ruolo cruciale nella definizione delle regole gravanti sull'intermediario, specie a fronte della violazione dei diritti di esclusiva degli autori, è giocato dalla giurisprudenza della Corte di giustizia, il che, d'altro canto, è ovvio in un sistema giuridico come quello europeo a impianto misto: in parte fondato sul diritto scritto, in parte judge made[46].

Col suo approccio pragmatico la Corte suggerisce un inquadramento della disciplina degli intermediari approssimativo, ma, per lo meno, non troppo distante dal vero. In Google France SARL, Google Inc. c. Louis Vuitton[47], meglio nota come caso AdWords, la Corte di giustizia ravvisa la ratio degli artt. 12 ss. dir. 00/31 nella limitazione delle «ipotesi in cui, conformemente al diritto nazionale applicabile in materia, può sorgere la responsabilità dei prestatori di servizi

[44] Sottolinea questo profilo D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p. 4 ss.

[45] Cfr. al riguardo l'impostazione al tema data da Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., pp. 237-238.

[46] Sia consentito il rinvio a F. Piraino, *L'inadempimento dello Stato all'obbligo di attuazione delle direttive europee e il problema del risarcimento del danno*, in *Europa dir. priv.*, 2012, p. 707 ss., in part. 715 ss.; ma v. anche C. Castronovo, *La codificazione*, in *Manuale dir. priv. eur.*, a cura di C. Castronovo-S. Mazzamuto, I, Milano, 2007, p. 182.

[47] Corte giust., Grande sezione, 23 marzo 2010, cause riunite da C-236/08 a C-238/08 – *Google France SARL, Google Inc.* c. *Louis Vuitton Malletier SA*C-236/08 e *Google France SARL* c. *Viaticum SA, Luteciel SARL*C-237/08 e *Google France SARL* c. *Centre national de recherche en relations humainesCNRRH SARL, Pierre-Alexis Thonet, Bruno Raboin, Tiger SARL*C-238/08, in *Giur. it.*, 2010, p. 1603 ss., con nota di M. Ricolfi, *Motori di ricerca,* link *sponsorizzati e diritto dei marchi: il caso* Google *di fronte alla Corte di giustizia*; in *Dir. inf.*, 2010, p. 707 ss., con nota di G. Spedicato, *La sottile linea di confine tra esclusiva sul segno e usi leciti: prime riflessioni sulla giurisprudenza comunitaria in materia di* keyword advertising; in *Dir. ind.*, 2010, p. 429 ss., con commento di M. Tavella-S. Bonavita, *La Corte di Giustizia sul caso "AdWords": tra normativa marchi e commercio elettronico*; in *Dir. comm. int.*, 2011, p. 507 ss., con nota di A. Montanari, *Contratto di AdWords e profili di responsabilità. Osservazioni a margine di Corte di giustizia 23 marzo 2010, cause riunite da C-236/08 a C-238/08*.

intermediari»[48]. La sentenza è interessante sotto più profili, primi tra tutti la portata del diritto di esclusiva legato al marchio e il perimetro degli usi leciti da parte di terzi, specie sotto forma dell'impiego di segni corrispondenti al marchio come parole chiave per la visualizzazione di link che rinviano a messaggi commerciali; ma, per limitarsi al problema che qui ci occupa, quello della posizione giuridica dell'internet provider, assume centralità l'orientamento assunto dalla corte sulla delimitazione dell'immunità garantita alla prestazione di hosting dall'art. 14 dir. 00/31. Quel che riveste maggior valore non è tanto la circoscrizione, oramai consolidata, della regola di esclusione della responsabilità alle prestazioni di memorizzazione delle informazioni fornite dal destinatario del servizio che si presentino meramente tecniche, automatiche e passive; quanto piuttosto la considerazione del servizio AdWords come non incompatibile con la fattispecie di hosting delineata dall'art. 14 dir. 00/31, e dunque con la relativa immunità. Com'è noto, AdWords è un servizio di posizionamento a pagamento che consente agli operatori economici interessati di ottenere la visualizzazione privilegiata[49] rispetto ai c.d. risultati naturali del motore di ricerca di un link pubblicitario, accompagnato da un breve messaggio commerciale, che indirizza verso il proprio sito e ciò grazie alla selezione da parte dell'inserzionista di una o più parole chiave o ideate da quest'ultimo o prescelte tra quelle suggerite da Google[50]. La visualizzazione privilegiata si determina quando una di queste parole chiave coincida con quella immessa da un qualunque utente di internet nel motore di ricerca gestito da Google[51]. Secondo la Corte di giustizia, la natura corrispettiva del servizio reso, la predisposizione da parte di Google delle modalità di pagamento o la fornitura di informazioni di ordine generale ai clienti non impediscono di includere il servizio AdWords nella fattispecie di hosting di cui all'art. 14 dir. 00/31 e la ragione risiede nella circostanza che il servizio viene erogato mediante software sviluppati da Google, e dunque in maniera automatica. Tali software si limitano, infatti, a trat-

[48] Corte giust., Grande sezione, 23 marzo 2010, causa riunite da C-236/08 a C-238/08, cit., punto 107, la quale prosegue osservando che «È pertanto nell'ambito di tale diritto nazionale che vanno ricercati i requisiti per accertare una siffatta responsabilità, fermo restando però che, ai sensi della sezione 4 di tale direttiva [la 00/31CE *n.d.a.*], talune fattispecie non possono dar luogo a una responsabilità dei prestatori dei servizi intermediari».

[49] La visualizzazione privilegiata dipende dalla collocazione dei link e dei messaggi commerciali, la quale è distinta dai risultati ottenuti mediante l'algoritmo generale di funzionamento del motore di ricerca e collocata in evidenza o in cima alla pagina su uno sfondo diverso da quello dei risultati c.d. naturali o sul lato destro.

[50] Per la descrizione del funzionamento del servizio AdWords e di talune significative condizioni del relativo contratto di servizio cfr. Montanari, *Contratto di AdWords e profili di responsabilità. Osservazioni a margine di Corte di giustizia 23 marzo 2010, cause riunite da C-236/08 a C-238/08*, cit., pp. 524-525, p. 529 ss.

[51] Il corrispettivo del servizio viene determinato in base al "prezzo massimo per click" che l'operatore si è dichiarato disposto a pagare al momento della conclusione del contratto di posizionamento e in ragione dei numero di click che il suo link ha ricevuto da parte degli utenti di internet. Nel caso non infrequente che una medesima parola chiave venga selezionata da più inserzionisti, l'ordine della visualizzazione viene determinato dal presso massimo per click pagato, dal numero di selezioni che il link ha ricevuto in precedenza e dalla qualità dell'annuncio come valutata da Google. L'inserzionista può in qualunque momento tentare di migliorare la propria posizione, innalzando il prezzo massimo per click o migliorando la qualità del proprio annuncio commerciale.

tare i dati inseriti dagli inserzionisti[52], il che non consente di ritenere che la società conosca o controlli le informazioni e i contenuti inseriti da questi ultimi e memorizzati sul suo server[53]. L'unico spiraglio lasciato dalla Corte di giustizia al coinvolgimento di Google nella violazione dei diritti di esclusiva compiuta per il suo tramite dagli inserzionisti, oppure nella commissione di atti di concorrenza sleale, è legato all'eventualità – da accertare coso per caso – che Google abbia giocato un qualche ruolo attivo nella redazione del messaggio commerciale abbinato al link pubblicitario oppure nella determinazione e selezione delle parole chiave[54].

Vero e proprio leading case in materia di tutela del diritto d'autore contro le violazioni realizzate nella rete è la sentenza Scarlet Extended SA c. Société belge des auteurs, compositeurs et éditeurs SCRLSABAM[55] non tanto sotto il profilo dei confini della fattispecie di hosting di cui all'art. 14 dir. 00/31, non foss'altro perché la prestazione di intermediazione oggetto della controversia consisteva in un'attività di mero accesso ad internet, quanto piuttosto sul versante della portata dell'esclusione del dovere preventivo di sorveglianza sulle comunicazioni veicolate di cui all'art. 15 dir. 00/31, che - com'è noto - si applica a tutte e tre le categorie di prestazioni tipizzate dalla direttiva sul commercio elettronico. Il nodo della controversia da cui è originata la domanda di pregiudiziale europea sottoposta alla Corte di giustizia riguarda l'impiego da parte degli utenti di internet di software "peer to peer" che consentono la condivisione, mediante invio e ricezione, di file contenenti opere musicali, cinematografiche o audiovisive coperte da diritto d'autore. L'aspetto di gran lunga più significativo – ma piuttosto negletto nella di-

[52] Corte giust., Grande sezione, 23 marzo 2010, causa riunite da C-236/08 a C-238/08, cit., punti 115-117.

[53] Critico è il giudizio su questa parte della decisione di Montanari, *Contratto di AdWords e profili di responsabilità. Osservazioni a margine di Corte di giustizia 23 marzo 2010, cause riunite da C-236/08 a C-238/08*, cit., p. 535 ss., sulla base di più argomenti sostanzialmente riconducibili alla comune ragione che il servizio AdWords non può essere equiparato *quoad effectum* sul versante risarcitorio alla prestazione ordinaria che conduce ai c.d. risultati naturali. Tali argomenti sono: *a* la sussistenza di un obbligo di protezione a favore dei terzi titolari di diritti di proprietà intellettuale in capo a *Google*, il quale si estrinseca nella necessità giuridica di verificare la legittimazione degli inserzionisti a utilizzare segni corrispondenti ad altrui marchi; *b* nella violazione del contratto di pubblicità stipulato in precedenza con il titolare del marchio illecitamente utilizzato che determina il contratto di AdWords laddove sia consentito di selezionare come parole chiave simboli identici o simili al marchio senza aver prima verificato il possesso della licenza; *c* nella violazione dell'ordine pubblico economico che la selezione senza verifica potrebbe determinare, giacché le diposizioni a tutela della proprietà industriale concorrono in misura significativa a delineare l'affetto del sistema economico europeo. Sostanzialmente adesivo è invece il giudizio di Tavella-Bonavita, *La Corte di Giustizia sul caso "AdWords": tra normativa marchi e commercio elettronico*, cit., pp. 448-449, pur con non poche cautele.

[54] Corte giust., Grande sezione, 23 marzo 2010, causa riunite da C-236/08 a C-238/08, cit., punto 118 e da ciò la conclusione che «l'art. 14 della direttiva 2000/31 deve essere interpretato nel senso che la norma ivi contenuta si applica al prestatore di un servizio di posizionamento su Internet qualora detto prestatore non abbia svolto un ruolo attivo atto a conferirgli la conoscenza o il controllo dei dati memorizzati. Se non ha svolto un siffatto ruolo, detto prestatore non può essere ritenuto responsabile per i dati che egli ha memorizzato su richiesta di un inserzionista, salvo che, essendo venuto a conoscenza della natura illecita di tali dati o di attività di tale inserzionista, egli abbia omesso di prontamente rimuovere tali dati o disabilitare l'accesso agli stessi».

[55] Corte giust., Sez. III, 24 novembre 2011, C-70/10, *Scarlet Extended SA* c. *Société belge des auteurs, compositeurs et éditeurs SCRL*, in *Dir. inf.*, 2012, p. 260 ss.; e in *Gior. dir. amm.*, 2012, p. 632 ss., con commento di F. Melis, *La Corte di giustizia Ue pone limitazioni alla tutela del* copyright *sulla rete*.

scussione che si è animata a seguito dell'emanazione della sentenza – investe il rapporto tra il rimedio inibitorio, esperibile ai sensi degli artt. 8, n. 3, dir. 01/29 e 11, terzo periodo, dir. 04/48 in caso di violazione di un diritto di proprietà intellettuale, e le disposizioni sul commercio elettronico relative alla posizione giuridica degli intermediari, di cui agli art. 12 ss. dir. 00/31[56]. Fedele al proprio approccio destrutturato, la Corte di giustizia non ricava dal collegamento tra questi blocchi normativi ricadute di ordine sistematico-concettuale, le quali, tuttavia, appaiono evidenti nell'offrire una conferma di quanto angusta sia la pretesa di inquadrare la disciplina degli intermediari come un complesso di regole di responsabilità. Il respiro di tale normativa è, infatti, ben più ampio, influenzando anche i presupposti e la portata di altri rimedi come l'azione inibitoria o l'azione di arricchimento ingiustificato, il che suffraga la convinzione che si sia piuttosto al cospetto della perimetrazione normativa della sfera di libertà di azione riconosciuta agli *internet service provider*. Più nel particolare, la richiesta di imporre per via giudiziaria un ordine inibitorio nei confronti dell'intermediario che offre un mero servizio di accesso che preveda l'adozione di un sistema di filtraggio di tutte le comunicazioni elettroniche che transitano mercé i propri servizi, da applicare indistintamente a tutta la propria clientela, con finalità anche preventive, a spese dell'intermediario e senza alcuna limitazione di tempo è giudicata dalla Corte in contrasto sia con l'art. 15 dir. 00/31 sull'assenza di un dovere preventivo di sorveglianza, sia con la previsione dell'art. 3 dir. 04/48 secondo cui i rimedi a tutela della proprietà intellettuale devono atteggiarsi come equi, proporzionati e non eccessivamente costosi[57]. Un sistema di filtraggio dotato dei caratteri richiesti dalla SABAM richiederebbe, infatti, che l'intermediario identifichi all'interno delle comunicazioni dei suoi clienti quelle che appartengono al traffico "peer to peer", che inoltre rintracci all'interno di questo insieme i file che contengono opere coperte da diritti di proprietà intellettuale, che scrimini quali di tali file sono scambiati in maniera illecita e, infine, che proceda al blocco degli scambi che si sono rivelati illeciti[58]. Nulla di meno distante da un'attività di vigilanza preventiva, che, peraltro, comporterebbe la compressione di taluni diritti fondamentali dei clienti, per lo più attinenti alla sfera della persona, quali il diritto alla protezione dei dati personali e la libertà di comunicazione, riconosciuti rispettivamente dagli artt. 8 e 11 della Carta dei diritti fondamentali[59]. Né si può trascurare che un tale sistema di filtraggio attuerebbe, inoltre, un'eccessiva limitazione della libertà di impresa dell'intermediario a causa dell'obbligo di adottare un sistema informatico complesso, costoso, permanente e, per di più, interamente a suo carico, in chiaro dispregio dei requisiti di equità, proporzionalità e non eccessiva onerosità di cui all'art. 3, n. 1, dir. 04/48[60]. Contro l'adozione di un rimedio inibitorio dal contenuto siffatto milita anche una ragione, per così dire, di ordine generale: il rischio di compressione della libertà di informazione nella rete collegato alla con-

[56] Corte giust., 24 novembre 2011, C-70/10, cit., punti 30-35.

[57] Corte giust., Sez. III, 24 novembre 2011, C-70/10, punti 35 e 36 ma già in tal senso Corte giust., 12 luglio 2011, C-324/09, *L'Oréal SA, Lancóme parfums et beauté & Cie SNC, Laboratoire Garnier & Cie, L 'Oréal UK Ltd c. eBay International AG e altri*, in *Dir. comun. scambi int.*, 2011, p. 510 ss.

[58] Corte giust., 24 novembre 2011, C-70/10, cit., punto 38.

[59] Corte giust., 24 novembre 2011, C-70/10, cit., punto 50.

[60] Corte giust., 24 novembre 2011, C-70/10, cit., punto 48.

creta eventualità che il sistema di filtraggio non sia in grado di distinguere sempre in maniera puntuale i contenuti leciti da quelli illeciti e che, dunque, possa bloccare anche i primi, tanto più che la comunicazione potrebbe originare da limitazioni del diritto d'autore che variano da Stato a Stato oppure dalla libera accessibilità di talune opere decretata in taluni Stati oppure ancora dalla scelta degli autori di metterle in linea gratuitamente[61]. Sullo sfondo della scelta di giudicare incompatibile con il diritto europeo una misura inibitoria che imponga un sistema di filtraggio dei contenuti elettronici del tenore in precedenza descritto si staglia un'operazione di bilanciamento giudiziario che esclude l'assoluta prevalenza del diritto di proprietà intellettuale sugli altri diritti fondamentali coinvolti nel fenomeno delle comunicazioni elettroniche, incluse quelle con vocazione commerciale[62], in termini coerenti con quanto già riconosciuto dalla Corte di giustizia in materia di tutela del diritto di proprietà[63].

Alla medesima conclusione, e per di più sulla base di un'operazione di bilanciamento che ricalca del tutto quella della sentenza Scarlet Extended SA c. SABAM, la Corte di giustizia giunge anche nei confronti di gestore di un social network, riconducendo i relativi servizi nello schema della prestazione di hosting. La sentenza è quella pronunziata nell'ambito del caso SABAM c. Netlog, nel quale la medesima richiesta di un ordine inibitorio preordinato a realizzare un controllo generalizzato, preventivo e illimitato nel tempo per contrastare la violazione dei diritti di utilizzazione economica connessi al diritto d'autore commessa in rete viene proposta nei confronti di una società di gestione di una piattaforma di rete sociale in linea che concede a ogni iscritto uno spazio personale, denominato "profilo", sul quale l'utente può immettere i contenuti più vari, accessibili a livello mondiale, così da comunicare con gli altri iscritti in modo da creare una comunità virtuale[64]. In questo caso la condivisione illecita di file musicali e audiovisivi, senza l'autorizzazione dei relativi titolari del diritto d'autore e senza pagamento di un alcun corrispettivo, avviene tra gli iscritti alla rete sociale per il tramite dei propri profili. Nonostante qui l'intermediario eroghi un servizio di memorizzazione duraturo e non già di mero accesso alla rete, lo scenario non muta: la Corte di giustizia reputa contrario al diritto europeo un'ingiunzione che costringa il prestatore di servizi di hosting a predisporre un sistema di filtraggio che comporti la sorveglianza, nell'interesse dei titolari di diritti d'autore, della totalità o della maggior parte delle informazioni memorizzate presso i propri server, per un periodo di tempo indeterminato e per di più coll'obiettivo di prevenire le future violazioni tanto delle opere esistenti quanto di quelle non ancora create al momento in cui il sistema viene predisposto[65].

Dall'esame degli orientamenti della Corte di giustizia si ricava l'impressione complessiva secondo cui la questione centrale nel contenzioso sinora formatosi non consista tanto nel ricono-

[61] Corte giust., 24 novembre 2011, C-70/10, cit., punto 52.
[62] Corte giust., 24 novembre 2011, C-70/10, cit., punto 43.
[63] Corte giust., 29 gennaio 2008, C-275/06, *Promusicae*, punti 62-68
[64] Corte giust., Sez. III, 16 febbraio 2012, causa C-360/10, *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBASABAM* c. *Netlog NV*, in *Dir. ind.*, 2012, p. 341 ss., con commento di M. Bellia-G.A.M. Bellomo-M. Mazzoncini, *La responsabilità civile dell'*Internet Service Provider *per violazioni del diritto d'autore*.
[65] Corte giust., 16 febbraio 2012, causa C-360/10, cit., punto 45.

scere o meno la tutela del diritto d'autore in rete anche nei confronti degli intermediari, specie mediante il ricorso al rimedio inibitorio, tenuto conto, peraltro, che il diritto europeo sul punto si esprime in senso apertamente favorevole; quanto piuttosto nella determinazione del contenuto, della portata e della proporzionalità del rimedio richiesto. In altri termini, non è questione di an ma di quomodo della tutela e, infatti, il nocciolo del contenzioso sinora formatosi si esaurisce in una dinamica endogena al sistema dei rimedi, risolvendosi per linee interne nella questione di quale sia la calibratura del mezzo di tutela che consenta di contemperare la salvaguardia della proprietà intellettuale con la preservazione di spazi sufficienti alla libertà d'impresa dei prestatori di servizi intermedi e dei gestori di siti internet, alla libertà di comunicazione, di informazione e alla protezione dei dati personali dei fruitori della rete e, infine, al più generale e generico esercizio della libertà individuale, specie di quella di pensiero e di informazione, nella rete.

Una puntuale conferma è offerta da due più recenti sentenze della Corte di giustizia[66] nelle quali i rimedi richiesti sono stati ritenuti compatibili col diritto europeo proprio in quanto reputati proporzionati e praticabili. In UPC Telekabel Wien GmbH c. Constantin Film Verleih GmbH una società di produzione cinematografica contestava il contributo su larga scala che un fornitore di accesso alla rete offriva alla violazione dei propri diritti di esclusiva su opere audiovisive che i destinatari del servizio di accesso realizzavano scaricandole o visualizzandole da un sito internet gestito da terzi[67]. Il rimedio inibitorio invocato mirava alla disabilitazione dell'accesso al sito pirata e il giudice austriaco di secondo cure, riformando la sentenza di primo grado, ha modificato il provvedimento inibitorio concesso, imponendo all'intermediario un divieto generico, sotto forma di un obbligo di conseguimento del risultato dell'inibizione all'accesso degli utenti al sito pirata, lasciando però l'individuazione delle concrete misure tecniche da adottare alla discrezionalità del medesimo provider, senza tuttavia imporre un vincolo assoluto di risultato, grazie al riconoscimento all'intermediario della possibilità di esonerarsi dalle sanzioni dimostrando che il suo mancato conseguimento è avvenuto nonostante l'adozione di misure ragionevoli. Proprio il contenuto indeterminato e l'assenza di un vincolo assoluto di risultato sono valsi a convincere la Corte di giustizia della compatibilità di tale rimedio al diritto europeo in virtù della sua attitudine a realizzare un giusto bilanciamento di tutti i diritti e libertà fondamentali coinvolti nella vicenda. Inclusi anche quelli degli utenti impossibilitati ad accedere al sito contestato in considerazione del carattere necessariamente mirato della misura inibitoria, tale da dover scongiurare il rischio di interdire l'accesso anche a contenuti leciti, e in ragione inoltre della possibilità riconosciuta agli utenti di impugnare le misure adottate dall'intermediario dinanzi al giudice nazionale, anche sotto il profilo della loro proporzionalità. In Tobias Mc Fadden c. Sony Music Entertainment Germany GmbH la prestazione dell'intermediario era, ancora una volta, un servizio di accesso, che però rivestiva natura accessoria rispetto allo svolgimen-

[66] Corte giust., 27 marzo 2014, C-314/12, *UPC Telekabel Wien GmbH* c. *Constantin Film Verleih GmbH*, in *Int. Rev. of intellectual property and competition law*, 2014, p. 826 ss. e Corte giust., 15 settembre 2016, C-484/14, *Tobias Mc Fadden* c. *Sony Music Entertainment Germany GmbH*, in *Dir. comun. scambi int.*, 2016, p. 315 ss.

[67] Corte giust., 27 marzo 2014, C-314/12, cit.

to di un'attività principale di vendita di impianti di illuminazione e che si sostanziava nella fornitura di una rete locale Wifi gratuita, mediante la quale era stata immessa in rete anonimamente un'opera musicale senza l'autorizzazione dell'autore. La Corte di giustizia esclude la responsabilità dell'intermediario per la violazione dei diritto d'autore compiuta tramite la rete locale e, in pari tempo, reputa non applicabile in via analogica quella parte della disciplina della prestazione di hosting che impone all'intermediario di rimuovere immediatamente le informazioni e i contenuti illeciti non appena ne abbia conoscenza a causa della differenza tra le due prestazioni: di durata quella di hosting e istantanea quella di mere conduit. Per quanto concerne invece la possibilità di emettere un provvedimento inibitorio, la Corte esclude sia l'ammissibilità della chiusura della connessione internet, a causa della sproporzione della misura, sia l'esame di tutte le informazioni che transita sulla rete locale, per contrasto con l'art. 15 dir. 00/31; mentre riconosce la conformità al diritto europeo dell'imposizione all'intermediario della protezione della connessione a internet tramite l'apposizione di una password. Anche in questo caso, il rimedio viene approvato in forza della sua proporzionalità e dell'attitudine a realizzare un acconcio bilanciamento della libertà di impresa dell'intermediario e della libertà di informazione dei destinatari del servizio di accesso, consentendo al contempo l'effettiva protezione del diritto fondamentale di proprietà intellettuale grazie all'introduzione di uno strumento che consente di risalire all'autore della violazione del diritto d'autore.

Al di là del carattere più o meno condivisibile delle soluzioni offerte dalle due pronunzie[68], la giurisprudenza della Corte di giustizia restituisce una disciplina degli intermediari nella quale, a parte l'ipotesi a sé stante del content provider, il rimedio principale consiste nell'azione inibitoria, di cui è, però, necessario che l'autorità giudiziaria delimiti accuratamente il contenuto in modo da scongiurare che l'ordine giudiziale finisca per imporre un dovere generico di sorveglianza in contrasto con l'art. 15 dir. 00/31 oppure che esso si riveli iniquo, sproporzionato ed

[68] Ne dubita Petruso, *Responsabilità degli intermediari di internet e nuovi obblighi di conformazione:* robotakedown, policy of termination, notice and take steps, cit., p. 499 ss., il quale ravvisa in *Telekabel* c. *Constantin Film* e *Tobias Mc Fadden* c. *Sony* un mutamento di rotta rispetto alle precedenti pronunzie *Scarlet Extended* c. *SABAM* e *SABAM c. Netlog*. Più nel particolare, P. reputa che la misura inibitoria avallata in *Telekabel* c. *Constantin Film* sia tutt'altro che proporzionata, a nulla giovando né il contenuto indeterminato della misura con facoltà all'intermediario di stabilire i mezzi ragionevoli, né l'ammissione dell'esonero da responsabilità di quest'ultimo ove provi di avere adottato tutte le misure ragionevoli, né ancora la previsione da parte degli ordinamenti nazionali di strumenti rimediali a disposizione dei destinatari del servizio per impugnare la misura di limitazione di accesso a certi siti adottata unilateralmente dal *provider*. Specie sotto quest'ultimo profilo P. osserva che «il controllo giurisdizionale circa il rispetto del principio di proporzionalità sarebbe, infatti, soltanto eventuale ed opererebbe *ex post*, esponendo a responsabilità il provider nei confronti degli utenti della rete per le erronee valutazioni effettuate al riguardo e lasciando cadere sui singoli l'onere di impugnare dinanzi al giudice nazionale le misure cautelativamente adottate dal provider». Non meno sproporzionata pare a P. la misura inibitoria avallata da *Tobias Mc Fadden* c. *Sony* in quanto l'introduzione di una password non offre una garanzia assoluta di efficacia, ossia non assicura che ad essere esclusi siano proprio gli utenti che abbiano in animo di violare gli altrui diritti di privativa. Peraltro, il meccanismo della password grava l'intermediario non soltanto dell'identificazione degli utenti, ma anche dell'archiviazione dei loro dati e del monitoraggio degli scambi *peer to peer*, il che sembrerebbe sospingere verso quella sorveglianza e quella ricerca di fatti e di circostanze da cui emergano elementi di illiceità che, però, l'art. 15 dir. 00/31 espressamente esclude.

eccessivamente oneroso in violazione dell'art. 3, n. 1, dir. 04/48[69]. E, d'altro canto, i requisiti della proporzionalità e della praticabilità del rimedio e soprattutto della conoscenza del fatto lesivo, assunti dalla Corte di giustizia come presupposti imprescindibili dell'ammissibilità dei rimedi nei confronti dell'intermediario, si accordano massimamente proprio con la tutela inibitoria la quale, specie quando non si limita a ripristinare lo status quo ante ma mira anche a precludere in via prospettica la reiterazione dell'illecito, esige la prevedibilità delle condotte da contrastare, strettamente dipendente dalla conoscenza delle cause delle violazioni già perpetrate. La responsabilità si atteggia, invece, come un rimedio secondario in una duplice accezione: o nel senso che il risarcimento del danno opera come misura posta a presidio dell'esatta attuazione del rimedio primario inibitorio, in funzione, dunque, di deterrenza rispetto alla trasgressione o alla ritardata esecuzione dell'ordine giudiziale di eliminazione degli effetti della condotta illecita, affiancandosi così alle astreintes, ossia alle misure di coercizione indiretta, che - com'è noto - presentano una natura diversa da quella risarcitoria[70]; o nel senso che il risarcimento del danno subentra quando la misura inibitoria si è rivelata, o si rivela in chiave prognostica, inefficace, lasciando sul campo pregiudizi che il ripristino dell'integrità dei diritti esclusivi violati non è in grado di rimuovere, o comunque di minimizzare.

In controluce, nella giurisprudenza della Corte di giustizia si intravvede un certo favore a far ricadere sugli intermediari la gestione delle contestazioni formulate dai terzi per violazione dei propri diritti di proprietà intellettuale o per lesione dei diritti della persona. Una tale impressione è suscitata specialmente dalla sentenza Telekabel c. Constantin Film, la quale, avallando una misura inibitoria di limitazione di accesso a un sito pirata di contenuto indeterminato, la cui specificazione è rimessa alla discrezionalità dell'intermediario, grava quest'ultimo del peso maggiore nella strategia di contrasto degli illeciti in rete, sostanzialmente – anche se non espressamente – rimettendo al confronto tra intermediario e titolare della posizione sostanziale lesa e alla condivisione delle misure da adottare il buon esito della tutela inibitoria. Questa linea di tendenza forse non corrisponde a quel progressivo spostamento sugli intermediari dei costi transattivi legati al controllo diffuso dei contenuti che qualcuno paventa, mentre potrebbe più cautamente definirsi come quell'obiettivo di politica del diritto che mira a coinvolgere gli intermediari nell'enforcement del diritto d'autore e delle altre situazioni soggettive suscettibili

[69] Una puntuale definizione delle coordinate alle quali i diritti nazionali si devono attenere per delineare il contenuto del rimedio inibitorio si ha in Corte giust., 12 luglio 2011, C-324/09, *L 'Oréal SA, Lancóme parfums et beauté & Cie SNC, Labora-toire Garnier & Cie, L 'OréalUK Ltd c. eBay International AG e altri*, in *Dir. com. scamb. int.*, 2011, p. 510 ss.

[70] Sul punto cfr. il dibattito sull'art. 614-*bis* c.p.c. e, *ex multis*, soprattutto S. Mazzamuto, *La comminatoria di cui all'art. 614* bis *c.p.c. e il concetto di infungibilità processuale*, in *Europa dir. priv.*, 2009, p. 947 ss.; Id., *L'*astreinte *all'italiana si rinnova: la riforma della comminatoria di cui all'art. 614*-bis *c.p.c.*, *ivi*, 2016, p. 11 ss.; M. Bove, *La misura coercitiva di cui all'art. 614*-bis *c.p.c.*, in *Riv. trim. dir. proc. civ.*, 2010, p. 781 ss.; E. Zucconi Galli Fonseca, *Misure coercitive fra condanna e tutela esecutiva*, *ivi*, 2014, p. 389 ss.; S. Recchioni, *L'attuazione forzata indiretta dei comandi cautelare ex art 614*-bis *c.p.c.*, *ivi*, p. 1477 ss. Lo riconosce anche la Suprema corte, chiamata a valutare la compatibilità delle *astreintes* previste dal diritto belga con l'ordine pubblico italiano: Cass., 15 aprile 2015, n. 7613, in *Riv. dir. proc.*, 2016, p. 243 ss., con nota di V. Giugliano, *Compatibilità delle* astreintes *con l'ordine pubblico italiano*.

di lesione in rete. Una puntuale conferma di questo orientamento, per lo meno sul terreno del diritto d'autore, lo si ricava dalla proposta di direttiva europea sul diritto d'autore nel mercato unico digitale del 14 settembre 2016[71], la quale sollecita, per un verso, la conclusione di accordi tra i titolari di diritti d'autore e gli intermediari per l'utilizzazione in rete delle opere dei primi, specie quando la memorizzazione riguarda una grande quantità di opere o di materiali comunque coperti da diritti di privativa, e, per altro verso, la cooperazione tra titolari e intermediari sia all'esecuzione di tali accordi sia alla reazione agli illeciti identificati dai primi e comunicati ai secondi. L'art. 13, par. 1, della proposta stabilisce, infatti, che «I prestatori di servizi della società dell'informazione che memorizzano e danno pubblico accesso a grandi quantità di opere o altro materiale caricati dagli utenti adottano, in collaborazione con i titolari dei diritti, misure miranti a garantire il funzionamento degli accordi con essi conclusi per l'uso delle loro opere o altro materiale ovvero volte ad impedire che talune opere o altro materiale identificati dai titolari dei diritti mediante la collaborazione con gli stessi prestatori siano messi a disposizione sui loro servizi. Tali misure, quali l'uso di tecnologie efficaci per il riconoscimento dei contenuti, sono adeguate e proporzionate. I prestatori di servizi forniscono ai titolari dei diritti informazioni adeguate sul funzionamento e l'attivazione delle misure e, se del caso, riferiscono adeguatamente sul riconoscimento e l'utilizzo delle opere e altro materiale»[72].

3. L'ordinamento italiano ha attuato la dir. 00/31 con il d.lgs. 9 aprile 2003, n. 70, recependo la disciplina degli intermediari negli artt. 14 ss.[73], con rilevanza sia nel diritto civile sia nel diritto penale[74]. L'adeguamento del diritto italiano al diritto europeo ha arginato il formasi di un orientamento particolarmente severo nei confronti degli intermediari e dei gestori di siti, giacché fondato sull'equiparazione, da un lato, tra sito internet e organi di stampa e, dall'altro

[71] COM2016 593 final. Sui prodromi di questa proposta v. S. Sarracco, *Il fenomeno* e-commerce *e i recenti sviluppi del mercato unico digitale in Europa*, in *Riv. dir. ec. trasp. amb.*, 2016, p. 89 ss. Sulla proposta v. invece T. Shapiro, *EU Copyright Will Never Be the Same: A Comment on the Proposed Directive on Copyright for the Digital Single MarketDSM*, in *European intellectual property Rev.*, 2016, p. 771 ss.

[72] L'art. 13 prosegue prevedendo che «2. Gli Stati membri provvedono a che i prestatori di servizi di cui al paragrafo 1 istituiscano meccanismi di reclamo e ricorso da mettere a disposizione degli utenti in caso di controversie in merito all'applicazione delle misure di cui al paragrafo 1. 3. Gli Stati membri facilitano, se del caso, la collaborazione tra i prestatori di servizi della società dell'informazione e i titolari dei diritti tramite dialoghi fra i portatori di interessi, al fine di definire le migliori prassi, ad esempio l'uso di tecnologie adeguate e proporzionate per il riconoscimento dei contenuti, tenendo conto tra l'altro della natura dei servizi, della disponibilità delle tecnologie e della loro efficacia alla luce degli sviluppi tecnologici».

[73] Cfr. Aa.Vv., *Commercio elettronico*, a cura di C. Rossello-G. Finocchiaro-E. Tosi, in *Tratt. dir. priv.*, diretto da M. Bessone, Torino, 2007, *passim* e, per l'inquadramento generale v. C. Rossello, *La nuova disciplina del commercio elettronico. Principi generali e ambito di applicazione*, *ivi*, p. 109 ss.; F. Delfini, *Il commercio elettronico*, in *Tratt. dir. econ.*, diretto da E. Picozza-E. Gabrielli, Padova, 2004, *passim*.

[74] Seminara, *Internetdiritto penale*, cit., p. 592 non dubita che le deroghe della responsabilità previste dalla dir. 00/31 e recepire dal d.lgs. 70/2003 operino anche nella sfera penale e ciò già in astratto alla luce della sussidiarietà della responsabilità penale, che per necessità deve risultare più restrittiva di quella civile e, dunque, non può reprimere fatti considerati civilisticamente leciti. In tal senso milita anche il considerando n. 8 dal quale per S. emerge in maniera chiara che il legislatore europeo abbia inteso delineare un autonomo sottosistema per gli intermediari, con rilevanza civile, penale e amministrativa.

lato, tra intermediario ed editore[75]. Quanto la disciplina italiana abbia fedelmente attuato il vincolo di risultato contenuto nella dir. 00/31 sarà oggetto di uno specifico approfondimento nel prosieguo. Qui è sufficiente specificare che il d.lgs. 70/2003 ha riprodotto le tre tipologie di prestazione dell'intermediario: mere conduitart. 14, cachingart. 15 e hostingart. 16; nonché le condizioni sotto le quali l'erogazione di tali servizi non implica il coinvolgimento del prestatore negli illeciti per avventura perpetrati per il loro tramite[76]. Infine, l'art. 17 d.lgs. 70/2003 ha recepito la previsione dell'assenza in capo all'intermediario di un dovere generale di sorveglianza sulle informazioni che trasmette o memorizza e, a maggior ragione, di un dovere di ricercare in via autonoma fatti o circostanze dai quali emerga la presenza di attività illecite. Al contempo, l'art. 17 d.lgs. 70/2003 impone i doveri attivi di: a informare senza indugio l'autorità giudiziaria o quella amministrativa avente funzioni di vigilanza, ove a conoscenza di presunte attività o informazioni illecite compiute o immesse da un destinatario delle proprie prestazioni di servizio online; b e fornire senza indugio, su richiesta delle autorità competenti, le informazioni in proprio possesso in grado di consentire l'identificazione del destinatario dei propri servizi di

[75] Trib. Napoli, 8 agosto 1996, in *Dir. inf.*, 1997, p. 970 ss.; in *Resp. civ. prev.*, 1998, p. 176 ss., con nota di R. Sanzo, e in *Giust. civ.*, 1998, p. 261 ss., con nota di L. Albertini, *Le comunicazioni via Internet di fronte ai giudici: concorrenza sleale ed equiparabilità alle pubblicazioni a stampa*, il quale proprio sulla base dei presupposti *supra* nel testo ha ravvisato la responsabilità del *provider* per violazione dell'obbligo di vigilare sul compimento di atti di concorrenza sleale realizzati mediante la pubblicazione di messaggi pubblicitari in concorso con l'autore principale, come ipotesi quindi di *culpa in vigilando*. Trib. Macerata, 2 dicembre 1998, in *Dir. ind.*, 1999, p. 35 ss., con nota di C. Quaranta. Nel senso, invece, dell'assenza di un dovere assoluto di controllo e di vigilanza in capo al provider: cfr. Trib. Cuneo, ord., 23 giugno 1997, in *AIDA*, 1997, p. 942 ss.; Trib. Roma, ord., 4 luglio 1998, in *Dir. inf.*, 1998, p. 807 ss. e in *NGCC*, 1998, I, p. 492 ss.; Trib. Roma, 22 marzo 1999, in *Dir. inf.*, 2000, p. 66 ss., con nota di P. Sammarco, *Assegnazione dei nomi a dominio su internet, interferenze con il marchio, domain grabbing e responsabilità del provider*, pur avendo riconosciuto nel caso di specie la responsabilità del *provider* per non essersi avveduto di aver consento all'utente l'apertura di un sito che aveva come nome di domino un noto acronimo.

[76] Tali condizioni - com'è sin troppo noto - consistono nel caso di prestazione di *mere conduit* nel fatto che: *a* l'intermediario non dia origine alla trasmissione; *b* non selezioni il destinatario; *c* non selezioni né modifichi le informazioni trasmesse. Nella fattispecie di *caching*, le condizioni previste per godere dell'immunità sono: *a* che l'intermediario non modifichi le informazioni; *b* si conformi alle condizioni di accesso a queste ultime; *c* si conformi alle norme d aggiornamento delle informazioni, indicate in modo ampiamente riconosciuto e utilizzato dalle imprese del settore; *d* non interferisca con l'uso lecito di tecnologia ampiamente riconosciuta e utilizzata nel settore per ottenere dati sull'impiego delle informazioni; *e* agisca prontamente per rimuovere le informazioni che ha memorizzato, o per disabilitarne l'accesso, non appena venga effettivamente a conoscenza del fatto che le informazioni sono state rimosse dal luogo in cui di trovavano inizialmente sulla rete o che l'accesso alle informazioni è stato disabilitato, oppure che un organo giurisdizionale o un'autorità amministrativa ne ha disposto la rimozione o la disabilitazione. Nel caso degli *hosting provider* gli elementi che concorrono a integrare la fattispecie di immunità sono: *a* che l'intermediario non sia effettivamente a conoscenza del fatto che l'attività o l'informazione è illecita; *b* per quanto attiene ad azioni risarcitorie, non sia al corrente di fatti o di circostanze che rendono manifesta l'illiceità dell'attività o dell'informazione; *c* non appena a conoscenza di tali fatti, su comunicazione delle autorità competenti, agisca immediatamente per rimuovere le informazioni o per disabilitarne l'accessoart. 16 d.lgs. 70/2003.

memorizzazione dei dati, nel quadro delle attività di individuazione e contrasto delle attività illecite[77].

Nella giurisprudenza italiana, non meno di quanto accaduto in quella europea, l'attenzione si è concentrata sugli elementi in presenza dei quali gli intermediari non possono giovarsi della sfera di immunità conferita loro dagli art. 14 ss. d.lgs. 70/2003. L'idea che si è andata diffondendo rinviene in tale regime giuridico un trattamento di spiccato favoreaddirittura qualcuno parla di un trattamento premiale[78] non sempre adeguato a tutelare tutti gli interessi coinvolti dagli illeciti commessi in rete, specie dei titolari di diritti d'autore. Proprio per ridimensionare l'ampiezza della sfera di immunità offerta dagli art. 14 ss. d.lgs. 70/2003, sull'onda di alcuni spunti offerti dalla Corte di giustizia anche sulla scorta del considerando 42 della dir. 00/31, la giurisprudenza ha elaborato la figura pretoria dell'intermediario hosting attivo[79], per lo più con l'obiettivo di sottrarre alla sfera di immunità i user generated contentes. YouTube, i social network e i motori di ricerca. L'assunto alla base della figura del provider attivo consiste nella convinzione che l'immunità dell'intermediario presupponga la sua assoluta neutralità rispetto alle informazioni e ai contenuti immessi dai terzi in rete, sicché in presenza di una qualche forma di compartecipazione nella gestione o anche soltanto nell'organizzazione, inclusa la presentazione, di tali contenuti si fuoriuscirebbe dalle fattispecie legali che tracciano il perimetro della liceità della condotta degli intermediari. Tanto più se si tiene conto che sovente alla presentazio-

[77] Sul punto, *ex multis*, cfr. Di Ciommo, *Evoluzione tecnologica e regole di responsabilità civile*, cit., p. 289 ss.; Nivarra, *La responsabilità degli intermediari*, cit., p. 307 ss.; Facci, *La responsabilità dei providers*, cit., p. 238 ss.; Tosi, *Le responsabilità civili*, cit., p. 516 ss.; Id., *Le responsabilità civili dei prestatori di servizi della società dell'informazione*, cit., p. 197 ss.; Id., *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., p. 688 ss.; Bocchini, *La responsabilità civile degli intermediari del commercio elettronico,* cit., *passim*, in part. p. 123 ss.; D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p. 18 ss.

[78] Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., cc. 637-638.

[79] Tra le diverse pronunzie cfr. Trib. Catania, 29 giugno 2004, in *Dir. inf.*, 2004, p. 466 ss.; Trib. Milano, 2 marzo 2009 e Trib. Roma, 15 dicembre 2009, *ivi*, 2009, p. 521 ss.; Trib. Roma, 11 febbraio 2010, *ivi*, 2010, p. 275 ss.; Trib. Milano, 24 febbraio 2010, n. 1972, caso "The Pirate Bay", in *Riv. dir. ind.*, 2010, p. 328 ss. e in *Cass. pen.*, 2010, p. 3994 ss.; Trib. Milano, sez. spec. proprietà ind. e int., 9 settembre 2011, in *Riv. dir. ind.*, 2011, II, p. 364 ss., con nota di A. Saraceno, *Note in tema di violazione del diritto d'autore tramite Internet: la responsabilità degli* Internet Service Provider. E nella giurisprudenza di legittimità, Cass., 23 dicembre 2009, n. 49437, in *Foro it.*, 2010, II, c. 144 ss., con nota di S. Di Paola, *Sequestro preventivo di sito* web *e inibitoria del giudice penale dell'attività del* provider; in *Dir. inf.*, 2010, p. 437 ss., con nota di F. Merla, *Diffusione abusiva di opere in Internet e sequestro preventivo del sito web: il caso «The Pirate Bay»*, la quale, in realtà, si è occupata più specificamente della responsabilità del gestore di un sito webwww.thepiratebay.org il quale consentiva la condivisione tra gli utenti, mediante la tecnologia *peer to peer,* di file contenenti opere protette dal diritto d'autore. La Suprema corte ha ritenuto che l'esclusione del concorso di persone nel reato *ex* art. 110 c.p. avrebbe presupposto che il gestore del sito si fosse limitato a mettere a disposizione il protocollo di comunicazione necessario alla condivisione e al trasferimento dei file di opere coperte da diritto d'autore; mentre nel caso di specie il gestore si è spinto più in là procedendo all'indicizzazione delle informazioni che gli provengono dagli utenti. Un tale apporto attivo del gestore giustifica l'emissione della misura del sequestro preventivo del sito web il cui gestore concorra al compimento dell'attività penalmente illecita di diffusione nella rete di contenuti protetti dal diritto d'autore.

ne dei contenuti è collegato un ritorno economico per l'intermediario, specie in conseguenza del loro sfruttamento pubblicitario. Ma anche dove non vi sia un guadagno diretto, l'intermediario di frequente affianca al servizio di memorizzazione duratura delle informazioni prestazioni ulteriori, come quelle di indicizzazione, di selezione, di organizzazione o di filtraggio dei contenuti, le quali sembrano far sporgere la posizione dell'intermediario rispetto a quella presupposta dalle norme che gli garantiscono la sfera di immunità. In tali casi, si sarebbe in presenza di un provider attivo, figura mezzana tra il content provider e le figure di provider delineate dalla legislazione europea e dal d.lgs. 70/2003. La conseguenza sul piano del trattamento giuridico è l'assoggettamento del provider attivo alle regole comuni di cui all'art. 2043 c.c. Nonostante un certo scetticismo in dottrina[80], la giurisprudenza si avvale ampiamente della figura, ma non si può fare a meno di registrare una certa incertezza negli esiti applicativi. Al riguardo è emblematico lo sviluppo travagliato del caso RTI s.p.a. c. Yahoo! Italia s.r.l. e Yahoo! Inc., che ha impegnato i giudici milanesi e che riguarda l'illecita messa a disposizione dei terzi da parte di taluni utenti, mediante la piattaforma di videosharing del portale Yahoo!, di frammenti di contenuti audio/video riferibili a filmati coperti dalle prerogative di cui agli artt. 78 ter e 79 l.d.a. a favore della RTI s.p.a. in qualità, rispettivamente, di produttrice delle opere audiovisive e di emittente televisivo: filmati accessibili inserendo come parola chiave il titolo delle trasmissioni televisive alle quali i filmati si riferiscono e, per di più, associati a molteplici messaggi pubblicitari tramite link collegati ai medesimi titoli[81]. In prime cure, il Tribunale di Milano ha ravvisato nelle prestazioni concretamente fornite da Yahoo! Italia s.r.l. i tratti di un hosting attivo in quanto quest'ultima si riserva per via contrattuale: a di creare algoritmi dei contenuti immessi dai suoi utenti, di modificarli, di tradurli in appropriati format multimediali, standard o media così da renderli integrabili su Yahoo! Video; b di utilizzare, distribuire, riprodurre, modificare, remixare, adattare, estrarre, preparare opere derivate, riprodurre in pubblico e visualizzare pubblicamente i Contenuti Video su Yahoo! Video [...] nonché su siti di terze parti; c di utilizzare i Contenuti Video per attività pubblicitarie o per promozioni commerciali [...]; e si attribuisce d un diritto di manleva nei confronti dell'utente per gli eventuali danni prodotti dalla pubblicazione dei video da questi immessi; nonché e il diritto nei confronti dell'utente di rimuovere immediatamente i Contenuti Video immessi o di rifiutare l'inclusione nella lista dei video disponibili nel caso di rilevazione della violazione dei diritti di Yahoo! o di terzi. Per di più Yahoo! ha f predisposto un servizio, visibile come link sotto ogni video pubblicato in rete, che consente a ogni visitatore di segnalare all'intermediario l'eventuale illiceità del contenuto immesso dall'utente e alla redazione di verificare la segnalazione e di rimuovere il contenuto contestato, il che denota per il Tribunale l'esercizio di un controllo da parte del provider sulla liceità delle informazioni memorizzate successivo alla loro immissione; e, inoltre, ha g incluso un servizio aggiuntivo di c.d. video correlati che consiste nella visualizzazione di ulteriori contenuti

[80] Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 639; D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p. 78 ss., p. 91.

[81] Trib. Milano, sez. spec. proprietà ind. e int., 9 settembre 2011, cit., p. 364 ss. e App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, in *Corriere giur.*, 2016, p. 811 ss., con nota di E. Bassoli, *Il diritto d'autore e la responsabilità del* provider: *evoluzioni tecniche e giurisprudenziali nell'appello* Yahoo vs. RTI.

audio/video non ricercata dal visitatore e offertagli, invece, in via automatica proprio in quanto correlata ai video specificamente ricercati, da cui si evince un'ulteriore e specifica attività di indicizzazione dei contenuti video che sottintende una selezione dei contenuti, amplificandone le possibilità di visione e di diffusione[82]. Dalla convergenza di tutti questi elementi il Tribunale ricava lo svolgimento da parte dell'intermediario di un'attività tutt'altro che neutrale di selezione e gestione dei contenuti immessi dagli utenti, finalizzata - ancorché mediante software - ad arricchirne e completarne la fruizione in vista di uno sfruttamento economico di tali contenuti che travalica la mera remunerazione del servizio offerto all'utente[83]. La qualificazione come hosting attivo comporta la sottrazione di Yahoo! alla sfera di immunità di cui all'art. 16 d.lgs. 70/2003, ma non implica certo l'imposizione di un obbligo di verifica preventiva dei materiali immessi sulle proprie infrastrutture, a causa dell'estrema difficoltà tanto tecnica quanto giuridica della sua attuazione. E tuttavia ciò non esclude l'insorgere dell'obbligo di attivazione a seguito della diffida da parte di chi si accrediti in maniera puntuale come titolare dei diritti di utilizzazione dei contenuti: un obbligo successivo che il Tribunale ha ritenuto non essere stato, però, adempiuto da Yahoo! nel caso di specie, nonostante la specifica e documentata segnalazione di RTI, il che rende fondata la richiesta del provvedimento inibitorio ex art. 156 l.d.a. e la connessa fissazione dell'obbligo di corrispondere una somma di danaro per ogni violazione o inosservanza successiva, che il collegio qualifica erroneamente come penale piuttosto che come astreinte[84].

I medesimi elementi che hanno indotto il Tribunale di Milano a collocare i servizi resi da Yahoo! nell'area dell'hosting attivo vengono valutati in senso diametralmente opposto dalla Corte d'appello, la quale è approdata, al termine di una motivazione certo non sintetica, alla conclusione che «le attuali tecnologie avanzate, in mancanza di altri elementi in grado di fare intravedere una vera e propria manipolazione dei dati immessi da parte dell'hosting provider, non siano da sole in grado di determinare il mutamento della natura del servizio di hosting provider di tipo passivosecondo la classificazione utilizzata dalla giurisprudenza nazionale richiamata dalla sentenza appellata, in servizio di hosting provider di tipo attivo, in ragione della mera presenza i di sofisticate tecniche di intercettazione del contenuto dei file caricati, attraverso un motore di ricerca, e ii delle più svariate modalità di gestione del sito e iii del particolare interesse del gestore a conseguire vantaggi economici»[85]. Più nel particolare, invocando la sentenza del caso L'Oréal c. eBay[86], la Corte d'appello ha considerato le diverse attività svolte da Yahoo!, sulla cui base il giudice di prime cure ha ritenuto di qualificare il provider come attivo, nulla più che manifestazioni di una più complessiva attività di trattamento dei dati immessi dagli utenti di natura meccanica e non manipolativa, come tali non sufficienti a sottrarre l'intermediario alla sfera di immunità delineata dall'art. 16 d.lgs. 70/2003[87]. La soluzione appare alla Corte

[82] Trib. Milano, sez. spec. proprietà ind. e int., 9 settembre 2011, cit., pp. 370-371.
[83] Trib. Milano, sez. spec. proprietà ind. e int., 9 settembre 2011, cit., p. 372.
[84] Trib. Milano, sez. spec. proprietà ind. e int., 9 settembre 2011, cit., pp. 374-375.
[85] App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., n. 24.
[86] Corte giust., 12 luglio 2011, C-324/09, cit.
[87] App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., n. 27. Poco oltre la sentenza precisa che «In particolare, si ritiene che i riscontrati servizi pubblicitari gestiti dal Fornitore di Accesso a InternetFAI,

d'appello, peraltro, in linea con le sentenze Scarlet Extended c. SABAM[88] e Netlog c. SABAM[89] e da ciò la Corte ricava un elemento di contraddittorietà nella sentenza del Tribunale annidato nella scelta di qualificare Yahoo! come un intermediario attivo e al contempo di riconfermare anche nei suoi confronti l'operatività dell'art. 17 d.lgs. 70/2003, ossia l'assenza di un dovere generale di sorveglianza[90]. Alla luce di questi rilievi e di ulteriori indicazioni, in realtà frutto di una lettura un poco esasperata[91], della più recente giurisprudenza europea, la Corte d'appello perviene all'accantonamento della figura dell'intermediario attivo[92], giudicata addirittura, alla luce del quadro giuridico e giurisprudenziale, una nozione «sicuramente fuorviante e sicuramente da evitare concettualmente in quanto mal si addice ai servizi di "ospitalità in rete" in cui il prestatore non interviene in alcun modo sul contenuto caricato dagli utenti, limitandosi semmai a sfruttarne commercialmente la presenza sul sito, ove il contenuto viene mostrato così come è caricato dall'utente senza alcuna ulteriore elaborazione da parte del prestatore»[93]. Al di là di affermazioni di sistema così impegnative, nel merito la Corte ritiene che la diffida rivolta a

i diritti di utilizzo e di riadattamento dei contenuti caricati a sé riservati, il diritto di manleva nei confronti dell'utente stabilito nelle condizioni integrative dell'accesso alla rete, nonché il potere di rimozione dei contenuti caricati e la facoltà di segnalazione degli illeciti da parte dell'utente, considerati nel loro insieme, non sono indici rivelatori di un'attività d'interferenza sui contenuti pubblicati nel sito, come tali in grado di mutare il regime di "responsabilità a posteriori" dell'*hosting provider* delineato nella direttiva esaminata e nelle pronunce della Corte di Giustizia, in quanto non essenzialmente in grado di alterare l'integrità dell'informazione contenuta nella trasmissione».

[88] Corte giust., 24 novembre 2011, C-70/10, cit.

[89] Corte giust., 16 febbraio 2012, causa C-360/10, cit.

[90] App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., n. 28. In verità, anche se si volesse accedere alla qualificazione compiuta dal Tribunale, la contraddizione non vi sarebbe: non si ravvisano, infatti, ragioni logiche e giuridiche che impediscano di sottrarre un intermediario di *hosting* dalla disciplina dell'art. 16 d.lgs. 70/2003 mantenendo fermo il disposto dell'art.17 d.lgs. 70/2003. L'applicazione del regime di responsabilità ordinario di cui agli artt. 2043 e ss. c.c. non comporta anche l'imposizione di un dovere generale di sorveglianza attiva, o per lo meno non lo comporta in tutte le circostanze. La colpa presupposta dall'art. 2043 c.c. potrebbe, infatti, sostanziarsi nella negligente e imprudente sottovalutazione di indici in grado, se adeguatamente considerati, di far emergere il carattere illecito delle informazioni o dei contenuti immessi dall'utente. In altri termini, è sul versante dello stato soggettivo di conoscenza dell'intermediario che potrebbe incidere l'applicazione del regime ordinario di responsabilità, rendendo più severa la valutazione che conduce a considerare colposa l'ignoranza degli illeciti commessi per il tramite dei propri

[91] Ci si riferisce in particolare al passaggio in cui la Corte d'appello ritiene di ricavare dalla sentenza a *Telekabel* c. *Constantin Film*Corte giust., 27 marzo 2014, C-314/12 un ordine assiologico che relega il diritto fondamentale d'autoreart. 17 Carta dei diritti fondamentali in posizione subordinata rispetto alla libertà d'impresaart. 16 Carta e alla libertà di espressione e di informazioneart. 11: «Conseguentemente, la regola generale che si ricava in una materia in cui il diritto d'autore, considerato quale valore fondamentale nell'art. 17 della Carta dei diritti fondamentali, si può contrapporre ad altri valori fondamentali, quali quelli attinenti alla libertà d'impresaart. 16 e alla libertà di espressione e d'informazioneart. 11, è nel senso di risolvere il conflitto tra i tre valori fondamentali in gioco dando prevalenza agli ultimi due»App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., n. 33.

[92] Affermazione, questa, che ha sollecitato le critiche dell'annotatrice della sentenza: E. Bassoli, *Il diritto d'autore e la responsabilità del* provider: *evoluzioni tecniche e giurisprudenziali nell'appello* Yahoo vs. RTI, cit., p. 823 ss.

[93] App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., n. 38.

Yahoo! da RTI non sia stata sufficientemente specifica nell'indicare i contenuti illeciti tanto da poter determinare l'insorgere a carico dell'intermediario dell'obbligo di rimozione immediata dei programmi illecitamente immessi in rete[94]. E inoltre i giudici del gravame hanno ritenuto che neppure la successiva azione giudiziale sia stata idonea a determinare l'autonoma insorgenza dell'obbligo di pronta rimozione, giacché tutt'al più capace di attivare i poteri inibitori del giudice volti ad imporre al provider la sola rimozione selettiva degli atti illeciti denunciati e non certo anche gravosi ordini generali o, peggio ancora, obblighi di sorveglianza generale[95]. Solo a seguito dell'incardimento dell'azione, RTI ha provveduto al deposito di un documento che conteneva la specifica individuazione dei dati illecitamente immessi in rete mediante la specificazione dei relativi URL. In conseguenza di tale produzione documentale Yahoo! ha spontaneamente rimosso i contenuti contestati, senza attendere l'ordine giudiziale, sicché il giudizio è proseguito relativamente a ulteriori URL indicati da RTI in corso di causa come illeciti per violazione dei propri diritti di sfruttamento economico delle opere fissate su supporto informatico. La richiesta di RTI si è indirizzata verso l'ottenimento di un provvedimento di inibizione di ogni uso e sfruttamento commerciale diretto o indiretto, anche per mezzo di soggetti/società da esse controllati e/o collegati e/o con cui comunque esistono rapporti/accordi imprenditoriali finalizzati alla gestione degli insert/link pubblicitari sul portale Yahoo!.it e dei relativi proventi, che si traduca nella violazione, in qualunque forma e con qualunque mezzo, dei diritti esclusivi di RTI. La Corte d'appello ha ritenuto che l'accoglimento di una tale richiesta comporterebbe l'imposizione a Yahoo! di realizzare un sistema di filtraggio di tutte le informazioni memorizzate nei propri server, applicato indistintamente a tutta la sua clientela, a titolo preventivo, a sue spese esclusive, senza limiti di tempo e che ciò si porrebbe in contrasto con la disciplina europea e nazionale dell'attività degli intermediari come interpretata dalla Corte di giustizia[96]. Di conseguenza non è neppure ravvisabile la responsabilità di Yahoo! per non aver predisposto un tale sistema di filtraggio né vi sono gli estremi per configurare la responsabilità da mancata rimozione dei contenuti illeciti non appena venuti a conoscenza dell'intermediario, in quanto la diffida non era sufficientemente specifica per determinate una tale conoscenza, che, invece, si è avuta soltanto in corso di causa, provocando l'immediata attivazione del provider[97].

Le oscillazioni della giurisprudenza sull'ammissibilità o meno della figura del provider attivo e, in ogni caso, sulla tipologia di attività e di servizi idonei a sospingere un intermediario in

[94] La Corte giudica infatti «pacifico che la diffida stragiudiziale non contenesse l'indicazione esatta degli URL o dei link dei video contenuti nel sito di YAHOO!I. La Corte, in merito, ritiene che una mera generica ricerca per nome o titolo commerciale dell'opera considerata illecita non sia sufficiente a determinare l'insorgere dell'obbligo di rimozione in capo al FAI, nel senso che una diffida siffatta non avrebbe mai potuto far venire meno la presunta neutralità del gestore, e quindi attivare la sua responsabilità a posterioricome sopra definita nei suoi contorni»

[95] App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., n. 52.

[96] giust., Sez. III, 24 novembre 2011, C-70/10, *Scarlet Extended* c. *SABAM* e Corte giust., 16 febbraio 2012, causa C-360/10, *SABAM* c. *Netlog*.

[97] App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., nn. 61-70. Per soprammercato, la Corte considera inammissibile l'estensione della domanda inibitoria a fatti illeciti successivamente individuati in corso di causa dal titolare del diritto d'autore, cui tuttavia l'intermediario si è spontaneamente conformato in corso di giudizio.

tale categoria dovrebbero mettere in guardia sull'attitudine di quest'ultima a fungere da vera e propria chiave di volta del sistema delle regole applicabili agli intermediari. A un'analisi più attenta, e a dispetto delle osservazioni dei primi commentatori[98], nell'economia delle due pronunzie la figura del provider attivo o il suo ripudio non appaiono in definitiva centrali, giacché il vero punto nevralgico della controversia sembra concentrarsi piuttosto sulla completezza e sulla specificità della diffida indirizzata all'intermediario dal titolare dei diritti di utilizzazione e di sfruttamento economico delle opere audiovisive. Il dato è estremamente significativo ai fini dell'intelligenza della corretta sistemazione della disciplina degli *internet service provider*.

4. La ricostruzione del sistema della c.d. responsabilità degli *internet service provider*, delineato dalla legislazione europea e recepito nei diritti nazionali, con particolare riguardo all'attività di hosting si presenta quanto mai controversa, a causa della diversità delle opinioni sul rapporto che intercorre tra i due presupposti di liceità della prestazione dell'intermediario fissati dall'art. 16 comma 1, lett. a e b, d.lgs. 70/2003. In breve, la questione si appunta sulla portata del requisito della conoscenza[99] previsto dall'art. 16, comma 1, lett. a d.lgs. 70/2003 e, in particolare, sul nodo relativo a quale forma di conoscenza faccia sorgere l'obbligo di intervento immediato per la rimozione delle informazioni comunicate in violazione di un diritto di esclusiva o di un diritto della persona o per la disabilitazione al loro accesso, sancito dall'art. 16, comma 1, lett. b d.lgs. 70/2003. Giurisprudenza e dottrina concordano sul necessario concorso di entrambi i presupposti per poter considerare liceità la condotta dell'hosting provider, sicché il quadro che ne emerge vede il provider estraneo alla violazione commessa dall'utente in due ipotesi: a se è senza colpa ignaro che la comunicazione di informazioni compiuta dall'utente consegue alla violazione di un'altrui posizione giuridica soggettiva; b se, una volta venuto a conoscenza dell'origine illecita della comunicazione, si attivi prontamente per rimuovere le informazioni o per disabilitarne l'accesso. La convergenza delle opinioni sull'interpretazione dell'art. 16 d.lgs. 70/2003 si arresta qui, lasciando campo aperto alla disparità di vedute sulla forma di conoscenza cui fa riferimento l'art. 16, comma 1, lett. a d.lgs. 70/2003. Da un lato si collocano coloro che ritengono sufficiente la conoscenza semplice dell'illiceità della condotta dell'utilizzatore, acquisita su segnalazione del titolare del diritto leso[100] o anche aliunde,

[98] Cfr. Bassoli, *Il diritto d'autore e la responsabilità del* provider: *evoluzioni tecniche e giurisprudenziali nell'appello* Yahoo vs. RTI, cit., p. 823 ss.

[99] Sottolinea la centralità della conoscenza da parte dell'intermediario del comportamento antigiuridico commesso dal terzo-utente Nivarra, *La responsabilità degli intermediari*, cit., p. 313.

[100] In tal senso Trib. Milano, ord., 25 gennaio 2011, *Bardolla* c. *Google Suggest*, confermata da Trib. Milano, 31 marzo 2011, in *Riv. dir. ind.*, 2011, p. 21 ss., con nota di E. Tosi, *La responsabilità civile per fatto illecito degli Internet Service Provider e dei motori di ricerca a margine dei recenti casi "Google Suggest" per errata programmazione del software di ricerca e "Yahoo! Italia" per "link" illecito in violazione dei diritti di proprietà intellettuale*; Trib. Milano, sez. spec. proprietà ind. e int., 7 giugno 2011; Trib. Roma, sez spec. proprietà ind. e int., 20 marzo 2011, *PFA Films* c. *Yahoo.it,* in *Dir. ind.*, 2012, p. 79 ss. con commento di L. Giove-A. Comelli, *Responsabilità del* provider *per mancata rimozione di* link *a materiale illecito* e in *Riv. dir. ind.*, 2012, p. 29 ss.revocata però da Trib. Roma, sez. spec. proprietà ind. e int., ord., 11 luglio 2011, *ivi*, p. 37 ss. in quanto la segnalazione da parte di *PFA Films* è stata ritenuta del tutto generica e così anche il ricorso nella prima fase del procedimento cautelare e poi il successivo reclamo, in quanto la *PFA Films* per ciascuno dei contenuti immessi

attribuendo rilievo preminente alla formulazione dell'art. 16, comma 1, lett. a d.lgs. 70/2003, il quale collega l'esclusione dell'hosting provider dal concorso nella responsabilità dell'autore della violazione alla circostanza che il primo «non sia al corrente di fatti o di circostanze che rendano manifesta l'illiceità dell'attività o dell'informazione». Dal lato opposto si collocano, invece, coloro che, con un approccio più favorevole agli intermediari, ritengono che la conoscenza destinata a far scattare l'obbligo di rimozione dei contenuti illeciti o di disabilitazione dell'accesso debba rivestire una forma qualificata[101]. In questa prospettiva, la segnalazione, anche circostanziata, da parte del titolare della situazione soggettiva lesa non si rivela sufficiente a far sorgere l'obbligo di intervento in capo al provider, essendo piuttosto necessaria la comunicazione dell'autorità amministrativa competente o dell'autorità giudiziaria. Tale posizione trae fondamento dalla scelta di conferire maggiore rilevanza alla formulazione dell'art. 16, comma 1, lett. b d.lgs. 70/2003, che prescrive all'intermediario di agire prontamente per la rimozione dei contenuti illeciti o per la disabilitazione dell'accesso «non appena a conoscenza di tali fatti, su comunicazione delle autorità competenti». Balza immediatamente agli occhi la profonda differenza delle conseguenze applicative collegate alle due letture.

Vi sono, però, fondate ragioni per respingere entrambe le interpretazioni e per accedere alla completa re-impostazione del meccanismo di immunità delineato dall'art. 16 d.lgs. 70/2003. Va in primo luogo segnalato che la formulazione dell'art. 16, comma 1, lett. b d.lgs. 70/2003 diverge dalle previsioni della dir. 00/31, la quale all'art. 14, par. 1, lett. b collega la liceità della condotta dell'hosting provider al presupposto che «non appena al corrente di tali fatti, [questi] agisca immediatamente per rimuovere le informazioni o per disabilitarne l'accesso». La medesima disposizione prevede al par. 3 la possibilità «per un organo giurisdizionale o un'autorità amministrativa, in conformità agli ordinamenti giuridici degli Stati membri, di esigere che il prestatore ponga fine ad una violazione o la impedisca nonché la possibilità, per gli Stati membri, di definire procedure per la rimozione delle informazioni o la disabilitazione dell'accesso alle medesime». È evidente che la previsione dell'art. 14, par. 3, dir. 00/31 integra una fat-

in rete, di cui essa affermi la provenienza da un soggetto non autorizzato, avrebbe dovuto fornire l'indicazione dei codici di identificazione telematicaURL in cui è disponibile il filmato contestato; e nello stesso senso v. anche Trib. Roma, 16 giugno 2011, *Yahoo! Italia* c. *Alfa Films*, in *Dir. ind.*, 2012, p. 75 ss.; App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., nn. 46 ss., dove in particolare è riconosciuto che «i rimedi sopra citati, di autotutela mediante diffida, e di eterotutela mediante ottenimento di un ordine di rimozione giurisdizionale o amministrativo, si pongono quali strumenti equivalenti e alternativi per chi si senta leso nei propri diritti, e il Tribunale, su questo punto, non ha fatto che confermare una linea interpretativa accolta dalla giurisprudenza di settore. Dirimente appare la stessa direttiva 2000/31 laddove, riguardo alla specifica responsabilità dell'hosting provider, nell'art. 14».

[101] Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., § 5: «una mera diffida se può ritenersi idonea nel caso degli ISP attivi - stante l'atipicità soggettiva che può, se adeguatamente motivata, fondare una deroga normativa, in particolare sotto il profilo della conoscenza effettiva dell'illecito - nel caso degli ISP passivi per i quali la conoscenza legale decorre - ai sensi dell'art. 16, 1° co. e art. 17, 3° co., d.lg. n. 70/2003 - esclusivamente dalla comunicazione dell'*autorità giudiziaria* o *amministrativa avente funzioni di vigilanza*». E in giurisprudenza: Trib. Firenze, 25 maggio 2012, in *Dir. inf.*. 2012, p. 1210 ss.

tispecie ulteriore rispetto a quella dell'art. 14, par. 1, lett. b dir. 00/31, ma nell'attuazione fattane in Italia le due previsioni sono state innestate e dalla crasi è fuoriuscito un radicale ridimensionamento della portata dell'art. 14, par. 1, lett. b dir. 00/31, giacché quell'obbligo di intervento dell'intermediario che la norma europea collega alla mera conoscenza, nell'art. 16, comma 1, lett. b d.lgs. 70/2003 scatta in presenza della conoscenza determinata dalla comunicazione delle autorità competenti. Si tratta con tutta evidenza di una trasposizione infedele della direttiva europea[102], frutto, però, con ogni probabilità, non già di un errore ma di una scelta deliberata volta a rendere meno rigoroso il regime applicabile agli hosting provider, comune peraltro ad altri ordinamenti europei, come, ad es., quello spagnolo[103]. Se ne ha una riprova nella soluzione escogitata dal legislatore italiano per conferire un senso alla previsione dell'art. 14, par. 3, dir. 00/31 relativa alla possibilità dell'autorità giudiziaria, o dell'autorità amministrativa di controllo, di imporre all'intermediario di mettere fine alla violazione delle altrui prerogative o addirittura di prevenirle. Svuotata di gran parte del suo significato dalla scelta di elevare la comunicazione delle autorità competenti a presupposto dell'obbligo di attivazione dell'intermediario, la norma europea è stata recuperata dal legislatore italiano all'art. 16, comma 3, d.lgs. 70/2003 innestandogli un riferimento all'intervento d'urgenza, che non figura nell'archetipo europeo e che sottrae il comma 3 alla sorte di presentarsi come una stanca ripetizione del comma 1 lett. b.

Il sistema di regole che emerge sulla base dell'interpretazione prevalente che assegna valore dirimente al tenore dell'art. 16, comma 1, lett. b d.lgs. 70/2003 non può essere accettata perché contrasta in maniera irriducibile con il diritto europeo[104]. Il recepimento non fedele dell'art. 14 dir. 00/31 da parte della normativa italiana non impedisce, però, di approdare a un'interpretazione conforme al diritto europeo, che – com'è sin troppo noto – costituisce oggi uno dei canoni ermeneutici prevalenti. Per conseguire un tale obiettivo si rende però necessario

[102] Eccede, dunque, App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., n. 20 quando afferma che «Le disposizioni della normativa nazionale sono una fedele trasposizione dei principi affermati dal legislatore europeo con la direttiva 2000/31/CE, ovvero agli artt. 12-15 della direttiva sul commercio elettronico» e ribadisce che la legislazione italiana si rivela una "fedele esecutrice" della normativa european. 25.

[103] L'art. 16 della *Ley* n. 34 dell'11 luglio 2002, *de Servicios de la Sociedad de la Información y de Comercio Electrónico*, collega l'effettiva conoscenza dell'intermediario al preventivo accertamento della natura illecita dei contenuti immessi in rete dal destinatario del servizio da parte dell'autorità competente. In Finlandia, la normativa distingue la violazione della proprietà industriale dalla violazione del diritto d'autore: nel caso di violazione del marchio, l'intermediario è tenuto a rimuovere i contenuti memorizzati nei propri server soltanto a seguito di un ordine giudiziario; mentre, nel caso di violazione del diritto d'autore, è sufficiente la segnalazione da parte di colui che si dichiara titolare delle relative prerogative, cui l'utente che ha immesso i contenuti contestati può replicare con un'opposizione. Differenti sono le soluzioni adottate negli ordinamenti inglese e francese, dove la diffida di parte è reputata sufficiente a far conseguire all'intermediario la conoscenza qualificata dell'illecito, a condizione che essa sia circostanziata e sufficientemente dettagliata. Sul punto v. la sintesi di Petruso, *Responsabilità degli intermediari di internet e nuovi obblighi di conformazione:* robo-takedown, policy of termination, notice and take steps, cit., p. 492 nt. 59.

[104] Una diversa ragione di difformità tra diritto italiano e diritto europeo è suggerita da Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 242 nt. 31, ma sul punto si tornerà di qui a breve.

abbandonare la più volte criticata impostazione che ravvisa negli artt. 14 e ss. d.lgs. 70/2003 una normativa sulla responsabilità degli *internet service provider* e accogliere la prospettiva più ampia che vi riconosce, invece, una normativa sui presupposti di liceità degli intermediari: in via diretta di quelli le cui prestazioni sono connotate da neutralità rispetto ai contenuti e in via indiretta anche di quelli che in varia misura sono a conoscenza dei contenuti immessi dagli utenti, se non addirittura concorrono a definirli.

5. La ridefinizione del significato e dei contenuti del sistema di regole delineato dagli artt. 14 e ss. d.lgs. 70/2003 esige, in primo luogo, di sottoporre a interpretazione restrittiva l'impiego del termine "responsabilità" compiuto dalle disposizioni italiane, qui d'altro canto in linea con le corrispondenti norme europee della sezione 4, le quali adoperano il termine "responsabilità". Dall'interpretazione complessiva delle disposizioni emerge, infatti, che quando gli artt. 14, 15 e 16 d.lgs. 70/2003 indicano le condizioni sotto le quali i prestatori di mere conduit, caching e hosting non sono "responsabili" delle informazioni trasmesse o della loro memorizzazione non fanno altro che delineare i presupposti in presenza dei quali gli intermediari non rispondono delle violazioni commesse dagli utenti per il loro tramite. Il predicato verbale "rispondere" evoca la categoria dell'ascrizione delle conseguenze negative della condotta illecita altrui e, quindi, si presenta ben più ampia dell'istituto della responsabilità che ne rappresenta soltanto una delle manifestazioni. In altri termini, gli art. 14 e ss. d.lgs. 70/2003 non si limitano a porre regole di responsabilità, la quale certamente può discendere dalla violazione di tali disposizioni ma certo non ne esaurisce il significato che, invece, si mostra ben più ampio, abbracciando l'intero spettro dei rimedi che discendono dall'illiceità della condotta dell'intermediario. Non a caso larga parte della casistica riguarda la possibilità o meno di ottenere nei confronti dell'internet provider provvedimenti di natura interdittiva o ripristinatoria, riconducibili alla figura dell'inibitoria, che in alcun modo si possono considerare effetti di responsabilità, prescindendo del tutto dai presupposti di quest'ultimadanno ingiusto, nesso causale, criterio di imputazione[105]. E gli stessi art. 14 e ss. prevedono espressamente che tanto l'autorità giudiziaria quanto l'autorità amministrativa di vigilanza possano indirizzare all'intermediario di mere conduit, caching e hosting, anche in via d'urgenza, provvedimenti volti a prevenire o a stroncare le violazioni commesse dagli utenti. In questo modo assume un significato più chiaro la più volte segnalata proposta di qualificare il c.d. sistema della responsabilità degli *internet service provider* non come un complesso di regole sull'esclusione e, correlativamente, sull'imputazione agli intermediari dei danni prodotti dalla violazione delle situazioni soggettive altrui da parte dei loro utenti, ma come un complesso di regole sul perimetro della liceità e, correlativamente, dell'illiceità della condotta degli intermediari al cospetto della violazione delle situazioni soggettive altrui da parte dei loro utenti. La conseguenza è che la gamma dei rimedi invocabili nei confronti dell'intermediario si presenta ben più ricca del solo risarcimento del danno[106], includendo anche l'inibitoria, la

[105] Sotto questo profilo, adotta un linguaggio scorretto sul piano dell'impiego delle categorie civilistiche Trib. Milano, sez. spec. proprietà ind. e int., 9 settembre 2011, cit., p. 374 quando parla di responsabilità con riferimento a una domanda di provvedimento inibitorio.

[106] Sul più ampio tema del risarcimento del danno per violazione della proprietà intellettuale cfr. M. Ricolfi, *Il danno da violazione di proprietà intellettuale nella giurisprudenza della Corte di Giustizia*, in *Giur. it.*, 2017,

restituzione di profitti, la nullità degli accordi sull'ampiamento della sfera di immunità degli intermediari in deroga della disciplina di cui agli artt. 14 e ss. d.lgs. 70/2003 etc. Per di più, non è affatto detto che l'inapplicabilità del regime delineato da tali disposizioni comporti l'inevitabile riconoscimento della responsabilità dell'intermediario, la quale potrebbe essere esclusa per altre ragioni: dall'assenza di qualcuno degli elementi costituitivicolpa, nesso causale, danno ingiusto oppure dalla presenza di scriminanti di natura diversa[107].

Nella prospettiva più larga dischiusa dalla coppia liceità-compartecipazione all'illiceità dell'altrui condotta acquista una forte credibilità la proposta di scindere il contenuto dell'art. 16, comma 1, d.lgs. 70/2003 in due diverse norme. L'art. 16, comma 1, lett. a d.lgs. 70/2003, nel collegare all'ignoranza incolpevole delle violazioni perpetrate dagli utenti servendosi delle proprie infrastrutture la liceità della condotta dell'intermediario, nei termini dell'immunità dal concorso nella condotta illecita posta in essere dall'utente, enuncia implicitamente anche una regola di condotta nel caso in cui l'internet provider acquisti invece conoscenza della violazione[108]. A dispetto dell'interpretazione che reputa lecita l'inerzia dell'intermediario in presenza di una conoscenza semplice finché non intervenga la comunicazione dell'autorità giudiziaria o di quella amministrativa di vigilanza, la condizione di conoscenza, comunque acquisita, impone all'hosting provider di adottare le regole di perizia professionale, che nella fattispecie in esame si traducono nell'adozione delle misure necessarie alla cooperazione per la verifica e il contrasto degli illeciti commessi per il tramite dei suoi servizi[109]. Al riguardo si traggono utili indicazioni dal considerando 46, il quale specifica che «Per godere di una limitazione della responsabilità, il prestatore di un servizio della società dell'informazione consistente nella memorizzazione di informazioni deve agire immediatamente per rimuovere le informazioni o per disabilitare l'accesso alle medesime non appena sia informato o si renda conto delle attività illecite»[110]. E qui si aprono due scenari.

Nel caso di conoscenza semplice, ossia qualora abbia contezza della possibile natura illecita della comunicazione o dell'utilizzo delle informazioni memorizzate, l'intermediario ha l'obbligo di segnalare all'autorità giudiziaria e all'autorità amministrativa l'ipotesi di illecito, offrendo gli elementi utili per risalire all'identità dell'autore della violazione e quelli a disposizione per accertare la sussistenza o meno dell'illecito. A seguito del positivo accertamento, giudiziale o

p. 680 ss.

[107] Un rilievo analogo è in Seminara, *Internetdiritto penale*, cit., pp. 592-593, con riferimento alla responsabilità penale.

[108] Pone, assai opportunamente, al centro della propria lettura della disciplina degli *internet service provider* lo stato di conoscenza dell'illecito da parte di questi ultimi Bocchini, *La responsabilità extracontrattuale del provider*, in *Manuale di diritto dell'informatica*, p. 540 ss.; Id., *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 636 ss.

[109] Sottolinea l'importanza del ruolo della diligenza professionale dell'intermediario, D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., pp. 39-41.

[110] Altri utili elementi vengono forniti dal considerando n. 48, secondo cui. «La presente direttiva non pregiudica la possibilità per gli Stati membri di chiedere ai prestatori dei servizi, che detengono informazioni fornite dai destinatari del loro servizio, di adempiere al dovere di diligenza che è ragionevole attendersi da loro ed è previsto dal diritto nazionale, al fine di individuare e prevenire taluni tipi di attività illecite».

amministrativo, dell'illecito scatta la fattispecie prevista dall'art. 16, comma 1, lett. b d.lgs. 70/2003 e, pertanto, a seguito della comunicazione dell'esito dell'esame da parte dell'autorità, sorge in capo all'intermediario l'obbligo di provvedere senza meno alla rimozione dei contenuti frutto della violazione delle altrui situazioni soggettive oppure alla disabilitazione dell'accesso a tali dati. Qualora non adempia nei tempi tecnicamente necessari all'eliminazione delle conseguenze della condotta illecita, l'intermediario è esposto alla responsabilità per gli eventuali danni prodotti dalla protrazione della violazioni anche a seguito della comunicazione dell'accertata violazione, configurandosi un'ipotesi di concorso di cause di cui all'art. 1227, comma 1, c.c. richiamato in sede aquiliana dall'art. 2056, comma 1, c.c., con applicazione del regime di responsabilità solidale di cui all'art. 2055 c.c. Si tratta di un'ipotesi di concorso colposo per mancata cooperazione nel contrasto di un illecito oramai appurato[111]. Non può tuttavia essere sottovalutata la questione relativa alla portata dell'ordine inibitorio, o come pure si dice del provvedimento ingiuntivo: quanto esso contiene misure ripristinatorie dello status quo ante, nulla quaestio; ma quando si estende all'adozione di accorgimenti per il futuro, tesi a prevenire nuove violazioni, sorgono i problemi relativi all'estensione delle misure esigibili, il che non può che riflettersi sul terreno della responsabilità e dell'esonero da essa per impossibilità dell'attuazione del provvedimento o per difetto di proporzionalità[112].

La seconda fattispecie ricavabile dall'art. 16, comma 1, d.lgs. 70/2003 non poggia, invece, sul combinato disposto delle lettere a e b, ma si impernia sul secondo periodo della lett. a, che è specificamente riferito al risarcimento del danno: «per quanto attiene ad azioni risarcitorie, non sia al corrente [l'hosting provider n.d.a.] di fatti o di circostanze che rendono manifesta l'illiceità dell'attività o dell'informazione». L'elemento posto a fondamento di questa seconda fattispecie consiste non nella mera conoscenza di una presunta attività illecita, ma nella conoscenza di un'illiceità manifesta[113], alla quale il hosting provider può pervenire nelle maniere più varie: a seguito di propri accertamenti più o meno casuali; della segnalazione particolarmente circostanziata da parte di colui che si proclama titolare del diritto leso; di una segnalazione generica dell'autorità amministrativa; di notizie di stampa etc.[114]. Il carattere manifesto dell'illiceità di

[111] Per il fondamento colposo della responsabilità del *provider* si dichiarano Nivarra, *La responsabilità degli intermediari*, cit., p. 314 e Ponzanelli, *Verso un diritto uniforme per la responsabilità degli* internet service providers*?*, cit., p. 10; Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 637, c. 643

[112] Nello stesso senso mi pare si muovano le considerazioni di Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 249-250.

[113] Coglie la centralità della nozione di "conoscenza" ma non ne trae le giuste conseguenze in sede di definizione delle norme applicabili agli intermediari: Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca, social network e aggregatori di contenuti di terzi*, cit., par. 5.

[114] Una diversa lettura della disposizione europea funge da matrice all'art. 16 d.lgs. 70/2003, ossia dell'art. 14, par. 1, lett. *a* dir. 00/31, è proposta da Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 214 nt. 31. Secondo questa prospettiva, la formula della versione inglese della direttiva «aware of facts or circumstances from which the illegal activity or information is apparent» riferita al rimedio risarcitorio va intesa, soprattutto in confronto all'«actual knowledge» riferita, invece, al rimedio inibitorio, nel senso che per ricorre alla tutela risarcitoria è sufficiente la mera conoscibilità della natura illecita dei contenuti,

cui l'intermediario è venuto a conoscenza fa sorgere, anche in questo caso, l'obbligo di adottare le regole della perizia professionale, le quali, nel frangente in esame, non si possono limitare, come nella fattispecie precedente, all'obbligo di segnalazione e di cooperazione con l'autorità giudiziaria e con quella amministrativa, imponendo piuttosto l'immediato intervento per eliminare le conseguenze della violazione, tramite cancellazione dei contenuti illeciti o disabilitazione dell'accesso. Qualora non provveda, trincerandosi dietro l'assenza di una comunicazione da parte delle autorità competenti, l'intermediario è esposto alla responsabilità nei confronti del titolare della situazione giuridica lesa in concorso con l'autore della violazione in una misura che, a seconda delle caratteristiche del caso concreto, potrebbe anche risultare paritaria. In tal modo viene recuperata in sede interpretativa la conformità del diritto italiano al diritto europeo, grazie all'elaborazione di una norma che ricalca in maniera puntuale la regola ricavabile dall'art. 14, par. 1, dir. 00/31. Se ciò vale per gli illeciti già commessi, resta il problema del contenuto e dell'ampiezza delle misure di vigilanza e di precauzione da adottare per scongiurare la reiterazione delle condotte lesive da parte del medesimo cliente e qui non si può che seguire la stella polare del carattere proporzionato, non eccessivamente oneroso e ragionevole degli accorgimenti dovuti[115].

Le due fattispecie ricavate dall'art. 16, comma 1, d.lgs. 70/2003 presentano una matrice comune: sono due forme di responsabilità di secondo grado[116], perché si innestano in un fatto dannoso già perfezionato nei suoi elementi costituitivi, acuendone le conseguenze lesive e quelle pregiudizievoli, e possono essere considerate come manifestazioni di una categoria unitaria rappresentata dalla responsabilità per mancato contrasto dell'altrui fatto illecito dagli effetti dannosi permanenti[117]. In dottrina si è proposta la qualificazione in termini di illecito pluri-

---

mentre per accedere a quella inibitoria si rivela necessaria la conoscenza effettiva. L'interpretazione è ingegnosa e, per di più, sembra accordarsi a quanto avviene in altre significative esperienze giuridiche, prima fra tutte quella statunitense; e tuttavia non persuade. Dall'analisi complessiva della disciplina europea del commercio elettronico sembra emergere uno scenario nel quale al risarcimento del danno è riservato il ruolo di rimedio secondario rispetto al rimedio inibitorio, variamente articolato: una sorta di *extrema ratio* dell'*enforcement* della proprietà intellettuale e dei diritti della persona. Se così è, fermo restando il presupposto della conoscenza effettiva per poter invocare la tutela inibitoria, sembrerebbe preferibile rintracciare la soglia di ingresso della responsabilità civile in un elemento più rigoroso rispetto alla mera conoscibilità, come si è tentato di prospettare *supra* nel testo, richiedendo addirittura la conoscenza di elementi che rendano manifesto il carattere illecito dei contenuti immessi in rete. D'altro canto lo stesso R. riconosce, in maniera del tutto condivisibile, che il presupposto della conoscibilità potrebbe entrare in contrasto con l'assenza di un dovere generale di sorveglianza in capo all'intermediario*ibidem*, p. 243.

[115] Sul punto v. Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 248.

[116] È a questo concetto cui forse intende alludere, per lo meno in parte, App. Milano, sez. spec. in materia di impresa, 7 gennaio 2015, n. 29, cit., n. 31 ss. quando etichetta quella del *provider* come una «responsabilità "a posteriori"». L'espressione non è però felice, perché la responsabilità è sempre un *ex post* rispetto a un'attività oramai realizzata e foriera di pregiudizi giuridicamente rilevanti, in quanto qualificazione di un fatto dannosso che discende da una valutazione retrospettiva di un fatto storico oramai fissato.

[117] Sulla natura di fatto illecito permanente dell'illecito commesso dall'utente in rete si è di recente pronunciata Cass. pen., Sez. V, 27 dicembre 2016, n. 54946.

soggettivo eventuale a formazione progressiva dell'intermediario di caching o di hosting che, venuto a conoscenza dell'illecito, omette di identificarne l'autore e di rimuovere l'informazione lesiva degli altrui diritti[118]. Si tratterebbe di un'ipotesi di concorso successivo, sostanzialmente doloso, nell'illecito di un terzo già compiuto in tutti i suoi elementi costitutivi, che deriva da una condotta omissiva che poggia sullo stato soggettivo di conoscenza da parte dell'intermediario[119]. Proprio quest'ultima caratteristica sospingerebbe la fattispecie della responsabilità dell'internet provider al di là degli istituti conosciuti dal codice civile, il quale ignora l'illecito permanente e prende in considerazione soltanto le fattispecie di concorso di cause nella produzione dell'evento lesivo, pur in assenza di una cooperazione psicologica, visto che è ammesso che i coautori possano essere chiamati a rispondere sulla base di criteri di imputazione diversi. Nella responsabilità dell'intermediario, il concorso, se c'è, si innesta – come si è già chiarito – su un evento lesivo già perfetto. L'opinione in esame si interroga allora sulla possibilità o meno di applicare a tale forma di responsabilità l'art. 2055 c.c., considerato che, per opinione prevalente, tale disposizione presuppone il concorso di cause già in sede di produzione dell'evento dannoso e ciò la renderebbe non estendibile in via diretta a una fattispecie di illecito permanente a formazione progressiva, che vede il concorso successivo in un illecito che nasce unisoggettivo. E tuttavia il dubbio viene sciolto nel senso dell'applicabilità dell'art. 2055 c.c. sia alla luce della portata ampia della disposizione in esame sia in considerazione del rapporto biunivoco che intercorre tra il comportamento del destinatario del servizio e il comportamento dell'intermediario[120]. Si osserva, infatti, che se l'illecito commesso in rete «è, per definizione, ad intermediazione necessaria, appare evidente che l'intermediario non è un agente esterno ed estraneo al fatto telematico, poi illecito, ma è proprio l'autore della intermediazione che ha creato il presupposto necessario del fatto telematico, poi illecito. Appare, allora, evidente che, rispetto all'intermediario, il fatto del destinatario del servizio non può essere considerato totalmente estraneo. Al punto che la dottrina afferma, coerentemente, la responsabilità oggettiva dell'host intermediario proprio perché il fatto illecito è in parte a lui attribuibile per aver posto in essere, attraverso la memorizzazione, la condicio sine qua non dell'illiceità»[121].

La tesi in esame coglie meglio di altre sistemazioni alcuni profili cruciali della fattispecie della responsabilità dell'intermediario, quali il suo carattere permanente, il concorso successivo, la trasformazione di un illecito unipersonale in un illecito plurisoggettivo. Questi indubbi elementi costituitivi della fattispecie vengono però inseriti in un quadro concettuale che non persuade. In primo luogo, il fondamento del concorso dell'intermediario nell'illecito commesso dall'utente

[118] Bocchini, *La responsabilità civile degli intermediari del commercio elettronico*, cit., p. 157 ss.; Id., *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, c. 640 ss.

[119] Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 243 solleva il problema, cruciale ma anche oltremodo arduo da sciogliere, relativo alla natura della conoscenza richiesta all'intermediario: si tratta della "conoscenza" degli umani o della "conoscenza" delle macchine? Sul tema v. D.L. Burk, *Towards an Epistemology of ISP Secondary Liability*, in *24 Philosophy & Technology*, 2011, p. 437 ss.

[120] Bocchini, *La responsabilità civile degli intermediari del commercio elettronico*, cit., p. 188.

[121] Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 642.

non risiede tanto nella natura ontologicamente infrastrutturale dell'illecito realizzato in internet quanto piuttosto nel suo carattere permanente, che dal punto di vista della responsabilità – che, come si è chiarito, non è l'unica né la principale dimensione della disciplina in esame – si traduce nella reiterazione nel tempo delle conseguenze dannose. Ed è qui che si innesta il concorso colposo dell'intermediario che, a causa della propria omissione nel rimuovere i contenuti illeciti, concorre alla produzione delle conseguenze dannose a partire, a seconda della fattispecie, dal momento della comunicazione o dell'ordine da parte delle autorità competenti o dal momento dell'acquisizione degli elementi che rendono manifesta l'illiceità. La dottrina specialistica più autorevole al riguardo ha sottolineato che, mentre il diritto d'autore, il diritto dei marchi e delle relative tutele rientrano nel diritto europeo e nel diritto nazionale armonizzato a quest'ultimo, il nodo del concorso nell'illecito resta affidato, invece, ai diritti nazionali che sotto questo profilo non sono ancora armonizzati e ciò rende tale profilo di estrema delicatezza[122]. Il rilievo sembra suggerire un approccio al tema più audace, così da non lasciarsi imbrigliare in letture tradizionali, elaborate in epoche e contesti molto distanti dall'attuale stagione degli illeciti digitali.

Ecco allora che, cercando di cogliere l'implicito invito, bisognerebbe tenere conto che l'art. 2055 c.c., nonostante il suo infelice riferimento al fatto dannoso[123], che in effetti suggerisce la circoscrizione della norma al concorso nell'evento lesivo, va coordinato con l'art. 1227, comma 1, c.c., espressamente richiamato in materia aquiliana dall'art. 2056, comma 1, c.c.[124]. Com'è noto, l'art. 1227, comma 1, c.c. esprime, in un contesto dedicato alla determinazione e alla liquidazione del danno, una regola causale e la collocazione non del tutto appropriata si comprende in considerazione del fatto che il tipo di concorso qui regolato riguarda proprio le conseguenze pregiudizievoli dell'evento dannoso. Se si volessero adottare le categorie della giurisprudenza[125] e di un'area sin troppo vasta della dottrina, si dovrebbe sostenere che l'art. 1227, comma 1,

[122] Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., pp. 237-238.

[123] Sul senso e la portata dell'art. 2055 c.c. cfr. C.M. Bianca, *Diritto civile. 5. La responsabilità*^{2}, Milano, 2012, p. 648 ss.; M. Franzoni, *Dei fatti illeciti*, in *Comm. cod. civ. Scialoja-Branca*, a cura di F. Galgano, Bologna-Roma, 1993, p. 713 ss.; Id., *L'illecito*^{2}, in *Tratt. resp. civ.*, diretto da M. Franzoni, I, Milano, 2010, p. 130 ss.; M. Orlandi, *La responsabilità solidale*. Profili delle obbligazioni solidali risarcitorie, Milano, 1993, p. 101 ss.; P.G. Monateri, *La responsabilità civile*, in *Tratt. dir. civ.*, diretto da R. Sacco, Torino, 1998, p. 192 ss.; A. Gnani, *La responsabilità solidale*, in *Il Codice Civile. Comm.* fondato da P. Schlesinger, diretto da F.D. Busnelli, Milano, 2005, *passim*, in part. p. 127 ss.; S. Marullo di Condojanni, sub *Art. 2055 - Responsabilità solidale*, in *Comm. cod. civ.*, diretto da E. Gabrielli, *Dei fatti illeciti*, a cura di U. Carnevali, artt. 2044-2059, Torino, 2011, p. 408 ss.; Id., *Il concorso di colpa nell'illecito civile*, Milano, 2012, *passim*.

[124] Sul punto cfr. N. Di Prisco, *Concorso di colpa e responsabilità civile*, Napoli, 1973, *passim* e, più di recente, G. Grisi, *Causalità materiale, causalità giuridica e concorso del creditore nella produzione del danno*, in *Contratti*, 2010, p. 617 ss.; B. Tassone, sub *Art. 1227 - Concorso del fatto colposo del creditore*, in *Comm. cod. civ.*, diretto da E. Gabrielli, *Delle obbligazioni*, a cura di V. Cuffaro, Artt. 1218-1276, Torino, 2013, p. 285 ss.; V. Caredda, *Concorso del fatto colposo del creditore*, in *Il Codice civile. Comm.* fondato da P. Schlesinger, diretto da F.D. Busnelli, Milano, 2015, *passim*, in part. p. 23 ss.; D. Farace, *Sul concorso colposo dei soggetti lesi*, in *Riv. dir. civ.*, 2015, p. 158 ss.

[125] Per una convinta adesione alla distinzione tra causalità materiale e causalità giuridica cfr. G. Travaglino, *La questione dei nessi di causa*, Milano, 2013, *passim*, in part. p.133 ss.

c.c. disciplina le concause che incidono sul nesso di causalità giuridica di cui all'art. 1223 c.c., ossia sul rapporto eziologico che collega le conseguenze dannose al fatto lesivo, il quale, a sua volta, è collegato alla condotta dannosa dal nesso di causalità materiale di cui agli artt. 40 e 41 c.p. L'armamentario concettuale che distingue due nessi causali si rivela però – come si è tentato di dimostrare anche in passato[126] – alquanto debole: il nesso causale è, infatti, unico[127] e non conosce frammentazione in quanto presuppone un rapporto di derivazione tra una condotta umana e le conseguenze pregiudizievoli da essa provocate, sicché l'art. 1223 c.c., a dispetto del linguaggio causale adottato, non integra una regola sulla causalità giuridica, ma una disposizione che enuncia un criterio di delimitazione del danno risarcibile che opera lungo l'unitaria traiettoria causale[128]. La causalità fonda un giudizio che istituisce un ordine di valutazioni giuridiche, fondate però su regole pre-giuridiche, desunte dai più svariati campi della conoscenza, grazie alle quali è possibile ricostruire quel rapporto di derivazione di un evento dalla sua causa che il diritto non sarebbe in grado di esprimere in piena autonomia[129]. Il diritto può privilegiare talune di queste regole causali pre-giuridiche, le può addirittura rielaborare, ma non ne può prescindere[130]. Non ha quindi molto costrutto immaginare due distinti nessi di causalità, uno materiale governato dalle leggi scientifiche di copertura, ossia da regole extra-giuridiche, e uno giuridico, governato, invece, più specificamente da criteri di valutazione interni al diritto.

Una volta chiarito questo aspetto concettuale, non sembra affatto che la logica interna del giudizio causale imponga che il concorso di cause operi soltanto in fase originaria, atteggiandosi inevitabilmente come la compartecipazione nell'azione che provoca la lesione e dà il la alla catena dei pregiudizi. L'apporto causale di una o più condotte ulteriori può, infatti, innestarsi in una struttura causale già instaurata e, dunque, in itinere, incidendo direttamente sull'entità delle conseguenze dannose. È proprio questa l'ipotesi che si verifica nella responsabilità degli intermediari, nella quale l'omissione dell'obbligo di rimuovere i contenuti illeciti immessi dall'utente si affianca alla condotta lesiva di quest'ultimo, concorrendo alla rinnovazione o al prolungamento nel tempo degli effetti dannosi scaturiti dalla violazione della posizione sostantiva altrui. Ancora meno persuasivo è il rilievo secondo cui l'incidenza causale della condotta dell'intermediario nella produzione del danno derivante dalla violazione dell'utente spiegherebbe la ragione – non condivisa tuttavia dell'opinione in esame[131] – per cui una parte della dottrina propone di ricondurre la responsabilità dell'intermediario nell'alveo della responsabilità oggettiva: il provider, infatti, crea con la memorizzazione delle informazioni immesse dall'utente la condicio sine qua

[126] F. Piraino, *I confini della responsabilità civile e la controversia sulle malformazioni genetiche del nascituro: il rifiuto del c.d. danno da vita indesiderata*, in *NGCC*, 2016, I, p. 450 ss., in part. p. 457 ss.

[127] In tal senso C. Castronovo, *Il risarcimento del danno*, in *Riv. dir. civ.*, 2006, p. 86 ss., in part. p. 89.

[128] Così – com' è noto – G. Gorla, *Sulla cosiddetta causalità giuridica: «fatto dannoso e conseguenze»*, in *Riv. dir. comm.*, 1951, p. 407 ss. e ora Castronovo, *Il risarcimento del danno*, cit., p. 87 ss.

[129] Cfr. A. Belvedere, *Causalità giuridica?*, in *Riv. dir. civ.*, 2006, I, pp. 8-10.

[130] R. Pucella, *La causalità «incerta»*, Torino, 2007, p. 40 ss.

[131] In un passaggio successivo del proprio ragionamento Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 643 esclude che quella dell'intermediario possa essere qualificata come una responsabilità oggettiva, poiché anzi essa si presenta come un regime premiale per il *provider*.

non dell'illecito di quest'ultimo[132]. Al di là delle plurime evidenze del fondamento colposo di tale forma di responsabilità in precedenza segnalate, non va mai dimenticato che la rappresentazione della responsabilità oggettiva come una tipologia di responsabilità per pura causalità è fieramente avversata da tutti gli specialisti della materia e la ragione, peraltro del tutto fondata, risiede nella circostanza che il nesso causale offre alla fattispecie di responsabilità la struttura ma nulla dice in ordine all'imputazione delle conseguenze dannose[133], la quale, anche quando non si identifichi con la colpa, deve poggiare su un dispositivo che esprima la scelta di valore che soggiace all'addebitamento dei danni a quel soggetto piuttosto che a un altro.

Se si volesse tentare un inquadramento di tale particolare fattispecie di concorso nelle categorie concettuali di common law, sembrerebbe confacente il richiamo alla Secondary Liability: una figura di responsabilità piuttosto ampia che abbraccia diverse ipotesi tutte però accomunate dall'esistenza di una condotta illecita primaria e dalla circostanza che un soggetto viene chiamato a rispondere del danno provocato da tale azione dannosa[134]. Si è soliti distinguere due forme di Secondary Liability: una Participant-based e una Relationship-based[135]. La prima presuppone una compartecipazione del responsabile secondario all'illecito dell'autore principale sotto forma di induzione, di concorso nella causazione o di facilitazione e il presupposto soggettivo di tale forma di responsabilità esige la conoscenza, o per lo meno la conoscibilità, del carattere antigiuridico della condotta dell'autore primario. In questa tipologia rientra la figura della Contributory Liability[136], che abbraccia le ipotesi di concorso, che, per opinione consolidata, non sono circoscritte alla compartecipazione nell'evento dannoso ma estese anche alle fattispecie in cui il contributo causale si esaurisce nella predisposizione dello strumento mediante il quale l'illecito viene compiuto. La Secondary Liability Relationship-based presuppone, invece, che il responsabile secondario tragga beneficio dall'illecito primario in forza della stretta relazione intercorrente con l'autore di quest'ultimo, tale da poterne prevenire l'azione dannosa: una prossimità che, dal punto di vista giuridico, giustifica il loro accomunamento sul piano della responsabilità[137]. Rientra in questo modello la Vicarious Liability, ossia la responsabilità per fatto altrui. La struttura della responsabilità degli intermediari delineata dalla legislazione

[132] Bocchini, *op. cit.*, c. 642.

[133] Cfr. Castronovo, *La nuova responsabilità civile*, cit., p. 302, pp. 337-338, il quale spiega che « pure nelle fattispecie di responsabilità oggettiva il rapporto di causalità non può essere lasciato a decidere da solo della responsabilità. La responsabilità oggettiva perciò non può essere pura assenza o irrilevanza dei criteri soggettivi di imputazione, bensì sostituzione di questi con altri di natura oggettiva, i quali svolgano nei confronti del rapporto di causalità la medesima funzione che da sempre è propria dei criteri soggettivi di imputazione nei fatti illeciti. Solo che, mentre nella responsabilità per colpa quest'ultima si asside su un nesso causale già individuato tra evento e fatto ai fini della qualificazione di quest'ultimo in funzione della responsabilità, nella responsabilità oggettiva sono i criteri di imputazione a individuare la sequenza causale alla quale occorre fare riferimento ai fini della responsabilità».

[134] G.B. Dinwoodie, *A Comparative Analysis of the Secondary Liability of Online Service Providers*, in *Secondary Liability of Internet Service Providers*, cit., p. 1 ss., in part. p. 10

[135] Sul punto cfr. G.B. Dinwoodie, *op. cit.*, p. 9.

[136] Dinwoodie, *op. cit.*, pp. 8-9.

[137] Dinwoodie, *op. cit.*, p. 9.

europea sembra iscriversi nella prima variante della Secondary Liability e non è infrequente che nelle trattazioni del tema si proponga l'inquadramento nel Contributory Infringement[138], che è una figura elaborata prevalentemente nell'ambito della tutela del brevetto e riconducibile al ceppo della Contributory Liability. Si ha Contributory Infringement quando la violazione dei diritti di utilizzazione economica derivanti da un brevetto o da un modello di utilità avviene mediante la predisposizione o l'offerta di predisposizione a favore di un soggetto diverso dal titolare del diritto di sfruttamento, e senza il consenso di questi, di mezzi relativi a una parte essenziale dell'invenzione, di per sé non coperta da brevetto, e univocamente finalizzati all'attuazione dell'invenzione medesima, con la consapevolezza della preordinazione del mezzo, o per lo meno della sua idoneità, a realizzare la contraffazione del trovato protetto dal diritto di privativa. Nel diritto statunitense il Contributory Infringement è delineato dal § 271c del Title 35 del United States Code[139] come un rimedio aggiuntivo a favore del titolare del brevetto e, nell'interpretazione fornitane dalla Corte Suprema[140], esso sussiste quando la violazione realizzata dal Contributory Infringer si riveli in grado di intaccare il mercato del trovato. Pur in assenza di una disposizione espressa, anche il diritto italiano conosce il Contributory Infringement, noto anche come contraffazione indiretta, con l'avallo della giurisprudenza di legittimità sin dal 1956[141].

Nonostante la frequenza del ricorso del Contributory Infringement per fornire le coordinate della posizione dell'intermediario, si può nutrire più di un dubbio sull'idoneità di tale figura a offrire un inquadramento fedele. Il Contributory Infringement si presta al limite a qualificare la posizione dell'intermediario nelle ipotesi di vera e propria compartecipazione di quest'ultimo, o del gestore del sito, nella violazione del diritto di utilizzazione realizzata dall'utente, mettendogli a disposizione gli strumenti tecnici indispensabili a compiere lo sfruttamento illecito dell'opera protetta o dell'invenzione[142]. La figura diviene inadatta, invece, a fornire una cornice concettuale adeguata alla posizione dell'intermediario che concorre alla violazione dell'utente tramite l'omessa rimozione dei contenuti illeciti a seguito dell'acquisizione degli elementi da cui

[138] Bellia-Bellomo-Mazzoncini, *La responsabilità civile dell'*Internet Service Provider *per violazioni del diritto d'autore*, cit., p. 357 ss.

[139] 35 U.S.C. § 271c: «Whoever offers to sell or sells within the United States or imports into the United States a component of a patented machine, manufacture, combination, or composition, or a material or apparatus for use in practicing a patented process, constituting a material part of the invention, knowing the same to be especially made or especially adapted for use in an infringement of such patent, and not a staple article or commodity of commerce suitable for substantial noninfringing use, shall be liable as a contributory infringer».

[140] U.S. Supreme Court, *Dawson Chem. Co.* v. *Rohm & Haas Co.*, 448 U.S. 1761980.

[141] Cass., 24 ottobre 1956, n. 3387, in *Riv. dir. ind.*, 1958, II, p. 3 ss., .

[142] Un'ipotesi di tal genere sembra quella decisa da Trib. Milano, sez. spec. in proprietà intellettuale ed industriale, 7 gennaio 2010, *Sky Italia S.r.l.* c. *Davide Boizza e Telecom Italia S.p.A.*, il quale ha riconosciuto la responsabilità del gestore del sito internet a causa dell'agevolazione alla visualizzazione dei filmati di partite del campionato di calcio coperte dai diritti di utilizzazione economica in capo a Sky realizzata non solo mediante la messa a disposizione di link a siti cinesi, ma anche grazie a dettagliate istruzioni per ottimizzare la visualizzazione, coordinando immagini e audio. Di contro, l'*internet service provider* è stato ritenuto esente da responsabilità alla luce della neutralità della sua prestazione, sostanzialmente limitata a fornire al sito Tvgratis il solo accesso a internet, ossia a un'attività di *mere conduit*, di cui all'art. 14 d.lgs. 70/2003.

emerga la loro manifesta illiceità oppure a seguito della comunicazione o dell'ordine delle autorità competenti. Nella concettualizzazione sinora prevalente, il Contributory Infringement si riferisce, infatti, a ipotesi di concorso nella commissione dell'evento dannoso mediante la messa a disposizione di una parte del trovato non coperta da brevetto ma finalizzata in maniera univoca all'utilizzazione dell'invenzione con consapevolezza da parte del realizzatore della parte. Nei termini qui suggeriti di una responsabilità per mancato contrasto dell'altrui fatto illecito dagli effetti dannosi permanenti, la posizione dell'intermediario concorre non già alla causazione dell'evento dannoso, ma alla protrazione nel tempo o all'accentuazione dei pregiudizi prodotti dalla violazione della situazione giuridica soggettiva del terzo, il che imprime alla fattispecie una connotazione del tutto peculiare.

Quel che va certamente condiviso nella costruzione dell'illecito plurisoggettivo eventuale a formazione progressiva è la soluzione offerta al problema della determinazione delle quote dell'obbligazione risarcitoria spettanti su ciascun concorrente ex art. 2055 c.c. La quota va determinata in misura della gravità della colpa di ciascun debitore in solido e dell'entità delle conseguenze che ne sono derivate e nella fattispecie del concorso dell'intermediario nell'illecito dell'utente tale operazione deve muovere dall'individuazione del momento a partire dal quale l'obbligo di rimozione dei contenuti illeciti è divenuto esigibile, perché è da qui che la condotta colposa dell'intermediario confluisce nella condotta lesiva dell'utente. È «lo scarto che separa l'ingresso nell'illecito dell'intermediario rispetto all'inizio dello stesso, che segna la misura della colpa dell'intermediario rispetto a quella del destinatario del servizio. Quanto maggiore sarà lo scarto temporale, allora, tanto minore sarà la misura della colpa dell'intermediario ai sensi dell'art. 2055, 2° comma, c.c.»[143].

6. È evidente che il profilo più delicato nell'ipotesi costruttiva qui delineata è rappresentato dal ruolo e dal valore da riconoscere alla segnalazione di colui che si dichiara titolare della situazione soggettiva lesa e la soluzione non può essere univoca: come non si può sostenere che la segnalazione della presunta vittima dell'illecito sia sempre idonea a svelare all'hosting provider elementi che rendono manifesta l'illiceità, così non si può neppure ritenere che essa sia per definizione inidonea a rendere manifesta la violazione di cui si lamenta il segnalante. È il concreto contenuto dell'avviso o della diffida a determinare se la segnalazione del presunto titolare del diritto violato integri la prima fattispeciequella che implica la segnalazione dell'illiceità alle autorità competenti da parte dell'intermediario oppure la secondaquella che comporta direttamente l'obbligo di eliminazione dei contenuti illeciti in capo all'intermediario[144]. E, d'altro canto, la conoscenza di un'illiceità che si presenta manifesta costituisce un elemento di natura congiunturale, in quanto legato ai caratteri specifici della violazione e dalla completezza della

[143] Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 643.

[144] Questa impostazione dovrebbe allentare le resistenze di chi si mostra contrario a riconoscere alla segnalazione della presunta vittima del valore giuridico di strumento di conoscenza qualificata idonea a far scattare a carico dell'intermediario l'immediato obbligo di cancellazione deli contenuti illeciti: in tal senso Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., par. 5.

segnalazione[145]. Va, peraltro, sottolineata la complessità della valutazione sul carattere illecito o meno dei contenuti immessi in rete, tale da aver indotto la dottrina a parlare di un apprezzamento non solo congiunturale, in quanto fortemente condizionato da elementi di contesto, ma anche multifattoriale, il che dovrebbe suggerire maggiore cautela nel caricare eccessivamente di contenuti il dovere di diligenza professionale degli intermediari in punto di acquisizione degli elementi dai quali ricavare la conoscenza della natura illecita delle informazioni memorizzate[146]. E, tuttavia, anche se inidonea a far emergere la natura manifesta dell'illiceità, la segnalazione da parte della presunta vittima dell'illecito riveste comunque un valore giuridico significativo, consentendo che si integri la prima fattispecie di cui all'art. 16, comma 1, d.lgs. 70/2003 e gettando così le basi per l'instaurazione di un meccanismo di notice-and-take down sul modello di quello statunitense, che però necessiterebbe di un intervento normativo specifico anche soltanto di natura regolamentare e che – andrebbe riconosciuto – non è detto che si riveli una panacea[147]. Un tentativo in tal senso è stato compiuto dall'Autorità per le Garanzie nelle Comunicazioni[148], per lo meno nel campo che qui maggiormente interessa, ossia quello della violazione online della proprietà intellettuale, con l'approvazione del Regolamento in materia di tutela del diritto d'autore sulle reti di comunicazione elettronica e procedure attuative del d.lgs. 9 aprile 2003, n. 70, delibera del 12 dicembre 2013[149]. Entrato in vigore il 31 marzo 2014, il regolamento disciplina la procedura dinanzi all'AGCOM esperibile da chi ritenga che un'opera digitale sia stata resa disponile su una pagina internet in violazione della legge sul diritto d'autore. L'art. 6 disciplina l'istanza e le condizioni di ricevibilità, procedibilità e ammissibilità. L'art. 7 regola il procedimento istruttorio dell'Autorità, che esordisce con la comunicazione dell'avvio del procedimento ai prestatori dei servizi di mere conduit o di hosting e, ove rintracciabili, all'uploader e ai gestori della pagina e del sito internet. L'art. 8 offre il ventaglio dei provvedimenti adottabili dall'Autorità che si traducono in ordini che, nel rispetto dei criteri di gradualità, di proporziona-

[145] E proprio per questo motivo ha perfettamente ragione Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 244 quando osserva che la disciplina degli intermediari «è basata su di un presupposto - la possibilità per il *provider* di distinguere con ragionevole grado di sicurezzae con costi non proibitivi fra legalità ed illegalità, liceità ed illiceità, dell'attività e dell'informazione - che è molto dubbio».

[146] Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 244 ss., il quale sottolinea che la complessità della valutazione della natura dei contenuti si accentua quanto la violazione concerne il marchio, a causa degli accertamenti fattuali che il *provider* dovrebbe essere chiamato in molti frangenti a compiere.

[147] Ne dubita, e con molto fondamento, Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., pp. 246-247.

[148] Non ne tiene conto Petruso, *Responsabilità degli intermediari di internet e nuovi obblighi di conformazione:* robo-takedown*,* policy of termination*,* notice and take steps, cit., pp. 491-492, il quale, quindi, ritiene che nel disciplina europea degli intermediari non vi sia traccia «financo di un sistema di *notice and takedown*a beneficio di chi sia stato danneggiato dalla presenza in rete dei più vari contenuti anche solo parzialmente regolamentato: qualcosa di solo lontanamente paragonabile ad un sistema di notifica e di rimozione quale quello disciplinato organicamente dal *Digital Millennium Copyright Act*, infatti, è solo indirettamente tratteggiato dall'art. 14 lett. *b* nel momento sanzionatorio della riattivazione del circuito della responsbailità».

[149] Delibera n. 680/13/CONS, sul quale v. Aa.Vv., *Il regolamento Agcom sul diritto d'autore*, a cura di L.C. Ubertazzi, Torino, 2014, *passim*.

lità e di adeguatezza, impongono agli intermediari di impedire la violazione del diritto d'autore o di porvi fine: ordini l'inottemperanza ai quali comporta l'applicazione da parte dell'Autorità delle sanzioni di cui all'art. 1, comma 31, l. 31 luglio 1997, n. 249. L'art. 9 introduce, infine, un procedimento abbreviato, applicabile quando la Direzione dei servizi media dell'Autorità, sulla base di una sommaria cognizione dei fatti oggetto dell'istanza, vi ravvisa un'ipotesi di grave lesione dei diritti di sfruttamento economico di un'opera digitale oppure un'ipotesi di violazione di carattere massimo. Il regolamento ha incontrato molta resistenza: impugnato dinanzi al Tar Lazio[150] e poi da questi sottoposto al giudizio della Corte costituzionale nell'ambito di una questione di legittimità costituzionale relativa agli art. 5, comma 1; 14, comma 3; 15, comma 2, e 16, comma 3, d.lgs. 70/2003 nonché all'art. 32-bis, comma 3, d.lgs., 31 luglio 2005, n. 177, testo unico dei servizi di media audiovisivi e radiofonici, così come modificato dal d.lgs., 15 marzo 2010, n 44[151]. Dichiarata inammissibile la questione di legittimità costituzionale[152], non si sono però assopite le perplessità sul regolamento di cui la dottrina specialistica tenta di circoscrivere il campo di applicazione dal punto di vista tanto soggettivo quanto oggettivo alla luce della portata della norma primaria che offre il fondamento al potere normativo in materia dell'AGCOM, il già ricordato art. 32-bis, comma 3, d.lgs. 177/2005[153]. Sotto il primo profilo, la normativa secondaria si dovrebbe applicare ai soli fornitori di servizi di media audiovisivi e radiofonici ai quali sia riconducibile la responsabilità editoriale sulla scelta dei contenuti trasmessi; mentre sotto il secondo profilo, il regolamento dovrebbe riguardare soltanto i contenuti audiovisivi e radiofonici e non anche quindi file musicali, immagini non in movimento, testi e software[154]. Non si può però dimenticare che, ancora più a monte, viene revocata in dubbio l'esistenza di un'adeguata copertura della normativa primaria alla pretesa dell'AGCOM di assumere poteri di vigilanza e ispettivi in ordine alla violazione del diritto d'autore compiute nella rete, che – com'è noto – alcuni fondano su una lettura giudicata particolarmente largheggiante dell'art. 182-bis l.d.a.[155] mentre altri sull'art. 32-bis, comma 3, d.lgs. 177/2005[156]. Quel che maggiormente si

[150] Tar Lazio, Sez. I., ord., 26 settembre 2014 n. 10020 e 10016, su cui cfr. P. Passaglia, *Corte costituzionale e diritto dell'Internet: un rapporto difficilee un appuntamento da non mancare*, in *Giur. Cost.*, 2014, p. 4857 ss.

[151] Corte cost., 3 dicembre 2015, n. 247.

[152] Le ragioni a sostegno della sentenza di inammissibilità affondano nei vizi di contraddittorietà, ambiguità e oscurità della motivazione e del *petitum*: in particolare, la Consulta reputa che l'ordinanza non chiarisca in maniera adeguata se il provvedimento richiesto sia una pronuncia ablativa o una pronuncia ablativo-modificativa.

[153] Per un approfondimento cfr. P. Costanzo, *Quale tutela del diritto d'autore in internet?*, in *Giust. cost.*, 2015, p. 2343 ss.

[154] Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., par. 9.

[155] Assai critico al riguardo D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p. 57 ss. Non sembrano ravvisare dubbi sulla competenza dell'AGCOM Bertoni-Montagnani, *La modernizzazione del diritto d'autore e il ruolo degli intermediari internet quali propulsori delle attività creative in rete*, cit., p. 144.

[156] E. Rosati-G. Sartor, *Social Networks e Responsabilità del Provider*, in *EUI Working Papers Law*, 2012, p. 15.

contesta è l'esercizio da parte dell'AGCOM del potere di irrogare ordini inibitori, muniti di sanzioni in caso di mancato rispetto da parte del destinatario[157].

Nel quadro della rilettura della posizione giuridica degli *internet service provider* emerge con maggiore precisione la portata dell'assenza del dovere generale e generico di sorveglianza in campo a questi e di ricerca attiva di condotte illecite, prescritta dall'art. 17 d.lgs. 70/2003. Escluse tanto l'assunzione da parte dell'intermediario di una posizione di garanzia quanto la sussistenza in capo a questi di un dovere di controllo preventivo sulle informazioni e sui contenuti immessi dagli utenti, va riconosciuta, invece, l'insorgere sull'intermediario di un obbligo particolareggiato di eliminare gli effetti della violazione successivo o a una comunicazione in tal senso delle autorità competenti o all'acquisita conoscenza della manifesta illiceità dei specifici contenuti o informazioni. L'art. 17 d.lgs. 70/2003 non si presenta in alcun modo ostativo alla configurazione di un obbligo successivo di cancellazione dei contenuti frutto della violazione di altrui diritti né al riconoscimento di un obbligo di intervento immediato in presenza di una violazione manifesta e in tal senso è condivisibile l'orientamento che si sta andando consolidando nella giurisprudenza di merito, anche meneghina[158], disattendendo le difese dei provider fondate proprio sull'attitudine preclusiva dell'art. 17 alla configurazione di un loro obbligo di intervento se non a seguito di un ordine giudiziale o dell'autorità amministrativa.

La configurazione di due distinte fattispecie, l'una fondata sulla conoscenza semplice dell'illecito commesso dall'utente e l'altra fondata sulla conoscenza di elementi che rendono manifesto tale illecito, disinnesca una delle obiezioni più frequenti al riconoscimento di un ruolo attivo all'intermediario: quella dell'inopportunità di delegare a quest'ultimo il giudizio di valore sull'illiceità o meno di determinati contenuti, come se questi debba operare da "sentinella" della rete: ecco perché si è suggerito di etichettare questo scenario non certo auspicabile come "responsabilità giuridica dei sicofanti". L'obbligo di intervento attivo, mediante rimozione selettiva dei contenuti frutto della violazione degli altrui diritti, sorge in prima battuta soltanto nel caso di conclamata illiceità, senza che il provider sia, quindi, costretto a compiere indagini

[157] D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., pp. 60-62 invoca come impedimenti all'esercizio della tutela in via amministrativa da parte dell'AGCOM in assenza di una disposizione primaria di delegazione espressa gli artt. 23, 97 e 113 Cost.; l'art. 182-*ter* l.d.a., che, in caso di accertamento della violazione delle norme del diritto d'autore, impone agli ispettori di compilare processo verbale da trasmettere senza meno agli organi di polizia giudiziaria per il compimento degli atti di cui all'art. 347 c.p.p.; l'art. 1, commi 5 e 6, l. 21 maggio 2004, n. 128, di conversione del d.l. 22 marzo 2004 n. 72 che reca interventi per contrastare la diffusione telematica abusiva di materiale audiovisivo, il quale fa esclusivo riferimento all'autorità giudiziaria. D'A. paventa che, in tal modo, venga esautorato il Parlamento, si determini un'eccessiva concentrazione di potere in capo all'Autorità, si registri una sostanziale espropriazione delle prerogative delle sezioni specializzate dall'autorità giudiziaria, si compia una disparità di trattamento dei titolari di diritti d'autore rispetto ai titolari di altri diritti, specie non patrimoniali, esposi a violazioni non meno frequenti in rete. Quest'ultimo profilo si presenta ancora più odioso se si pensa che la disciplina sul commercio elettronico, che rappresenta il blocco normativo più rilevante della disciplina degli intermediari, ha una portata generale, applicandosi trasversalmente a qualsiasi illecito commesso in rete, senza prevedere trattamenti preferenziali per i titolari di diritti di proprietà intellettuale.

[158] Trib. Milano, ord., 3 ottobre 2013; Trib. Napoli Nord, 3 novembre 2016, in *Giur. it.*, 2017, c. 629 ss., con nota di R. Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti.*

e accertamenti sulla natura delle informazioni memorizzate, in effetti inusuali a carico di un imprenditore nel settore dei servizi. Quando la violazione delle altrui prerogative non si presenta così lampante, l'intermediario non è per l'appunto chiamato a verifiche improprie, ma deve limitarsi a segnalare il presunto illecito all'autorità giudiziaria o a quella amministrativa e, soltanto a seguito dei loro accertamenti positivi, dovrà provvedere alla cancellazione o alla disabilitazione dell'accesso. La responsabilità per mancato contrasto dell'altrui fatto illecito dagli effetti dannosi permanenti qui delineata bilancia i vantaggi del capitalismo estrattivo tipico delle attività economiche di servizio di intermediazione nelle comunicazioni a distanza con il perseguimento di una forte tutela per i diritti dei terzi, non ultime le prerogative autorali violate nella rete, alternando il dovere generale di intervento in via autonoma con l'obbligo specifico di cooperazione con le autorità competenti per la reintegrazione, la salvaguardia e la prevenzione di specifiche situazioni giuridiche violate.

La tesi qui proposta presenta, inoltre, il vantaggio, certo non marginale, di rendere sostanzialmente irrilevante il problema della responsabilità dell'intermediario per la rimozione di contenuti segnati come illeciti e poi rivelatisi rispettosi delle altrui prerogative[159]: quel problema che il DMCA risolve addossando all'autore della segnalazione la responsabilità nei confronti di colui che ha immesso i contenuti contestati e rimossi per le eventuali conseguenze dannose provocate dalla cancellazione o dalla disabilitazione dell'accesso. Nei termini suggeriti, il diritto italiano impone all'intermediario l'immediata rimozione soltanto quando la loro origine illecita si presenti manifesta e, se anche – in casi limite – tale apparenza dovesse apparire infondata a un esame più approfondito, l'intermediario resterebbe esente dalle doglianze risarcitorie da parte dell'autore dei contenuti rimossi in quanto protetto dalla sfera di immunità stesa dall'art. 16, comma 1, lett. a d.lgs. 70/2003. Non è, dunque, neppure necessario ricorrere all'escamotage di distinguere tra violazione dei diritti patrimoniali e dei diritti della persona, immaginando che la rimozione in via autonoma dei contenuti da parte del provider operi soltanto in presenza di elementi che inducono a ritenete violati i secondi, mentre si arresti in presenza della lamentata violazione dei primi, proprio a causa del rischio di determinare il capo all'utente dei contenuti contestati un danno nell'eventualità che la segnalazione del presunto titolare o l'apparenza di illiceità si rivelino infondate[160]. La diversificazione non si giustifica né in base al diritto positivo né sulla scorta dei principî né ancora in ragione della coerenza del sistema e, per di più, può essere contestata anche l'idea che la rimozione di contenuti che appaiano lesivi di altrui diritti

[159] Sotto questo profilo, appare troppo rinunciataria la lettura di Tosi, *Responsabilità civile per il fatto illecito degli* Internet Service Provider *tra tipizzazione normativa ed evoluzione tecnologica: peculiarità e criticità del regime applicabile alle nuove figure soggettive dei motori di ricerca,* social network *e aggregatori di contenuti di terzi*, cit., par. 5, il quale rinuncia a setacciate le potenzialità ermeneutiche della disciplina italiana, preferendo auspicare un intervento normativo sul modello del DMCA.

[160] Così Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 637, il quale ritiene che occorra «evitare di trasformare il *provider*, che è un imprenditore che agisce secondo le regole del mercato, in un primo giudice dei conflitti che si creino sulla piattaforma. È ben diversa, infatti, la circostanza che rischiare la lesione sua un diritto personalissimoonore, decoro, riservatezza da quella che sia un diritto patrimoniale, quale può essere il diritto d'autore, nell'ipotesi in cui la segnalazione sia infondata e si finisca, dunque, per rimuovere un contenuto *aufond* del tutto lecito».

della persona da parte dell'intermediario non esponga quest'ultimo al rischio di un'azione risarcitoria ad opera di colui che li ha immessi per il danno morale conseguente alla violazione del proprio diritto di espressione del pensiero. Si percepisce un anelito alla giustizia nella posizione che distingue tra diritti patrimoniali e non in considerazione del fatto che i primi, anche se non protetti non appena l'internet provider venga a conoscenza della lesione nell'attesa di un provvedimento dell'autorità, possono pur sempre trovare adeguata protezione in sede risarcitoria proprio alla luce del loro contenuto economico, mentre i diritti della persona non possono ricevere una tutela successiva altrettanto adeguata a causa della loro natura di diritti di natura non economica[161]. Ciononostante, la via corretta da percorrere pare un'altra: quella che collega l'insorgere immediato dell'obbligo di cancellazione o di disabilitazione al carattere manifesto dell'illiceità, quale che sia la natura del diritto leso.

7. Questa rilettura della disciplina degli *internet service provider* consente anche di abbandonare la distinzione pretoria tra ISP attivi e passivi[162], che si è rivelata molto utile per impedire a taluni operatori della rete di trincerarsi dietro l'esclusione di responsabilità prevista dalla normativa europea, invocando la propria qualificazione di hosting provider, e per far mergere l'estrema varietà delle prestazioni rese dagli intermediari e delle possibili situazioni lesive. Tuttavia, la distinzione manifesta una certa rozzezza[163], perché nasce da un processo di soggettivizzazione, che è assai frequente nelle fasi iniziali di inquadramento normativo di fenomeni nuovi, e che pretende di concettualizzare e tipizzare in chiave soggettiva elementi di fattispecie che invece sono situazionali. Per di più - e si è avuto modo di constatarlo - la coppia intermediari attivi-intermediari passivi si è rivelata alla prova dei fatti anche fomite di incertezze applicative. Nello specifico ambito della disciplina degli intermediari, la mossa della diversificazione degli scenari normativi su base soggettiva comporta un'interpretazione forse eccessivamente restrittiva dell'art. 16 d.lgs. 70/2003 da parte della giurisprudenza, la quale, pur di restringere la portata della norma, considera la prestazione di hosting che ne forma oggetto sostanzialmente circoscritta al mero immagazzinamento di informazionistorage in tal modo forzando la realtà socio-economica di riferimento, in quanto l'attività di hosting non si atteggia mai in senso autenticamente passivo, includendo sempre soluzioni tecniche per la ricerca dei dati e la loro sistematica organizzazione[164]. Quel che la giurisprudenza, con notevole

[161] Bocchini, *op. cit.*, c. 639.

[162] Nella medesima direzione si muove Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 637 ss., in part. 639, il quale correttamente osserva che «non è tanto l'attività peculiare e ulteriore, rispetto alla mera resa del servizio di *hosting* che qualifica il *provider* come attivo, ma l'obbligo che scatta legalmente una volta che si acquisisca la conoscenza del contenuto illecito presente sui server e dal quale trae anche profitto. La normativa non opera una differenziazione soggettiva del suo ambito di applicazione, ma semplicemente impone un'ulteriore attivitàed è in questo momento che l'*hosting* diventa attivo successivamente alla venuta a conoscenza dell'illecito. E questi risultati, evidentemente sono desumibili non già da una lettura evolutiva del Considerando n. 42 della direttiva, ma dal regime di responsabilità così come previsto dalla disciplina ed illustrato in precedenza».

[163] Critico nei confronti della categoria dell'*hosting attivo* anche D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p. 80 ss.

[164] Cfr. A. Mantelero, *Caso RTI c. Yahoo: il Tribunale di Milano insiste sull'idea dell'hosting attivo. Repetita*

intuito e sensibilità[165], ha concepito in chiave soggettiva, elaborando la figura dell'intermediario attivo, va invece in maniera più proficua e rigorosa rielaborato in chiave oggettiva, distinguendo tra fattispecie concrete nelle quali gli intermediari - quale che sia la natura e il tipo di servizi erogati - sono a conoscenza, o avrebbero potuto essere a conoscenza adottando la diligenza professionale, di elementi di fatto che rivelano in modo manifesto la commissione di un illecito da parte di loro utenti e fattispecie concrete nelle quali una tale evidenza non vi sia. Sia chiaro: è ben possibile prefigurare una tipizzazione delle situazioni nelle quali l'illiceità appare manifesta e quelle in cui è ragionevole ritenere che l'intermediario sia a conoscenza dell'illecito e si può procedere in tal senso muovendo dalle caratteristiche della prestazione dell'intermediario, a condizione che non si dimentichi che la chiave di volta della disciplina rimane la condizione di conoscenzadell'illiceità o della manifesta illiceità e non già il ruolo più o meno attivo svolto dall'intermediario. Tentativi in tal senso sono già stati compiuti in dottrina, ma in un quadro ambiguo nel quale non è chiaro se le proposte di tipizzazioni mirino o meno a suffragare e ad alimentare la distinzione tra hosting attivi e passivi. Si è suggerito di considerare attivo: a il provider che non si limita solo ad associare contenuti pubblicitari ai materiali immessi in rete dagli utenti ma che offre agli inserzionisti un servizio che consente di visualizzare i messaggi pubblicitari in relazione agli specifici contenuti dei video immessi dagli utenti, mediante l'utilizzo di parole chiave; b il provider che acquisisce il diritto di utilizzare i video immessi dagli utenti, di modificarli, di distribuirli, di adattarli e che riorganizza i materiali caricati sulla propria piattaforma; c il provider che predisponga un servizio visibile come link sotto ogni video pubblicato in rete che consente al visitatore di segnalare al prestatore del servizio l'eventuale illiceità del contenuto immesso dall'utente e alla redazione di verificare la segnalazione e di provvedere alla rimozione: servizio definito "segnala abuso"; d il provider che fa sottoscrivere ai suoi utenti dei contratti che prevedono sia una licenza non esclusiva per l'esercizio dei diritti di riproduzione e di adattamento relativi ai video caricati sia la possibilità per l'intermediario di rimuoverli; e il provider che fornisca un servizio automatico di "video correlati" che consiste nella visualizzazione a lato o sotto il video in riproduzione di altri contenuti a esso associa-

iuvant?, in *www.ictlawanddataprotecionit.wordpress.com*; D'Arrigo, *Recenti sviluppi in tema di responsabilità degli* Internet Service Providers, cit., p. 80.

[165] In Italia – com'è noto – la figura dell'*hosting provider* attivo è giunta anche agli onori della cronaca nella causa *Google* v. *Vividown* in ordine alla responsabilità penale della prima per illecito trattamento dei dati personali *ex* art. 167 d.lgs. 169/2003, collegata all'inserimento di contenuti illeciti consistenti in un video con finalità denigratorie nei confronti di un disabile e connessa all'omissione all'interno delle condizioni generali del servizio delle informazioni relative agli obblighi imposti dalla disciplina sulla protezione dei dati personali: Trib. Milano, 12 aprile 2010, in *Cass. pen.*, 2010, p. 1288 ss., con nota di R. Lotierzo, *Il caso* Google - Vivi Down *quale emblema del difficile rapporto degli* internet providers *con il Codice della* privacy. Com'è noto, la sentenza è stata poi riformata in appelloApp. Milano, 27 febbraio 2013, in *Vita not.*, 2013, p. 609 ss., con nota di G.M. Riccio, *Google/Vividown:* leading case *o abbaglio giurisprudenziale?* e l'assoluzione è stata confermata in Cassazione: Cass. pen, 3 febbraio 2014, n. 5107, in *Dir. inf.*, 2014, p. 225 ss., con nota di F. Resta, *La rete e le utopie regressivesulla conclusione del caso Google/Vividown*, e in *Vita not.*, 2014, p. 663 ss., con nota di M. Iaselli, *Caso Vividown: la decisione della Cassazione nel solco della legalità*.

bili[166]. Secondo una diversa e più ampia tipizzazione, elaborata sulla sorta delle indicazioni ricavabili dalla giurisprudenza della Corte di Giustizia, il provider esercita un ruolo attivo, se non addirittura compartecipativo, che non giustifica l'applicazione del regime di immunità previsto dagli artt. 12 e ss. dir. 00/31 e 14 e ss. d.lgs. 70/2003: a nel caso in cui l'intermediario offra un spazio virtuale per la visualizzazione di contenuti caricati dall'utente ma controllati dall'intermediario medesimo prima della loro immissione in rete; b nel caso in cui l'intermediario offra uno spazio virtuale con un motore di ricerca il quale consente in risposta all'inserimento di parole chiave la visualizzazione di link di contenuti presenti sul server dell'intermediario medesimo e caricati in precedenza da un altro utente con il controllo dell'intermediario[167]; c nel caso in cui l'intermediario offra uno spazio virtuale con un motore di ricerca il quale consente in risposta all'inserimento di determinate parole chiave la visualizzazione di determinati link di contenuti presenti sul server dell'intermediario medesimo sulla base di un accordo con l'utente che stabilisce i contenuti caricati nello spazio virtuale messo a disposizione, il contenuto dei link mediante cui vi si accede e le parole chiave che ne permettono la visualizzazione; d nel caso in cui l'intermediario offra uno spazio virtuale con un motore di ricerca il quale consente in risposta all'inserimento di determinate parole chiave la visualizzazione di determinati link per accedere a dei contenuti caricati sul server dell'intermediario medesimo senza alcuna forma di controllo da parte di questi sulla base di un accordo con l'utente che stabilisce sia il contenuto dei link sia le parole chiave; e nel caso in cui l'intermediario offra uno spazio virtuale con un motore di ricerca il quale consente in risposta all'inserimento di determinate parole chiave la visualizzazione di determinati link di siti web presenti nella rete e ciò in forza di un accordo che stabilisce sia il contenuto dei link sia le parole chiave[168].

Al di là del pregevole sforzo di isolare situazioni tipo che possano facilitare la valutazione del giudice sul coinvolgimento o meno dell'intermediario nell'illecito dell'utente, non si può non ammonire dal rischio che operazioni del genere contribuiscano a distogliere dalla corretta prospettiva incentrata invece sullo stato di conoscenza dell'illecito da parte del provider, nelle due varianti in precedenza enunciate. L'elaborazione di casi-tipo ha senso soltanto se concepiti come congegni che agevolano la prova presuntiva della conoscenza da parte dell'intermediario[169]. La centralità di una nozione così tanto elastica come quella di conoscenza, se, da un lato, incrementa la discrezionalità giudiziale, dall'altro lato, consente però di piegare in senso protettivo le sempre nuove conquiste della scienza, e in particolare della tecnologia[170]. Destinata a molti-

[166] Bocchini, *La responsabilità civile degli intermediari del commercio elettronico*, cit., p. 560 ss.

[167] In tale ipotesi, il *provider* non risponde dell'attività del motore di ricerca, in quanto egli non esercita alcuna forma di controllo sulle parole chiave e sulle visualizzazioni dei *link* associate, ma potrebbe essere esposto a responsabilità peri contenuti immessi dagli utenti nello spazio messo a disposizione dell'utente.

[168] Montanari, *Prime impressioni sul caso SABAM c. Netlog NV: gli* Internet Service Provider *e la tutela del diritto d'autore* online, cit., pp. 1090-1091.

[169] Sul problema della ripartizione dei temi di prova e sull'onere del soggetto che si dichiara leso o danneggiato dal fatto illecito dell'utente di provare la conoscenza da parte dell'intermediario si sofferma assai opportunamente Bocchini, *La responsabilità di Facebook per la mancata rimozione di contenuti illeciti*, cit., c. 640.

[170] Sui rapporti tra diritto e scienza cfr. Aa.Vv., *Giurisprudenza e scienza*Roma, 9-10 marzo 2016, Atti dei Convegni Licei, Roma, 2017, *passim*.

plicare le possibilità di azione dell'uomo, la tecnologia si presenta il più delle volte – o per lo meno in prima battuta – come un complesso di mezzi, sempre più sofisticati, che ampliano le occasioni di lesione di situazioni giuridiche o di produzione di danni. In campi più condizionati di altri dal progresso scientifico, qual è l'attività degli *internet service provider*, la tecnologia dovrebbe diventare oggetto di un precetto di salvaguardia, una sorta di dispositivo di chiusura del sistema delle regole, volto ad estendere la tutela giuridica sin dove la tecnologia consenta: un dispositivo al quale ben si attaglia l'appellativo di "variabile tecnologica". Se questo è l'obiettivo, la tipizzazione sulla base del contenuto della prestazione dell'intermediario non si rivela la più opportuna dal momento che adotta una prospettiva statica, incentrata su costanti di azione e destinata fatalmente a vanificare le potenzialità di salvaguardia offerte dal progresso tecnologico. La nozione di conoscenza consente al giudice, in sede di valutazione del comportamento tenuto dall'intermediario e, dunque, in chiave retrospettiva, di verificare se sussistano strumenti tecnici che avrebbero permesso all'intermediario di acquisire elementi di conoscenza dell'illecito commesso online di cui non si è tenuto conto nei precedenti giudiziari o amministrativi perché in passato non acquisibili a causa dello sviluppo tecnologico dell'epoca. Dal punto di vista della più stretta tecnicalità giuridica, la "variabile tecnologica" potrebbe rendere più agevole per il soggetto leso o danneggiato la prova per presunzioni della conoscenza dell'intermediario e, per converso, più ardua per quest'ultimo la mera difesa, detta anche eccezione in senso lato, ossia la confutazione dell'idoneità dei fatti addotti dall'attore a fornire la prova critica della conoscenza dell'illiceità dei contenuti da parte del convenuto.

Se non impostati in questi termini i rapporti tra tecnica e diritto rischiano di assumere un assetto non auspicabile, nel quale le valutazioni giuridiche non godono di una sufficiente autonomia ma sono rigidamente condizionate dalle condizioni tecniche. Una sorta di capovolgimento in cui le evidenze del progresso scientifico e tecnologico finiscono per rivelarsi più prescrittive delle disposizioni normative. E di ciò si intravvedono già le tracce: nella regolazione giuridica degli intermediari le condizioni offerte dalla tecnica e la ragionevolezza dei relativi costio l'impatto sull'organizzazione dell'impresa influenzano l'individuazione della regola di diritto e dell'apparato dei rimedi. Il canone europeo prescrive infatti: 1 l'impossibilità di configurare un obbligo di vigilanza attiva su tutti i dati immessi sulla piattaforma telematica dagli utenti; 2 la necessità di individuare rimedi effettivi a tutela dei diritti, specie dei diritti d'autore; 3 ma anche la necessità che essi si rivelino equi, proporzionati, non eccessivamente costosi; 4 e tali da scongiurare il rischio dell'introduzione di barriere al commercio legittimo[171].

[171] Al punto che Ricolfi, *Contraffazione di marchio e responsabilità degli* internet service providers, cit., p. 250 osserva correttamente che l'esclusione di un dovere di monitoraggio generalizzato dei contenuti memorizzati continua a operare a favore dell'intermediario anche dopo la conoscenza di elementi che rendano manifesto il carattere illecito di tali contenuti e, addirittura, anche dopo l'emanazione dell'ordine inibitorio da parte dell'autorità giudiziaria o dell'autorità amministrativa competente.