ASA 2021 Statistics and Information Systems for Policy Evaluation

BOOK OF SHORT PAPERS
of the on-site conference

edited by
Bruno Bertaccini
Luigi Fabbris
Alessandra Petrucci

ASA 2021 Statistics and Information Systems for Policy Evaluation

BOOK OF SHORT PAPERS
of the on-site conference

edited by
Bruno Bertaccini
Luigi Fabbris
Alessandra Petrucci

ASA 2021 Statistics and Information Systems for Policy Evaluation

BOOK OF SHORT PAPERS
of the on-site conference

edited by
Bruno Bertaccini
Luigi Fabbris
Alessandra Petrucci

Cover graphic design: Lettera Meccanica SRLs
Front cover: © man64|123rf.com

ASA 2021 On-site Conference on
STATISTICS AND INFORMATION SYSTEMS
FOR POLICY EVALUATION

University of Florence,   September 6 - 8, 2021

Bruno Bertaccini, Luigi Fabbris and Alessandra Petrucci

(Editors)

Partners

Table of contents

Preface

The Association for Applied Statistics (ASA) and the Department of Statistics, Computer Science, Applications DiSIA "*Giuseppe Parenti*" of the University of Florence, jointly with the partners AICQ-CN (Italian Association for Quality Culture North and Centre of Italy), AISS (Italian Academy for Six Sigma), ASSIRM (Italian Association for Marketing, Social and Opinion Research), Comune di Firenze (the Florence Municipality), SIS (the Italian Statistical Society), Regione Toscana (the Tuscany Region) and Valmon – Evaluation & Monitoring Ltd, have organised a scientific conference titled "*Statistics and Information Systems for Policy Evaluation*", aimed at promoting new statistical methods and applications for the evaluation of policies.

Due to the health emergency caused by the COVID-19 pandemic, the Scientific and the Local Organizing Committees decided to reschedule the conference appointment in two different scientific events: an on-line Opening Conference held in February and March 2021 and a postponed on-site Conference held in September 2021.

This Book includes 40 peer-reviewed short papers discussed during the on-site Scientific Conference. This event was spread over 3 days and organized in thematic sessions; each session, led by a chair, collected works on the following homogeneous issues: "Evaluation Of Educational Systems", "Decision Making", "Health and Well-Being", "Tourism and Gastronomy". The papers published in this book are organized in those sessions.

On behalf of the Scientific Program Committee, we would like to thank the authors for submitting and presenting their interesting and inspiring works in the context of the evaluation of policies, the partners, the chairs, the discussants and the Local Organizing Committee. Finally, we are thankful to the members of the Scientific Committee for helping with the peer-reviewing process.

Florence (Italy), October 2021

Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci

EVALUATION OF EDUCATIONAL SYSTEMS

Determinants of the transition to upper secondary school: differences between immigrants and Italians

Patrizio Frederic, Michele Lalla

1. Introduction

Education decisions that teenagers in 13-15 age range face are the first important steps in the lifecycle which determine their educational achievements and job trajectory. These choices occur at a particular stage in their lives, when influences inside the home are still strongly felt and knowledge about their interests and abilities or skills is vague and unstable. In this sense, such decisions strongly depend on both individual and family characteristics involving their socio-economic conditions, as well as on the environment or contextual background of the area where they reside.

The objective of this paper is to pinpoint differences with respect to citizenship, a binary variable distinguishing between immigrants and non-immigrants (hereinafter also referred to as Italians), and the *secondary* binary variable, defined as equal to one for individuals who were not enrolled in an upper secondary school and equal to zero otherwise. The Bayesian approach has been applied to investigate the determinants of the secondary variable. The prior distribution was set to be a Laplace distribution with parameter λ. Hence, the Bayesian estimation of the model parameters corresponds to the Lasso estimation procedure. The latter is a popular method that simultaneously allows for the selection of the explanatory variables and their interactions and the estimation of the model coefficients. Starting from an initial model, which includes all the selected quantitative and categorical variables and all the interactions between the categorical variables, the applied method led to a very parsimonious model, but surprisingly it did not include family income.

2. Data sources and descriptive statistics

The data were extracted from two surveys, with the reference year being 2009, carried out by the Italian National Institute of Statistics (Istat): one being the European Union Statistics (or Surveys) on Income and Living Conditions (EU-SILC) restricted to Italy alone, IT-SILC (Istat, 2008; Eurostat, 2009), and the other being the Italian Survey on Income and Living Conditions of families with Immigrants (IM-SILC), which is a single cross-sectional survey (Istat, 2009) that involved families with at least one immigrant component residing in Italy. The IT-SILC sample was added to the IM-SILC sample to obtain a sample with a consistent number of immigrants with respect to non-immigrants. For further details about these two data sets and about the main variables introduced in the model, see Lalla and Frederic (2020). The target sample was obtained by first selecting individuals in the age range of 16 to 19, obtaining a sample of 2,702 cases. Then, among the latter, the eligible cases were only those individuals whose highest attained ISCED (International Standard Classification of Education) level was equal to 2 (=lower secondary education). The final target sample was made up of 2,039 individuals.

The relationship between the secondary (binary) dependent variable and the ISCED Level Currently Attended (ILCA) showed that 16.9% of individuals were not enrolled in further education (termed "not-attending"), while 79.7% were currently attending an upper secondary school (Table 1).

Patrizio Frederic, University of Modena and Reggio Emilia, Italy, patrizio.frederic@unimore.it, 0000-0001-9073-2878
Michele Lalla, University of Modena and Reggio Emilia, Italy, michele.lalla@unimore.it, 0000-0002-1639-7300

Table 1. Absolute frequencies and row percentages of secondary (binary) dependent variable by the ISCED level currently attended (ILCA)

Secondary\ ILCA	Not-attending	Vocational school	Upper secondary	Total
Secondary = 1	344	69		413
	83.3	16.7		100.0
Secondary = 0			1626	1626
			100.0	100.0
Total	344	69	1626	2039
	16.9	3.4	79.7	100.0

The ILCA was examined with respect to several qualitative variables and revealed many significant relationships. For the sake of brevity, only some of them are cited. The ILCA showed a significant relationship with respect to citizenship, $CS(2)= 45.177$ ($p<0.000$), where $CS(g)$ stands for Chi-Square with g degrees of freedom: the percentage of immigrants attending upper secondary education was lower than that of Italian citizens (74.3% versus 81.7%), while the percentage of immigrant not in school was higher than that of Italians (24.9% versus 14.0%). There was a significant relationship between the ILCA and self-perceived health, $CS(2)= 8.351$ ($p<0.015$), implying that individuals perceiving fair or bad or very bad health tended to discontinue their education with respect to those perceiving good or very good health (Ichou and Wallace, 2019). The ILCA was also related to the index of the total self-perceived health of parents, $CS(6)= 27.356$ ($p<0.000$). The ILCA proved to be linked to the Italian macro-regions $CS(8)= 39.092$ ($p<0.000$), as industrialisation and the possibility of finding employment increased, the percentage of individuals not in school decreased. The ILCA was related to the maximum ISCED level attained by parents, $CS(12)= 179.908$ ($p<0.000$). As the education of parents increased, the percentage of young individuals in school increased. The ILCA yielded significant relationships also with several variables describing the working conditions of the parents, although the strength of such relationships was often weak.

The ILCA was also analysed with respect to the main quantitative variables.

The age of fathers analysed according to the ILCA and citizenship showed that the fathers of immigrants were younger than the fathers of Italians by about four years. Similarly, the mothers of immigrants were younger than the mothers of Italians by about four years and nine months. The Disposable Family Income (DFI) per capita (in thousands of euros) is reported in Table 2 by the ILCA and citizenship. On the average, the DFI per capita for immigrants was significantly lower than that of Italians by about four thousand euros: about 39.8%.

Table 2. Absolute frequencies, means, and standard deviations (SD) of the disposable family income per capita (in thousands of euros) by citizenship and by the ISCED level currently attended (ILCA) by their children

Citizenship\ ILCA	Not-attending	Vocational school	Upper secondary	Total
Italian citizen, *n*	211	65	1229	1505
Means	*8.103*	*9.202*	*10.835*	*10.381*
SD	5.413	9.914	7.867	7.730
Foreign citizen, *n*	133	4	397	534
Means	*5.393*	*2.249*	*6.573*	*6.247*
SD	3.630	1.824	4.338	4.201
Total, *n*	344	69	1626	2039
Means	*7.055*	*8.799*	*9.794*	*9.298*
SD	4.975	9.764	7.396	7.213

The other types of income considered in the models revealed various structures of relationships and levels of significance. For example, the gap between immigrant and Italian fathers amounted to about ten thousand euros, i.e., –37.4%. The mothers' disposable personal income presented similar statistically significant differences for both marginal effects, with a gap amounting to about five thousand nine hundred euros, i.e., –39.5%. However, the disposable personal income gender gaps were –51% for Italians and –54% for immigrants.

The size of immigrant families proved to be slightly larger than those of Italians and was statistically significant for both marginal effects, i.e., citizenship and the ILCA.

Citizenship was examined with respect to some other variables, even if it was not a target dependent variable. Its relationship with the maximum ISCED level attained by parents was statistically significant, CS(6)= 97.73 (p<0.000) (Bertolini and Lalla, 2012; Bertolini et al., 2015). Citizenship was significantly related to the degree of urbanisation, CS(2)= 24.225 (p<0.000): immigrants tended to settle in densely populated areas more than Italians (38.4% versus 35.5%) or in moderately populated areas (44.6% versus 39.3%). Citizenship also showed a significant relationship with the Italian macro-regions and yielded a significant relationship with the index summarising the total self-perceived health of parents, CS(3)= 29.832 (p<0.000) (Ichou and Wallace, 2019). Citizenship proved to be associated with many variables describing working conditions and revealed a significant relationship with the maximum position of parents on the job, CS(4)= 173.877 (p<0.000).

3. Model by Bayesian Lasso selection of regressors

Let Y be the binary variable coding if the i-th individual is not attending upper secondary education, or he/she is. Let $\mathbf{x}_i$ be a vector of regressors. Let π_i be the probability that Y=1 given $\mathbf{x}_i$. Let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)$ be the parameters vector of the model. The logit model is

$$\pi_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})}. \tag{1}$$

A common method that performs estimation and model selection at the same time is the *Lasso* method (Tibshirani, 1996), which is a procedure involving an additional penalization term, L_1, summed up to the negative log-likelihood of the model that depends on an additional parameter named λ, $\lambda \geq 0$. Many penalized methods can be interpreted as the negative logarithm of a posterior distribution in a purely Bayesian fashion. Let $p(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$ be the usual logit model in the usual Bayesian notation, and let $p(\boldsymbol{\beta} | \lambda) \propto \exp\left(-\lambda \sum_{j=0}^{K} |\beta_j|\right)$ be the Laplace prior distribution on coefficients $\boldsymbol{\beta}$; then the posterior distribution is

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{x},\mathbf{y},\lambda) &\propto p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})\ p(\boldsymbol{\beta}|\lambda) \\ &= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \exp\left(-\lambda \sum_{j=0}^{K} |\beta_j|\right) \end{aligned} \tag{2}$$

The choice of parameter λ plays a crucial role in the estimation procedure. Many different studies have focused on this issue. Besides the classic AIC and BIC criteria, a k-fold Cross Validation (CV) procedure and the One Standard Error Rule (1SE) have been proposed. The applied estimation method consists in two steps:

1. The model was first estimated using the *glmnet* (Friedman et al., 2010) package in R (R Core Team, 2019). Then the *optimal* lambda (λ_{1SE}) and the mode estimations $\left(\hat{\boldsymbol{\beta}}_{\lambda_{1SE}}\right)$ were evaluated.

2. Using the R package *MCMCpack*, N=10,000 samples were drawn from the posterior distribution $p(\boldsymbol{\beta}|\mathbf{x},\mathbf{y},\lambda_{1SE})$ to perform a full Bayesian analysis, where $p(\boldsymbol{\beta}|\lambda_{1SE})$ was chosen to be Laplace distributed.

Note that the model matrix of the starting model consists in 2039 rows by 943 columns, and classical methods can be affected by the *curse of dimensionality*. Instead, the Lasso method is very stable and quick, and shrinks 923 values (out of 943) of $\hat{\boldsymbol{\beta}}_{\lambda_{1SE}}$ to zero; thus only 20 betas have a posterior distribution which is not symmetric to zero.

4. Outcomes of the logistic model

The interpretation of coefficients in a logit model is not easy. The odds ratios (OR) are reported in Table 3, which presents only interaction terms of the first order because the analysis of interactions orders was limited to the first order to simplify interpretation. The interactions are indicated by the symbol ×, which may be read as "by".

A binary variable having an odds ratio greater than 1 implied that the group represented by the binary variable equal to 1 had a higher probability of having $y=1$ than the group identified by the binary variable equal to 0. The binary variables $(\mathbf{x}_b)$ with an odds ratio greater than 1 were observed for interactions only. For example, the odds ratio of the interaction term the "father is limited in activity because of health problems" (x_1) × "father with a permanent contract" (x_2), denoted by x_{12}, was equal to 1.826 meaning that the odds of the event $y=1$, when $x_{12}=1$ (both x_1 and x_2 are equal to 1), are +82.6% greater than the odds of the event $y=1$, when $x_{12}=0$. Let $\mathbf{x}_c = \boldsymbol{\mu}$ be the mean values of the continuous regressors. Note that: (1) the product of two binary variables is a binary variable again, (2) the percentage of increment of the reference probability, $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$, is given by [100*(1–OR)] and is reported below in parentheses, (3) the corresponding value of OR may be found in Table 3. The probability of having $y=1$ (i.e., of discontinuing their education) was equal to $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}} = \mathbf{0.160}$, calculated at the mean values of the continuous regressors $(\mathbf{x}_c = \boldsymbol{\mu})$ and the binary variables equal to 0 $(\mathbf{x}_b = \mathbf{0})$. Therefore, for x_{12} the result was a probability of $\pi_{x_{12}|\bullet} = 1.826\times0.160= 0.292$, nearly double the probability for $x_{12}=0$. Similarly, significant high probabilities of discontinuing one's education or dropping out were observed for other interaction terms: "father is limited in activity because of health problems" × "family living in the macro-region Islands" (+149.0%), "father is limited in activity because of health problems" × "family living in a moderately populated area" (+56.3%), "assets reduction for needs" × "young individual with self-perceived bad health" (+234.0%), "assets reduction for needs" × "mother with self-perceived bad health" (+56.3%), "family living in a densely-populated area" × "parents are unemployed or inactive" (+122.7%), "mother only is employed" × "parents skill level on the job is labourer" (+82.3%). In synthesis, real and self-perceived health conditions heavily affect the probability of discontinuing one's education in the transition from lower secondary to upper secondary school and throughout all the secondary school years, although this happens through the interactions with other factors. Note that the "number of helps requests for aid because the family lives in need", which is formally a continuous variable, interacts with "mother suffering from any chronic (long-standing) illness or condition" and yielded an odds ratio greater than 1 at the mean of the first term.

The binary variables having an odds ratio lower than 1 implied that the represented group had a lower probability of having $y=1$ than the complementary group. In Table 3 there are only two binary variables with an odds ratio lower than 1. For example, the binary variable "both parents employed" (BPE) had an odds ratio equal to 0.736 and hence the corresponding complement to one, expressed as a percentage, was equal to [100*(1–0.736)] = 26.4%. Therefore, the probability

of discontinuing one's education amounted to –26.4% (the negative value indicates the reduction quantity) of the probability of the complementary group, which did not have both parents employed, $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$. In other words, the group with BPE equal to 1 had a probability $\pi_{\text{BPE}|\bullet} = 0.736\times0.160 = 0.118$, implying that the probability of the group with BPE equal to 1 decreased the probability of discontinuing their education by an amount of $100\times(1-0.736)=$ 26.4% with respect to the complementary group, which had a probability given by $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}} = 0.160$. Similarly, a significant low probability of discontinuing their education was observed for the interaction term the "assets reduction for needs" × "mother with permanent employment contract" (–51.5%). The constant of the model was not statistically significant, even if its magnitude was comparable with other parameters.

Table 3. Logistic regression with Lasso method and Bayesian approach: Estimated odds ratio (OR), standard errors (SE), p-values (*p*), and means

Variables	OR	SE	*p*	mean
(Intercept)	1.899	1.441	0.1876	
(Age/10)^2	2.766	0.626	0.0000	2.981
(Age father)/10	0.531	0.120	0.0000	4.906
Father: education level in years	0.935	0.328	0.0044	10.598
Mother: education level in years	0.895	0.203	0.0000	10.557
Number of objects owned in home	0.894	0.292	0.0022	4.323
Both parents employed	0.736	0.304	0.0154	0.361
Interactions of first order				
(No. of help requests) × (Mother: chronic illness)	1.146	0.414	0.0056	0.207
(Father: health problems) × (Father: permanent contract)	1.826	0.656	0.0054	0.073
(Father: health problems) × (Macro-region: Islands)	2.490	1.068	0.0198	0.018
(Father: health problems) × (Moderately populated area)	1.563	0.752	0.0376	0.060
(Assets reduction) × (Young individual: bad health)	3.340	1.247	0.0074	0.011
(Assets reduction) × (Mother: bad or poor health)	1.563	0.768	0.0420	0.063
(Assets reduction) × (Mother: permanent contract)	0.485	0.193	0.0118	0.089
(Urban high density) × (Unemployed & inactive)	2.227	0.664	0.0008	0.036
(Only Mother employed) × (labourer)	1.823	0.531	0.0006	0.081
Pseudo-R square	0.171	*n* =	2039	

The continuous variables. The individual age (range 16-19), expressed in decades, showed a parabolic and positive impact on the interruption of education paths before completion of upper secondary school. The high impact may occur for specific reasons: the survey protocols did not interview individuals under the age 16, the vocational school data were not collected well. The other continuous single variables entering the model showed significant effects on the interruption of education. As the ages of fathers and the parents' education levels increased, the probability of discontinuing education decreased. If the number of objects owned in home (dishwasher, refrigerator, telephone, television, and so on) increased, then the risk of interrupting one's education decreased. As indicated above, the increase in the "number of helps requests because the family lives in need" for individuals having a "mother suffering from any a chronic (long-standing) illness or condition" yielded an increase in the risk of dropping out of school. This empirical evidence highlights the importance of welfare programmes to help families experiencing economic and physical difficulties, with the specific aim of reducing the number of students interrupting their education.

The main fault of the Lasso method in selecting significant explanatory variables concerns the lack of some income variables in the model because various income components have frequently been found to be significant in the literature (Ochsen, 2011; Krause et al., 2015).

In the applications, the interactions should be supported by social, behavioural, psychological or economic theories. Otherwise, they may be obtained automatically just by using an adaptive procedure like the Lasso method and only as empirical findings. In fact, few models with interactions exist in the literature. Probably, the interactions may be easily found among binary or categorical variables, but this case is relatively interesting because they can be replaced with specific typologies. The same holds true for the interactions of a continuous variable with other explanatory binary variables, but the interaction between two continuous variables is very difficult to grasp immediately. In general, it is useful to find a theoretical justification for the existence of the interactions, instead of blindly searching for interaction terms. However, it is highly plausible that almost all phenomena are outcomes of interactions among many variables, but knowledge about and explanations of these results may become very complicated and challenging.

References

Bertolini, P., Lalla, M. (2012). Immigrant Inclusion and Prospects through Schooling in Italy: An Analysis of Emerging Regional Patterns, in *Punishing Immigrants: Policy, Politics and Injustice*, eds. C. E. Kubrin, S. Marjorie, R. M. Jr. Zatz, New York University Press, New York (pp. 178-206).

Bertolini, P., Lalla, M., Pagliacci, F. (2015). School enrollment of first- and second-generation immigrant students in Italy: A geographical analysis, *Papers in Regional Science*, **94**(1), pp. 141-160.

Eurostat (2009). Description of Target Variables: Cross–section and Longitudinal, EU–SILC 065 (2009 operation), Directorate F, Unit F–3, Eurostat, Luxembourg.

Forster, A.G., van de Werfhorst, H.G. (2020). Navigating Institutions: Parents' Knowledge of the Educational System and Students' Success in Education, *European Sociological Review*, **36**(1), pp. 48-64.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, **33**(1), pp. 1-22.

Ichou, M., Wallace, M. (2019). The Healthy Immigrant Effect: The role of educational selectivity in the good health of immigrants, *Demographic Research*, **40**(4), pp. 61-94.

Istat (2008). Ceccarelli C., Di Marco M., and Rinaldelli C. (Eds.). L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc). Metodi e Norme n. 37, Istat, Rome.

Istat (2009). Reddito e condizioni di vita delle famiglie con stranieri [electronic resource]. Istat, Rome. https://www.istat.it/it/archivio/52405. Accessed 3 January 2020.

Krause, A., Rinne, U., Schüller, S. (2015). Kick it like Özil? Decomposing the native-migrant education gap, *International Migration Review*, **49**(3), pp. 757-789.

Lalla, M., Frederic, P. (2020). Tertiary education decisions of immigrants and non-immigrants in Italy: an empirical approach, *DEMB Working Papers Series*, N. 168, pp. 1-39. University of Modena and Reggio Emilia - Department of Economics.

Ochsen, C. (2011). Recommendation, class repeating, and children's ability: German school tracking experiences, *Applied Economics*, **43**(27), pp. 4127-33.

R Core Team (2019). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna (AT). URL http://www.R-project.org/. Accessed 18 April 2020.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, *Series B*, **58**(1), pp. 267-288.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso, *The Annals of Statistics*, **35**(5), pp. 2173-2192.

The top candidate is an intermediate one: An analysis of online posts of Veneto industries

Luigi Fabbris

1. Introduction

In this work[2], we examine the results of an experiment carried out in year 2018 by sending a number of fictitious CVs in response to a sample of job vacancies posted by Veneto industries. The experiment aimed at highlighting which applicant characteristics influence the recruiters' call back rate and speed (Brocco et al., 2021).

Common sense dictates that the best candidate for a job is the person who, among those who showed up, possesses the most qualifying characteristics to fill the vacancy. So, "best" is relative to the vacancy. For this reason, a company's recruiter tends to match the expected with the exhibited characteristics and ignores the candidates whose skills are at all irrelevant to the vacancy even if their human and social characteristics stand out.

We hypothesise that another factor affects the initial stage of the recruitment process: the applicant's expectations in terms of benefits and career as perceived by the recruiters. Our hypothesis is that recruiters match the applicants' expectations, as perceivable from their CVs, with the benefits the company is prepared to offer to the future employee and, for this reason, they discard both the worst applicants and the ones who are too good, thus favouring the intermediate ones.

The rest of this paper is organised as follows: Section 2 briefly describes the available data and the survey methodology, Section 3 presents the main results on reverse discrimination and Section 4 interprets the outcomes, refers to the mainstream literature and offers a conclusion.

2. Data and methods

The experimental survey was carried out by sending a number of fictitious CVs in response to online posts for 120 job vacancies in Veneto industries. The experiment consisted of creating and sending five different CVs to each job opening and waiting for the company to call back. The CVs differed according to a fractional factorial design aimed to control a total of ten applicant characteristics: gender, place of origin, field of study, academic degree level and final mark, English and computer skills, driving their own car, and being a music lover or a youth group volunteer. The rate and speed of call backs was expected to reflect the acceptability—or conversely, the social discrimination—likelihood of certain characteristics or combinations of characteristics. As a whole, 600 CVs were emailed in response to online posts and 59 call backs were obtained.

The job openings were drawn from a specialised job-search website—subito.it. In a preliminary comparison with other websites and newspapers, we verified that all job openings advertised through local newspapers were available through the chosen website. Hence, we decided to only use the internet to collect the ads. As a whole, these belonged to the manufacturing industry (20.5%), the service industry (43.2%), other service sectors (21.8%) and commerce (14.5%). Job vacancies were related to the following five activities: administrative offices, human resource offices, marketing activities, commercial offices and information

---

[1] The author wishes to thank Professor Maria Cristiana Martini for her precious help with the segmentation analysis of the data.

systems. These were jobs for which all hypothesised graduates might be appropriate. By design, each type of activity received the same number of openings: one-fifth of the sample.

The criterion variable Y of our analyses is having obtained a response to a mailed CV. So, the Y variable has two possible values: 1 if the company called back and 0 otherwise. For practical purposes, the telephone call backs were equal to the email ones. Moreover, the CVs were randomly selected from a pool defined by logically crossing the ten experimental factors, so all responses were equally important. A certain level of intra-post correlation is possible due to a partial similarity of CVs mailed in response to the same post.

We applied both a segmentation and a multilevel regression analysis. The segmentation, or regression-tree analysis, consists of a stepwise partitioning of the 600 CVs in subsamples according to a predictor at a time to maximize the between-subsample distance of the criterion variable. The segmentation procedure ends if either no partition is statistically significant or the size of a possible subsample is below a predefined minimum. This technique is particularly appropriate to highlight multiple interactions, that is the interaction between a plurality of predictors. In this work, we adopted the CHAID algorithm of the SPSS package, which allows the partitioning of the sample to any number (≥ 2) of subsamples and of categorical predictors. The CHAID results are presented in the following; other multivariate analysis results have been published elsewhere (Brocco et al., 2021).

3. Results

The synthetic results of call backs (Table 1) show the following highlights:

a) The country of origin of candidates who received a call back was predominantly Italy (16% from Veneto and 13% from Southern Italy), followed by Central Africa (8%), Northern Africa (7%), Eastern Europe (7%) and the Middle East (4%). For estimation efficiency, from now on, the candidates' origin is categorised into Italians and born abroad.

b) Gender was not significant, though women were invited to a job interview more frequently than their male counterparts. This result hides a significantly higher preference for social sciences and humanities degrees instead of scientific and technical ones. With a strong correlation between the feminine gender and the attainment of social sciences and humanities degrees, this makes it clear why both women and graduates in social sciences or humanities obtained more call backs than their male and scientific-technical counterparts. In addition, this is not an unexpected result since some vacancies may have been posted to replace a maternity leave.

c) Another expected result is that, *ceteris paribus*, a Master's graduate is preferred to one with only a Bachelor's. The between-level difference of both the call back rate and speed is large. This may mean that, when a recruiter has knowledge of only the CVs of two graduates, she prefers inviting the candidate to a job interview who, in probabilistic terms, is more productive.

d) What is puzzling in our experiment is that graduates with intermediate skills in English and computer use showed higher preferences and obtained a call back quicker than those who stated they were fluent in English and experts in computer use. Indeed, the call back rate was 9% for fluency in English and 12% for intermediate English skills and 8.5% and 10.5% for higher and intermediate computer skills, respectively. Of course, higher or intermediate skill levels were consistently preferred to low levels. Besides, neither of the two distributions of these two basic skills was significant at 10%.

e) The three social characteristics—possession of a car, playing music and volunteering as youth entertainers—were much less relevant as call back correlates. However, the possession of a car showed an unexpected relationship with the call back probability since the applicants who stated they could use one at work obtained far fewer calls (7.2%) than those who did not mention it in their CVs (11%). In a multilevel regression analysis

(Brocco et al., 2021), this variable was significant at 10%.

Table 1. Per cent call back rate by experimental factor, Veneto region, 2018, n=600.

	Overall rate	7-day rate
Gender: Male	7.7	7.0
Female	10.7	8.0
Origin: Veneto	16.0*	16.0**
South Italy	13.0	11.0
Eastern Europe	7.0	6.0
North Africa	7.0	5.0
Central Africa	8.0	5.0
Middle East	4.0	2.0
Degree level: Bachelor	7.0°	4.7*
Master	11.3	10.3
Discipline: Engineering	6.7**	4.8**
Science	3.3	2.7
Social science	14.0	12.7
Humanities	12.7	10.0
Degree mark: High	9.0	8.1
Medium	9.5	8.6
Low	9.0	8.1
English skills: Fluent	9.0	8.1
Intermediate	12.0	10.7
Basic	6.5	6.0
Computer skills: Expert	8.5	7.1
Intermediate	10.5	9.9
Basic	8.5	7.6
Car owner: Available to drive own car	7.2	6.9
Unspecified	11.0	9.6
Music skills: Plays an instrument	10.7	9.5
Unspecified	7.6	7.1
Juvenile groups: Children's entertainer	10.3	8.8
Unspecified	8.0	7.7

*Significance level: *** 1‰, ** 1%; * 5%; ° 10%.*

A multivariate analysis was realised to better understand the reason recruiters showed higher preferences for intermediate levels of competencies and for graduates without a car. The segmentation analysis of the Y variable produced Figures 1 and 2 for the overall and the seven-day rates, respectively. The tree configurations showed the following:
- o   We take for granted that the main result of the experiment is that graduates born abroad are only occasionally preferred to Italian ones.
- o   Among Italians, the highest call back rates were obtained by male graduates in a social sciences or humanities discipline (31.2% vs. 8.7% for female graduates) and by graduates in a STEM (science and engineering) discipline who were intermediate-

level users of computers (17.4% vs. 1.5% for basic and expert users, respectively).

o Among graduates born abroad, the highest call back rates were obtained by those who had good or medium-range grades at graduation and were intermediate in English (16.7% with a peak of 33.3% among social sciences graduates) and by only 5.6% of those who were either basic or fluent in English and had a non-scientific degree.

o The analysis of the seven-day tree only highlights the preference for graduates born abroad who graduated with a Master's Degree with an intermediate knowledge of English, with a 15.4% call back rate. It can be noticed that this rate is much higher than the average call back rate of Italians and more than twice that of all applicants born abroad.

Figure 1. Regression tree obtained partitioning the sample of CVs. Criterion variable: total proportion of call backs. (Significance: *=5%; **=1%; ***=1‰; minimum group size=18)
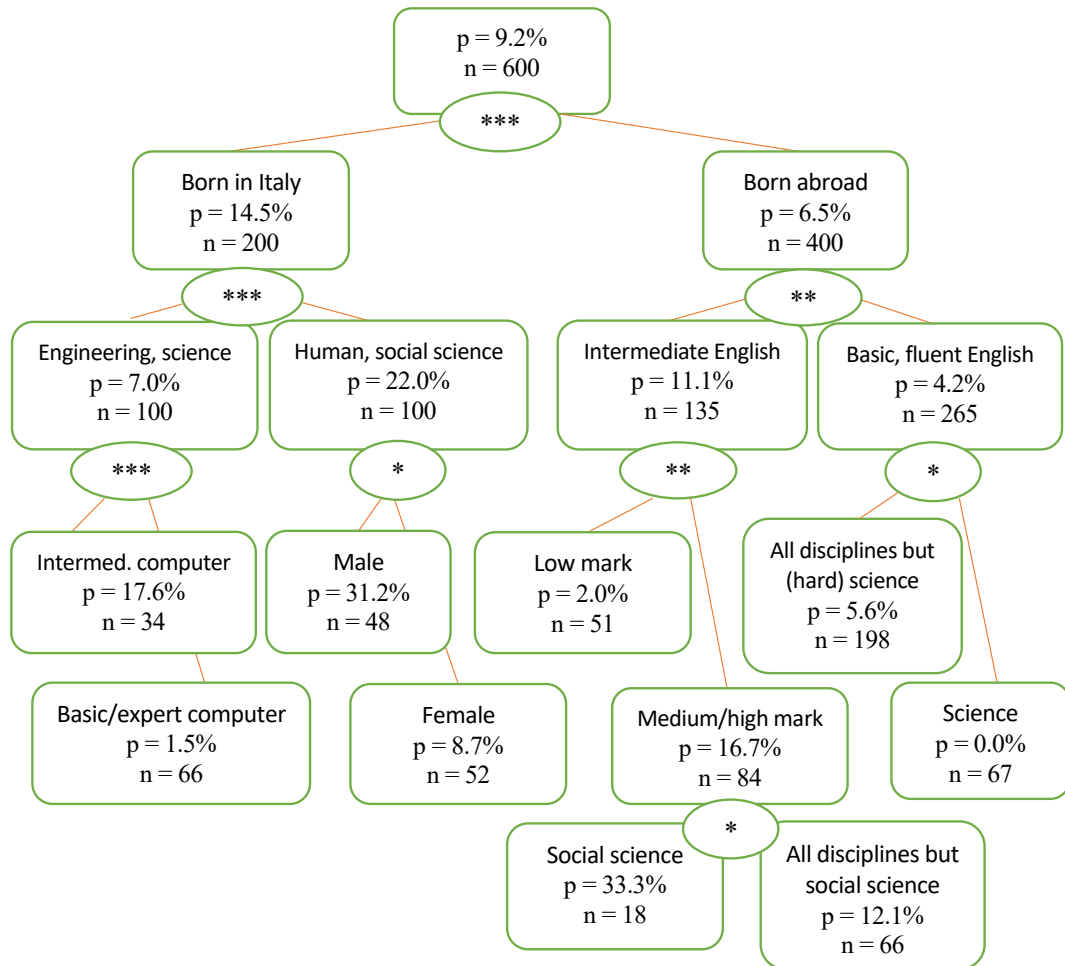
4. Discussion and conclusion

We hypothesised that criteria adopted by recruiters while selecting, through examination of CVs, applicants for an invitation to a job interview are complex. The golden standard of the selection process is the set of activities pertinent to the vacant job. Namely, when confronted with applicants with various competencies, recruiters restrict their choice to those that are pertinent to the vacancy.

Figure 2. Regression tree obtained segmenting the sample of CVs. Criterion variable: call back by seven days from mailing (Significance: *=5%; **=1%; ***=1‰; minimum group size=18)
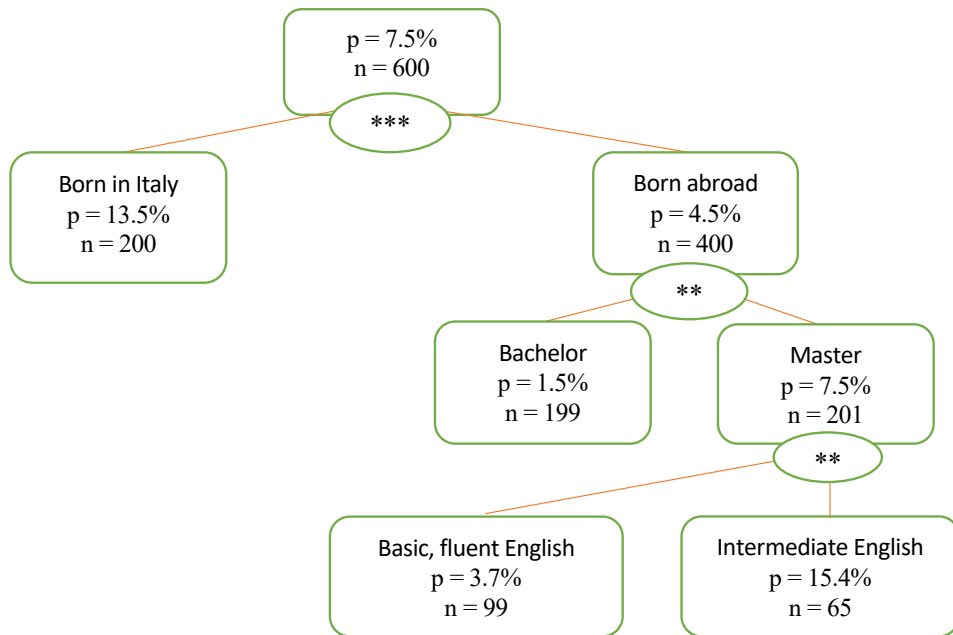
However, competencies are a matter for interpretation because recruiters—with the possible help of line operators—even if they know exactly what the job entails, are called to state if the competence of the best candidate fits the organisation's expectations (Taylor and Bergmann, 1987; Rynes and Barber, 1990; Autor et al., 2003; Thebe and Van der Waldt, 2014). For instance, if they are confronted with two applicants, one with a Bachelor's degree and the other with a Master's in the same discipline, given equivalent financial standing, they tend to prudently choose the latter.

Besides competencies, the perceived attitudes and job-related values of applicants are the basic parameters for recruitment. At this level, the recruiters' tastes may discriminate against certain candidates. A vast body of research indicates that gender, race, age, and physical, moral and cultural characteristics may cause discrimination. Even in this research, the applicants' ethnicity appeared to be a cause of discrimination.

Our research highlighted a sort of reverse discrimination that pushed us to unveil why recruiters implicitly preferred women to men as well as candidates possessing a degree in social sciences or humanities to those with a degree in a STEM field and/or an intermediate rather than a higher level of English and computer skills and, finally, showed aversion against fresh graduates possessing their own car. Indeed, the multivariate analysis showed that the preference for women masked a prevalence of social sciences and humanities degrees among call backs and this means that gender is not a cause of discrimination.

The analysis of multiple interactions involving linguistic and/or computer skills showed a higher preference for applicants perceived as likely to be less demanding. The lower preferences for graduates owning a car can be considered a further symptom of the attitude not to call back wealthier people. We could conclude that recruiters, in opposition to job market common sense (Autor et al., 2003), considered the risk of losing an exceptional but demanding candidate a minor regret. The practical implications of this outcome are that applicants should consider writing their CV accordingly.

In decision theory, this type of attitude refers to the so-called *minimax regret*, or *avoidance of*

regret criterion, which is typical of a risk-neutral decision-maker. An analogous theory in the recruitment field called *uncertainty avoidance* was initially developed by Hofstede (1980) with reference to country cultures and adapted to organizations by Barber (1998) and House et al. (2004). With regard to recruitment, the theory claims that companies should prevent applicants from dropping out during the selection process because the maintenance of candidacies is a factor that improves the organization's reputation.

According to Hofstede, Italy is a country with a high uncertainty avoidance culture. For instance, Italian companies prefer predictability and dislike ambiguous situations. So, in general, recruiters are frightened that changes in applicants' pursuit intentions could cause a loss of image for their organization and could negatively affect their career (Barber, 1998). Indeed, it is easy to imagine that top candidates are given more occupational opportunities than others and are more prone to drop out of candidate pools (Highhouse et al., 2003).

These cultural considerations[3] interact with technology. Online job postings and the company's website are now a main source of organizational information. Therefore, applicants are aware of the reputation of the company advertising the job and recruiters are cognizant that applicants know this. This job market transparency contrasts with the hypothesis that recruiters prefer the good instead of the best candidates. However, graduates apply for job interviews even if they are called back for an interview at another company. Uncertainty avoidance theory is relevant to job applicants since, while trying to avoid the risk of unemployment, they apply even for dead-end and low-paying jobs.

Definitely, we resorted to the hypothesis that even the quality of vacancies could influence the search for a good-instead-of-the-best candidate. Unfortunately, job quality was not a factor in our experimental construct. We suggest that, in future work, the type and quality of job offers be considered as a recruiter's ulterior motive influencing their call back rate and speed.

References

Autor, D.H., Levy, F., Murnane, R.J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, **118(4)**, pp. 1279-1334.

Barber, A.E. (1998). *Recruiting Employees: Individual and Organizational Perspectives*. Thousand Oaks, CA: Sage Publications, Inc.

Brocco, R., Fabbris, L., Martini, M.C. (2021). Il CV da esibire ai selettori per ottenere un colloquio di lavoro in Veneto. In: L. Fabbris (a cura di) *I posti di lavoro, gli imprenditori, i neolaureati* (pp. 159-188), Padova: Cleup.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., Gupta, V. (2004). *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. Thousand Oaks, CA: Sage.

Highhouse, S., Lievens, F., Sinar, E.F. (2003). Measuring attraction to organizations. *Educational and Psychological Measurement*, **63**, pp. 986-1001.

Hofstede, G. (1980). *Culture's Consequences: International Differences in Work-Related Values*. Beverly Hills, CA - London: Sage.

Rynes, S.L., Barber, A.E. (1990). Applicant attraction strategies: An organizational perspective. *Academy of Management Review*, **15**, pp. 286-310.

Taylor, M.S., Bergmann, T.J. (1987). Organizational recruitment activities and applicants' reactions at different stages of the recruitment process. *Personnel Psychology*, **40**, pp. 261-285

Thebe, T.P., Van der Waldt, G. (2014). A recruitment and selection process model: The case of the Department of Justice and Constitutional Development. *Administratio Publica*, **22(3)**, pp. 6-29.

Venaik, S., Brewer, P. (2010). Avoiding uncertainty in Hofstede and GLOBE. *Journal of International Business Studies*, **41**, pp. 1294-1315.

---

[3] Cultural studies in business are developing continuously. See, among others, Venaik and Brewer (2010).

Psychometric properties of a new scale for measuring academic positive psychological capital

Pasquale Anselmi, Daiana Colledani, Luigi Fabbris, Egidio Robusto, Manuela Scioni

1. Introduction

The understanding of the factors that may influence the academic performance of students and the effectiveness of fresh graduates to stand the labor market is a crucial objective to develop adequate educational policies. Individual dispositions and personality traits are among the most important variables that should be considered to achieve this goal. Scholars attributed a relevant role to a set of traits developed within the framework of positive psychology (Seligman & Csikszentmihalyi, 2014), named "psychological capital" (PsyCap; Luthans et al., 2007). PsyCap is defined as an individual's positive psychological state of development, which is characterized by four traits: Self-efficacy, resilience, optimism, and hope. Self-efficacy (or confidence) represents one's awareness of having all the abilities and resources needed to accomplish his own tasks and duties. Resilience indicates the ability to overcome difficulties and "bounce back" from adversities and failure. Optimism reflects the subjective tendency to positively interpret events and circumstances and to consider both positive and negative aspects of reality to drawn new bits of knowledge (Youssef & Luthans, 2005). Finally, hope defines a positive motivational state that is typical of those people who are determined toward their goals and able to redirect, if needed, their strategies to achieve them.

Several instruments for the assessment of these traits can be found in the literature. The most popular is the PsyCap Questionnaire (PCQ; Luthans et al., 2007). Since these instruments are meant for workers, they may not be appropriate for assessing PsyCap traits among fresh graduates who are only about to enter the labor market. To overcome this limitation, a new instrument has been recently developed for measuring PsyCap among students and fresh graduates: The Academic PsyCap (Anselmi et al., 2021; Robusto et al., 2019). It includes four scales that measure the traits of the psychological capital (i.e., self-efficacy, resilience, optimism, and hope) and has been found to be significantly associated with several variables (e.g., entrepreneurial disposition and the number of actions taken to search for a job) that are relevant for students and young workers at the beginning of their careers.

In its last version, the Academic PsyCap includes 24 items, selected from an initial pool of 37, and is characterized by satisfactory psychometric properties (Anselmi et al., 2021). In this work, we present and discuss a refinement of the instrument through a bifactor approach aimed to improve it. The bifactor method allows for modeling the structure of a questionnaire through a general factor and a set of domain-specific factors. In the case of PsyCap, the general factor is the positive psychological capital, whereas the domain-specific factors are the four distinct dimensions it consists of. Using this method to refine the scale would allow for a better understanding of the structure of the positive psychological capital and for developing an instrument that, while assessing the four dimensions of PsyCap, also provides an effective measure of its general factor. This makes sense also in light of the findings of several studies that suggested the existence of a core underlying factor accounting for the overlap between the four PsyCap dimensions (Baron et al., 2016; Choisay et al., 2021; Luthans et al., 2007). The research supported the usefulness of considering the single

Pasquale Anselmi, University of Padua, Italy, pasquale.anselmi@unipd.it, 0000-0003-2982-7178
Daiana Colledani, University of Padua, Italy, daiana.colledani@unipd.it, 0000-0003-2840-9193
Luigi Fabbris, University of Padua, Italy, luigi.fabbris@unipd.it, 0000-0001-8657-8361
Egidio Robusto, University of Padua, Italy, egidio.robusto@unipd.it, 0000-0002-7583-2587
Manuela Scioni, University of Padua, Italy, manuela.scioni@unipd.it, 0000-0003-3192-4030

PsyCap components but also showed that they often act synergistically and that a broader construct may be more effective than the distinct components in predicting individuals' attitudes and performances (Baron et al., 2016; Dawkins et al., 2013; Luthans et al., 2007, 2016).

2. Method

Participants

A sample of 1,603 fresh graduates (Males 38.5%, Mean age = 24.44, SD = 4.36), recruited in the context of the PETERE project, took part in the study. All participants were surveyed within one month after graduation at the University of Padua. The survey was administered via a CAWI (Computer-Assisted Web-based Interviewing) system. Students from medicine and nursing courses were not included in the sample.

Measures

The original pool of 37 items was used to measure the four facets of PsyCap: resilience (11 items), self-efficacy (9 items), optimism (9 items), and hope (8 items). All items were scored on a four-point Likert scale (from 1 "Completely disagree" to 4 "Completely agree").

Analytic approach

A bifactor Exploratory Factor Analysis (EFA) was run on the 37 items. Relying on the results of this model and the investigation of item content, 20 items (five for each dimension) were selected to compose the new Academic PsyCap. Thus, starting from the original full item pool, a new version of the scale was obtained that was based on a bifactor approach. This new scale differed from that developed by Anselmi et al. (2021) with a different (non-bifactor) approach.

The factor structure of the resulting scale was investigated through confirmatory factor analysis (CFA). Three models were tested and compared: a one-factor model, a correlated four-factor model, and a bifactor model. In the first model, all the 20 items of the scale were loaded on a single dimension (PsyCap). In the second model, four different and correlated factors were defined (i.e., self-efficacy, resilience, optimism, and hope), each consisting of five items. Finally, a bifactor model was run that included one general factor (i.e., positive psychological capital) measured by all the 20 items of the scale, and four domain-specific factors (i.e., self-efficacy, resilience, optimism, and hope), each measured by five items.

All models were run using Mplus7 (Muthén & Muthén, 2012), and the WLSMV estimator (weighted least squares mean and variance-adjusted; Muthén & Muthén, 2012), which is recommended for categorical observed data (e.g., Flora & Curran, 2004; Brown, 2006). The goodness-of-fit of the three models was evaluated using several fit indices: χ^2, Comparative Fit Index (CFI), Standardized Root Mean Square Residual (SRMR), and Root Mean Square Error of Approximation (RMSEA). A non-significant χ^2 ($p \geq .05$) suggests adequate fit. Since this statistic is sensitive to sample size, other fit measures were also considered. CFI indices close to .90 (over .95 for excellent fit), SRMR values less than .08, and RMSEA smaller than .06 (.06 to .08 for reasonable fit) are indicative of a good model fit (Marsh et al., 2004). To compare these competing factor structures, the Akaike Information Criterion (AIC; Akaike, 1974) was considered. To this aim, following Olatunji et al. (2019) and Rhemtulla et al. (2012), the 4-point Likert scale data were temporarily treated as continuous and the Robust Maximum Likelihood estimator (Muthén & Muthén, 2012) was used. Concerning AIC, smaller values are indicative of a better fit. Relative differences were considered meaningful if models differed in AIC (ΔAIC) by 10 or more (Burnham et al., 2011). Concerning the bifactor model, a series of indices were also considered, namely the Explained Common Variance (ECV; Sijtsma, 2009; Ten Berge and Sočan, 2004), and McDonald's coefficients (1999) omega (ω) and hierarchical

omega (ωh). The ECV represents the ratio of the common variance explained by the general factor to the total common variance (Reise, Bonifay et al., 2013; Reise, Scheines et al., 2013; Rodriguez et al., 2016). High values (.70 to .80) indicate that the factor loadings obtained from a unidimensional model well approximate those on the general factor obtained from the bifactor solution, and suggest that the scale is substantially one-dimensional (Rodriguez et al., 2016). McDonald's (1999) ω and ωh are factor-analytic "model-based" estimates of internal consistency. The former represents the proportion of variance of the scores that can be attributed to all sources of variance (i.e., general and domain-specific factors), whereas the latter quantifies the amount of variance that is accounted for by the general factor (Revelle & Zinbarg, 2009; Zinbarg et al., 2005, 2007). Both ω and ωh were computed for the general factor. Conversely, ω was also computed for the domain-specific factors. For this coefficient, values close to or greater than .70 are satisfactory. Concerning ωh, values larger than .75-.80 indicate that a factor can be interpreted as the measure of a single construct despite multidimensionality (Reise, Bonifay et al., 2013; Reise, Scheines et al., 2013).

The invariance of the scale across males and females and across bachelor and master graduates was tested through Multiple-Group Confirmatory Factor Analysis (MG-CFA). In the first step, the model was simultaneously fitted to the specific subsamples (males and females; bachelor and master graduates) to test configural invariance (i.e., the same pattern of fixed and free factor loadings were specified across groups). Subsequently, a series of constrained models were tested and compared to evaluate scalar (i.e., invariance of both factor loadings and item thresholds) and strict invariance (i.e., invariance of factor loadings, item thresholds, and residual variances). The test of change in CFI (ΔCFI) was used to compare nested models. Invariance was indicated by ΔCFI values lower than or equal to |.01| (Cheung & Rensvold, 2002).

3. Results

Table 1 shows the factor loadings of the three models that were run on the 20 items selected by applying the bifactor EFA, whereas Table 2 shows the fit indices of these models. The one-factor model did not fit the data, while the other two models obtained a better fit. In the four-factor model, consistently with theoretical expectations, all items showed meaningful loadings on the intended dimensions (λs from .505 to .887, ps ≤ .001), even though correlations between factors were large (rs = from .580 to .985, ps ≤ .001). With regard to the bifactor model, all items significantly loaded on the general factor (λs = from .328 to .799, ps ≤ .001) and on the relative domain-specific factors (λs from .095 to .705, ps ≤ .05). The inspection of ΔAICs indicated that the bifactor model was superior compared with the other two models (ΔAIC between the one-factor and correlated four-factor models = 1892.64; ΔAIC between the one-factor and bifactor models = 2462.13, and ΔAIC between the correlated four-factor and bifactor models = 569.49). Moreover, given the high correlations between the latent factors in the correlated four-factor model, the bifactor solution seems to be the most suitable option to represent the structure of the scale.

In the bifactor model, the ECV of the general factor was .67, indicating that the scale should be intended as multidimensional. However, the value of the ωh coefficient was high (.86), and this suggests that, despite multidimensionality, the general factor could be interpreted as the measure of a single common construct (Reise, Bonifay et al., 2013; Reise, Scheines et al., 2013).

With regard to internal consistency, ω coefficients were satisfactory for both the general and domain-specific factors (ωs = .95, .88, .90, .83, and .81 for general, self-efficacy, optimism, resilience, and hope factors, respectively).

The invariance of the bifactor model was tested across males and females and across bachelor and master graduates. The results are reported in Table 3. All models reached a successful fit in all samples and the value of the ΔCFI supported the considered levels of invariance.

4. Discussion and conclusion

In this work, a 20-item version of the Academic PsyCap was developed adopting a bifactor approach. The resulting scale was found to adequately assess the four dimensions of self-efficacy, resilience, optimism, and hope, as well as to appropriately define a general factor of psychological capital. In the bifactor model, both the domain-specific and the general factors showed adequate internal consistency and factorial validity.

The results of this work are in line with the literature that indicates that PsyCap components often act synergistically as a broader construct that may be more effective than the distinct components in predicting individual's attitudes and performances (Baron et al., 2016; Luthans et al., 2007; see Dawkins et al., 2013; Luthans et al., 2016).

Future studies are advocated to explore the relationships of the Academic PsyCap scales with indicators of students' and fresh graduates' achievements.

Table 1. Factor loadings and correlations between factors

		Bifactor Model		Four-correlated-factor model	One-factor model
	Items	General factor	Domain-specific factors		
Self-efficacy	Usually, when I face a problem, I am able to identify different solutions.	.588	.437	.696	.632
	I have the resources to handle even unforeseen situations.	.625	.309	.705	.636
	If I were in a difficult situation I would be able to find a way out.	.663	.322	.746	.68
	In difficult situations, I feel effective in finding a way out.	.799	.299	.887	.801
	I believe I am able to analyze a problem and identify a possible solution.	.664	.507	.782	.711
Optimism	I'm usually optimistic about the future.	.622	.500	.784	.699
	I always try to believe that behind every cloud there is a blue sky.	.644	.531	.816	.734
	I am convinced that my willpower will prevail over bad luck.	.655	.147	.721	.634
	I always try to see the glass half full.	.642	.632	.839	.761
	Even in difficult situations, I try to take the best opportunities and the bright side.	.724	.391	.853	.757
Resilience	Until now, my successes have largely depended on the choices I made.	.472	.520	.587	.548
	I'm proud of everything I have achieved by now.	.636	.326	.698	.658
	My efforts and my skills are the basis of the results I have achieved.	.507	.537	.625	.584
	Usually, in one way or another, I try to overcome difficulties.	.689	.095*	.701	.663
	I always try to give my best in all the things I do without getting discouraged in the face of obstacles.	.698	.150	.722	.678
Hope	The goals I have achieved so far are due to my planning skills.	.363	.705	.587	.502
	I think I will be able to achieve my current goals by counting on my determination.	.647	.326	.767	.673
	I have a hard time planning things to do when I have to reach a goal. (R)	.328	.536	.505	.430
	Willpower was key to obtaining an academic degree.	.512	.445	.662	.581
	At present, I think I'm a successful person in carrying out my duties.	.619	.189	.699	.619
	Correlations between latent factors				
	Self-efficacy - Optimism	-.318		.639	
	Self-efficacy - Resilience	.110†		.832	
	Optimism - Resilience	-.240		.686	
	Hope - Self-efficacy	.216**		.737	
	Hope - Optimism	-.228		.580	
	Hope - Resilience	.829		.985	

Note. All parameters were significant at $p \leq .001$, excluding those indicated with $*p \leq .05$ and $**p \leq .01$. The parameter indicated with † was non-significant ($p > .05$).

Table 2. Model fit indices

	χ^2	*df*	*p*	RMSEA	C.I. RMSEA	CFI	SRMR	AIC
One-factor model	4960.981	170	.000	.133	.129, .136	.806	.100	58175.32
Correlated four-factor model	1916.709	164	.000	.082	.078, .085	.929	.062	56282.68
Bifactor model	731.961	144	.000	.050	.047, .054	.976	.034	55713.19

Table 3. Fit indices of multiple-group confirmatory factor analyses for invariance

	Gender invariance						Bachelor/Master invariance					
	χ^2	*df*	*p*	RMSEA	CFI	ΔCFI	χ^2	*df*	*p*	RMSEA	CFI	ΔCFI
Configural	896.48	288	.000	.051	.976		843.953	288	.000	.049	.978	
Scalar	880.08	358	.000	.043	.980	-.004	805.032	358	.000	.039	.982	-.004
Strict	824.74	378	.000	.038	.983	-.003	838.677	378	.000	.039	.982	.000

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), pp. 716-723.

Anselmi, P., Colledani, D., Fabbris, L, Robusto, E., Scioni, M., (2021). Measuring content validity of academic psychological capital and locus of control in fresh graduates, in *ASA 2021 Statistics and Information Systems for Policy Evaluation, Vol. 127*, eds. B. Bertaccini, L. Fabbris, and A. Petrucci, Firenze University Press, Firenze, (IT), pp. 23-28.

Baron, R.A., Franklin, R.J., Hmieleski, K.M. (2016). Why entrepreneurs often experience low, not high, levels of stress: The joint effects of selection and psychological capital. *Journal of Management*, **42**(3), pp. 742-768.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. Guilford Press, New York, (NY).

Burnham, K.P., Anderson, D.R., Huyvaert, K.P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, **65**(1), pp. 23-35.

Cheung, G.W., Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, **9**(2), pp. 233-255.

Choisay, F., Fouquereau, E., Coillot, H., Chevalier, S. (2021). Validation of the French Psychological Capital Questionnaire (F-PCQ-24) and its measurement invariance using bifactor exploratory structural equation modeling framework. *Military Psychology*, **33**(1), pp. 50-65.

Dawkins, S., Martin, A., Scott, J., Sanderson, K. (2013). Building on the positives: A psychometric review and critical analysis of the construct of psychological capital. *Journal of Occupational and Organizational Psychology*, **86**(3), pp. 348-370.

Flora, D.B., Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, **9**, pp. 466-491.

Luthans, F., Avolio, B.J., Avey, J.B., Norman, S.M. (2007). Positive psychological capital: Measurement and relationship with performance and satisfaction. *Personnel Psychology*, **60**(3), pp. 541-572.

Luthans, K.W., Luthans, B.C., Palmer, N.F. (2016). A positive approach to management education: The relationship between academic PsyCap and student engagement. *Journal of Management Development*, **35**(9), pp. 1-22.

Marsh, H.W., Hau, K.T., Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling*, **11**(3), pp. 320-341.

McDonald, R.P. (1999). *Test Theory: A Unified Approach*. Erlbaum, Hillsdale, (NJ).

Muthén, B.O., Muthén, L.K. (2012). *Mplus Version 7: User's Guide*. Muthén & Muthén, Los Angeles, (CA).

Olatunji, B.O., Ebesutani, C., Tolin, D.F. (2019). A bifactor model of obsessive beliefs: Specificity in the prediction of obsessive-compulsive disorder symptoms. *Psychological Assessment*, **31**(2), pp. 210-225.

Reise, S.P., Bonifay, W.E., Haviland, M.G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, **95**(2), pp. 129-140.

Reise, S.P., Scheines, R., Widaman, K.F., Haviland, M.G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement,* **73**, pp. 5-26.

Revelle, W., Zinbarg, R.E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, **74**(1), pp. 145-154.

Rhemtulla, M., Brosseau-Liard, P.É., Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, **17**(3), pp. 354-373.

Robusto, E., Maeran, R., Colledani, D., Anselmi, P., Scioni, M. (2019). Development of two scales for measuring academic psychological capital and locus of control in fresh graduates. *Italian Journal of Applied Statistics*, **31**(1), pp. 13-28.

Rodriguez, A., Reise, S.P., Haviland, M.G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, **21**(2), pp. 137-150.

Seligman, M.E., Csikszentmihalyi, M. (2014). Positive psychology: An introduction, in *Flow and the Foundations of Positive Psychology. The Collected Works of Mihaly Csikszentmihalyi,* ed. M. Csikszentmihalyi, Springer, Dordrecht, (NL), pp. 279-298.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, **74**(1), pp. 107-120.

Ten Berge, J.M., Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, **69**(4), pp. 613-625.

Youssef, C.M., Luthans, F. (2005). Resiliency development of organizations, leaders and employees: Multi-level theory building for sustained performance, in *Authentic Leadership Theory and Practice: Origins, Effects and Development (Monographs in leadership and management, Vol. 3),* eds. W. Gardner, B. Avolio, and F. Walumbwa, Elsevier, Oxford, (UK), pp. 303-343.

Zinbarg, R.E., Revelle, W., Yovel, I. (2007). Estimating ω h for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement*, **31**(2), pp. 135-157.

Zinbarg, R.E., Revelle, W., Yovel, I., Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, **70**(1), pp. 123-133.

Gender and Information and Communication Technologies interest: results from PISA 2018

Mariangela Zenga

1. Introduction

The information and communications technology (ICT) is a growing presence in the modern society, so knowledge and skills related to ICT should become an integral part of education. Moreover the development of digital literacy also takes place firstly at school but also in the informal learning at home, among peers and in other out-of-school contexts (Fraillon et al. (2014); Juhanak et al. (2019); Erstad (2012)). The literature on gender and ICT is a thriving topic in the last years. Previous research pointed out that the differences in gender are much more pronounced for ICT usage at home, instead of at school (BECTA (2008)): boys use more often ICT outside of school for leisure purposes (as for playing computer or console games) than girls. On the contrary, girls have a greater use of ICT for school work and online social networking. Considering attitudes, confidence and self-efficacy girls show lower level on ICT in comparison to boys.

In 2015, a new construct, *ICT Engagement*, was introduced by Zylka et al. (2015). ICT Engagement is theoretically based on self-determination theory (Deci and Ryan (2000)) and it is assumed to be "a crucial individual factor for developing and adapting ICT skills in a self-regulated way" that "facilitates learning and acquiring new knowledge and skills through the life span by using ICT in both formal and informal learning environments" (Goldhammer et al. (2017)). The ICT Engagement involves ICT interest, Perceived ICT competence, Perceived autonomy related to ICT use, and ICT as a topic in social interaction (Goldhammer et al. (2017)). In this work we are interested in the ICT interest (ICTI) that represents a "content-specific motivational disposition" and describes "individuals' long-term preference for dealing with topics, tasks, or activities related to ICT" (Goldhammer et al. (2017)). Six items are included in the construct using a four-point Likert response scale ranging from 1 to 4 (where: 1=Strongly disagree, 2= Disagree, 3=Agree, 4=Strongly Agree):

- I forget about time when I'm using digital devices;
- The Internet is a great resource for obtaining information I am interested in (e.g. news, sports, dictionary);
- It is very useful to have social networks on the internet;
- I am really excited discovering new digital devices or applications;
- I really feel bad if no internet connection is possible;
- I like using digital devices.

The overall index of ICTI based on the previous six items is scaled using a generalized partial credit model (Muraki (1992)) and values of the index correspond to Warm likelihood estimates (Warm (1989)) that are standardized in a second moment. In this way, the index shows the average equal to zero and the standard deviation equal to one across OECD countries (PISA (2018)).

Using 2018 PISA data, the relationship between gender and ICTI for 15-year-olds in OECD countries will be analyzed. Moreover a three-level multilevel model will show the effects on

the ICTI of the characteristics of the respondent, of the school in which the student is and of the country in which the student's living.

2. Data

The OECD Programme for International Student Assessment (PISA) is a triennal international survey and it aims to assess the performance in mathematics, reading, science and financial literacies. It provides the most comprehensive and rigorous international assessment of 15-aged students learning outcomes to date. In several countries the questionnaire has questions on students' familiarity with ICTs and engagement in ICT. The assessments are also supplemented by background questionnaires. Pupils are asked about their motivations for study, attitudes to school, views on reading, and their socio-economic background. Another questionnaire asked headteachers about the challenges facing their schools, organisation and factors that they believe affect their students' performance.

In this paper we analyze 109,106 15-aged students (49.1% of female) of 8,115 schools who were sampled for PISA 2018 in the 23 OECD countries. Approximately 38% of students started to use a digital device when they were 6 years old or younger, 37% when they were 7-9 years old and 25% when they were 10 years old or older.

3. Statistical method

The multilevel models are used in literature when the aim of analysis is to investigate the relationships between outcomes and variables when data presents naturally a hierarchical structure (Goldstein (2011); Hox et al. (2017) and Rice and Leyland (1996)). In this work, a model with three levels is proposed: students within school belonging to a OECD country. In particular, the multilevel models will control for the presence of a possible effect of school, which may render students within the same school more alike in terms of experienced outcome than students coming from different schools, everything else held equal. Moreover, it is possible to consider also the influence of country. As aforementioned, the proposed model includes three levels: students i as level-1 unit ($i = 1, ..., n_j$), school j as level-2 unit ($j = 1, ..., J$) and country k as level-3 unit ($k = 1, ..., K$). The aim of the analysis is the identification of some relationships between ICTI and some characteristics related to the students, schools and countries. Let Y_{ijk} be the score of the ICTI, for student i within school j belonging to country k. Following Hox et al. (2017), let $X^{(1)} = \{X_{ijk}\}$ be the matrix for the explanatory variables at the level-1, $X^{(2)} = \{X_{jk}\}$ the matrix for the explanatory variables at the level-2 and $X^{(3)} = \{X_k\}$ the matrix for the explanatory variables at the level-3. The level-1 model states a linear relationship between the observed response and the level-1 covariates:

$$Y_{ijk} = \alpha_{0jk} + X^{(1)}\alpha_{1jk} + \epsilon_{ijk}. \tag{1}$$

At the level-2, the intercept of the level-1 model (eq. 1) can be written as:

$$\alpha_{0jk} = \beta_{00k} + X^{(2)}\beta_{1k} + u_{0jk}. \tag{2}$$

Finally, the level-2 intercept in equation 2 can be modeled as:

$$\beta_{00k} = \lambda_{00} + X^{(3)}\lambda_1 + \gamma_{0k}. \tag{3}$$

Combining equation 3 and equation 2 in equation 1, it yields the following:

$$Y_{ijk} = \lambda_{00} + X^{(1)}\alpha_{1jk} + X^{(2)}\beta_{1k} + X^{(3)}\lambda_1 + u_{0jk} + \gamma_{0k} + \epsilon_{ijk}. \tag{4}$$

In eq. 4, the fixed effects are given by the the overall intercept (λ_{00}), the student level, school level and country level covariates, the random effects are given by the school level and the country level ($u_{0jk} + \gamma_{0k}$), and the residuals are represented by ϵ_{ijk}. Respect to the meaning of the levels, u_{0jk} is the unobserved school random effect of the intercept amongst schools, with $u_{0jk} \sim N(0, \sigma_u^2)$; γ_{0k} is the random variation of the intercept amongst countries, with $\gamma_{0k} \sim N(0, \sigma_\gamma^2)$; moreover $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$. Random components at different levels are assumed uncorrelated, whilst non null correlations are assumed for students in the same school or in the same country. The random effect among schools can be interpreted as the mean score in ICTI of schools with respect to outcome adjusted for fixed coefficients related to student, school and country characteristics. The u_{0jk} estimates show the contribution of the j-th school to mean score in ITCI. Using the model in eq. 4, the intraclass correlation (ICC) is defined as:

$$\begin{cases} ICC_{Level-2} = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_\gamma^2 + \sigma_\epsilon^2} \\ ICC_{Level-3} = \dfrac{\sigma_\gamma^2}{\sigma_u^2 + \sigma_\gamma^2 + \sigma_\epsilon^2} \\ \end{cases} \tag{5}$$

In general, the ICC indicates the proportion of the variance explained by the grouping structure in the population. In particular, ICC in eq. 5 identifies the proportion of variance at the school level ($ICC_{Level-2}$) and at the country level ($ICC_{Level-3}$).

4. Results

First of all, we analyze the difference in gender for the ICTI. In 2018, the ICTI mean for the female is equal to -0.10, while it is equal to 0.10 for male group and the *F* test for Gender $F = 11.47, p = 0.0007$ suggests that there are differences in ICTI between male and female. Considering the interaction of gender and country on the ICTI, the *F* test ($F = 90.29, p < 0.0001$) suggests significant effects. The Fig.1 reports the countries difference in means for gender for the ICTI as

$$\Delta(ICTI) = ICTI_{Female} - ICTI_{Male} \tag{6}$$

If $ICTI < 0$ then Female group shows lower value in mean for ICTI respect to the male group, while if $ICTI > 0$ then female group shows greater value in mean for ICTI respect to the male group. If $ICTI = 0$ female group has the same level of interest in ICT than the male group. Czech Republic students show the lowest difference in means for ICTI (-0.27), followed by Luxembourg students (-0.23) and Belgian students (-0.25), while Greek students have the highest difference in means for ICTI (0.23), followed by Korean students (0.17) and Irish students (0.15).

As the second aim, we consider a three-level multilevel model considering the effect on several explanatory variables on ICTI. The explanatory variables are: GENDER, AGE (Age when the respondent first used a digital device), ESCS (index of family economic, social and cultural status), WEALTH (family wealth possession), ICTHOME (availability of ICT devices in student's home), ICTSH (availability of ICT devices in student's school) and USESH (ICT use at school). Table 1 reports the results for the models. The test on the random effects shows that a three levels model is required underlining that the ICTI depends both on the shool effect and on the country effect. As shown in Model 0, the ICC values for the three levels model indicate that approximately 19% of the variability in the ICTI is accounted by the school and 3% by the country.
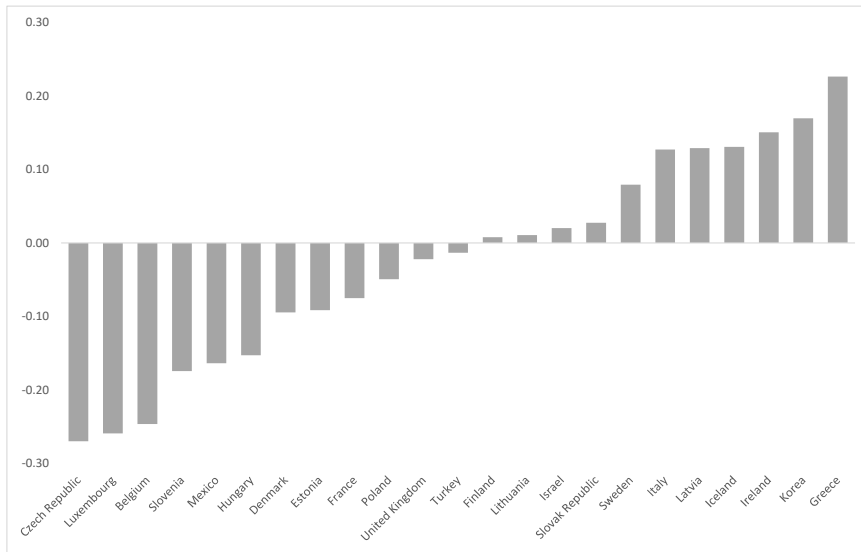
Figure 1: The difference in Gender in ICTI for OECD countries.

The results of the complete model (Model 1) show that the ICTI level of female respondents is higher than the level of male respondents. The age at which children start using a digital device has a significant relationship with the level of the ICTI. In particular, the sooner children approach a digital device, the higher their interest in ICT will be. The availability of ICT devices in student's home a significant positive relationship with ICTI, on the contary the availability of ICT devices in student's school have a negative impact. The family wealth possession results to have a positive impact on ICTI. Moreover as the use of ICT at school increases, the level of ICTI increases too.

5. Conclusion

The ICT interest is an individual preferring (long-term) participation in activities related to ICT and its use (Goldhammer et al. (2017)). For this reason it is important to investigate the relationship among the level of ICT interest of 15-years old students and variables in a family environment considering the influences of the school by a multilevel model. We verified that the gender difference exists, but it depends also by the country in which the students live. The results also seem to underline that the ICT interest depends in large scale on the school effect more than a country effect. Other factors were considered: gender, age at which children start using a digital device, availability of ICT devices in student's home, family wealth possession and use of ICT at school increases, the level of ICTI represent interesting explanatory variables for the model. No evidences seem to be related to the economic, social and cultural status of the student's family.

Table 1: Results of the three-level multilevel models. Source: calculations on 2018 PISA data. In table "rc" means "reference category"

Variable	Model 0			Model 1		
	Est.	**SE**	*p-value*	**Est.**	**SE**	*p-value*
Fixed Effects						
Constant	-0.041	0.040	0.298	0.0106	0.042	0.800
Gender (rc: Male)						
Female				0.014	0.006	0.023
Age (rc : 7-9 y. old)						
3 years old or younger				0.274	0.012	< .0001
4-6 years old				0.140	0.007	< .0001
10-12 years old				-0.079	0.008	< .0001
13 years old or older				-0.275	0.014	< .0001
ESCS				0.004	0.004	0.382
ICTHOME				0.006	0.005	0.001
ICTSCH				-0.018	0.001	< .0001
WEALTH				0.065	0.005	< .0001
USESCH				0.175	0.003	< .0001
Random Effects						
$\sigma^2_{Country}$	0.034	0.011	0.001	0.033	0.010	0.001
σ^2_{School}	0.233	0.006	< .0001	0.222	0.005	< .0001
$ICC_{Country}$	0.029			0.034		
ICC_{School}	0.192			0.193		
Log likelihood	-173,800.70			-171,061.30		
AIC	347,607.4			342,122.6		
N	109,106			109,106		
Country	23			23		

References

BECTA (2008). How do boys and girls differ in their use of ICT? Coventry: BECTA.

Christoph, G., Goldhammer, F., Zylka, J., and Hartig, J. (2015). Adolescents' computer performance: The role of self-concept and motivational aspects. *Computers & Education*,81, pp. 1–12.

Deci, E. L., and Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, **11**(4), pp. 227–268.

Erstad, O. (2012). The learning lives of digital youth-beyond the formal and informal. *Oxford Review of Education*, **38**(1), pp. 25–43.

Fraillon, J., Ainley, J., Schulz, W., and Friedman, T. (2014). Students' Use of and Engagement with ICT at Home and School, in *Preparing for Life in a Digital Age: The IEA International Computer and Information Literacy Study International Report*, eds. J. Fraillon, J. Ainley, W. Schulz, T. Friedman, and E. Gebhardt Springer International Publishing, pp. 125–166.

Goldhammer, F., Gniewosz, G., and Zylka, J. (2017). ICT Engagement in learning environments, in *Assessing contexts of learning world-wide–extended context assessment framework and documentation of questionnaire material*, eds. S. Kuger, E. Klieme, N. Jude, and D. Kaplan Heidelberg: Springer International Publishing, pp. 331–351

Goldstein, H. (2011). Multilevel statistical models. John Wiley & Sons.

Hox, J. J., Moerbeek, M. and Van de Schoot, R. (2017). Multilevel analysis: Techniques and applications. Routledge.

Juhanak L., Zounek J., Zaleska K., Barta O., and Vlckova K. (2019) The relationship between the age at first computer use and students' perceived competence and autonomy in ICT usage: A mediation analysis. *Computers & Education*, **141**, 103614.

Muraki, E. (1992) A generalized partial credit model: application of an algorithm. *ETS Research Report Series*, **i-30**.

PISA 2018 Technical Report. PISA. OECD publishing, Paris.

Rice, N. and Leyland, A. (1996). Multilevel models: applications to health data. *Journal of health services research & policy*, **1**,(3), pp. 154–164.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, **54**, pp. 427–450.

Zylka, J., Christoph, G., Kroehne, U., Hartig, J., and Goldhammer, F. (2015). Moving beyond cognitive elements of ICT literacy: First evidence on the structure of ICT engagement. *Computers in Human Behavior*, **53**, pp. 149–160

A structural equation model to measure logical competences

Silvia Bacci, Bruno Bertaccini, Riccardo Bruni, Federico Crescenzi,
Beatrice Donati

1. Introduction

Logical abilities are a ubiquitous ingredient in all those contexts that take into account soft skills, argumentative skills, or critical thinking. However, there is a substantial lack of research that addresses the actual possession of such logical abilities by students. With this aim, since October 2020 the University of Florence has promoted a three-stage initiative to collect data in order to measure the logical abilities of students when enrolling at the University. The first stage is an entrance test for assessing the students' initial abilities. This test comprises ten questions, each investigating a specific reasoning construct.

At the second stage, students attend a short training course to strengthen their logical abilities. As third step, in order to evaluate the effectiveness of the course, they take an exit examination, replicating the structure and the difficulty of the entrance test.

This paper builds on the previous work by Bertaccini et al. (2021) where the effectiveness of the course was tested via Item Response Theory (DeMars, 2010; Bartolucci et al., 2019) and test-equating techniques (Battauz, 2015). Building on an enlarged database of students that took the training course and examinations in the second half of 2021 and leveraging auxiliary information about students' characteristics, we estimated a Structural Equation Model (SEM; Duncan, 2014; Bollen, 1989) to have a better comprehension and interpretation of the results reported by Bertaccini et al. (2021).

2. Data and methods

Data

The data that we analyse in this work are obtained from the 80 students that took both the tests and the short training course in the second semester of the academic year 2021-2021. The items of each test aimed at investigating the same logical constructs, namely: Double negation (item code N); Disjunction negation (item code D); Conjunction negation (item code C); Hypothetical reasoning (item code IMPL); Sufficient and necessary conditions (item code NEC); Negation of the universal quantifier (item code NU); Negation of the existential quantifier (item code NE); Modus tollens (item code MT); Syllogism (item code S); Multiple steps deduction (item code DED). The students who respond correctly to a given item are given a score equal to 1, otherwise they are given a score equal to 0.

In addition, we were able to obtain exogenous information on students' characteristics such as their age, the grade obtained at the secondary school, the scientific area the student has enrolled in (i.e. science, social, technic, humanistic) and the years of university enrolment.

Compared to the previous work by Bertaccini et al. (2021), the novelty of this study consists in investigating the role of auxiliary information on students to explain their logic abilities and the effectiveness of the training course. The authors assessed the effectiveness of the course by

first estimating two Item Response Theory (IRT) models, one for the entrance test and one for the exit test, and then it tested whether there was a significant shift in the distribution of the logical abilities through a test equating procedure. Given the availability of auxiliary information, we opted for a one-step procedure based on SEM so as to take into account both the measurement issue of logical abilities before and after the training course and the structural relations among the observed (i.e., student characteristics) and the latent variables (i.e., logical abilities).

Methods

A SEM is a multivariate technique used to test complex relationships between observed (manifest) and unobserved (latent) variables as well as relationships between two or more latent variables. Special observed variables, named indicators or items, are used to measure the latent variables. In turn, observed and latent variables distinguish in exogenous variables, which are not explained within the model, and endogenous variables that are affected by other variables in the model (plus an error term). A SEM is characterised by a system of multiple equations, discerning between two sub-models: (i) a structural model, designed to explain the relationships among latent variables as well as among endogenous latent variables and observed variables, and (ii) a measurement model, to link the latent variables to the items. In more detail, the structural model can be expressed by the following equation

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{1}$$

where we model the latent logic ability at the exit test $\boldsymbol{\eta}$ as depending on the latent logic ability at entrance, $\boldsymbol{\xi}$. Also, in (1) $\boldsymbol{B}$ is a matrix of regression coefficients of the endogenous latent variables; $\boldsymbol{\Gamma}$ is the matrix of regression coefficients among the endogenous and exogenous latent variables and $\boldsymbol{\zeta}$ is vector of errors.
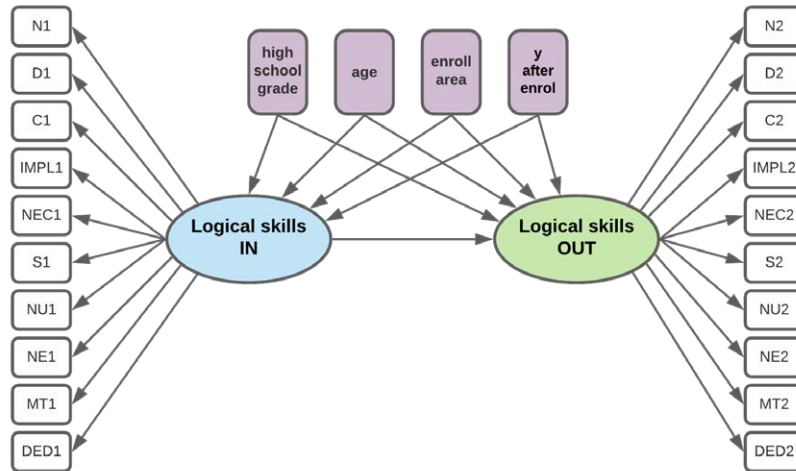
Figure 1: Structural part of the theoretical SEM.

The measurement model is defined by two equations, respectively for the endogenous (2) and exogenous (3) latent variables:

$$\boldsymbol{x} = \boldsymbol{\Delta}_{\boldsymbol{x}}\boldsymbol{\zeta} + \boldsymbol{\sigma}, \tag{2}$$

$$y = \Delta_y \eta + \epsilon, \tag{3}$$

where y is a vector of the item responses, x is a vector of exogenous individual characteristics. In both (2) and (3) Δ_y and Δ_x are matrices of factor loading while δ and σ are vectors of error terms.

Note that when one or more exogenous variables are not affected by measurement errors, the structural (1) is simplified as:

$$\eta = B\eta + \Gamma x + \zeta \tag{4}$$

3. Results

The proposed SEM with all the significant variables is reported in Figure 1. More detailed estimates are shown in Table 1. The estimated SEM presents a good fit to the data with Comparative Fit Index (CFI) equal to 0.944, Tucker-Lewis Index (TLI) of 0.935 and a Root Mean Squared Error of Approximation (RMSEA) of 0.054. All estimates were obtained using the R-package Lavaan (Rosseel, 2012).
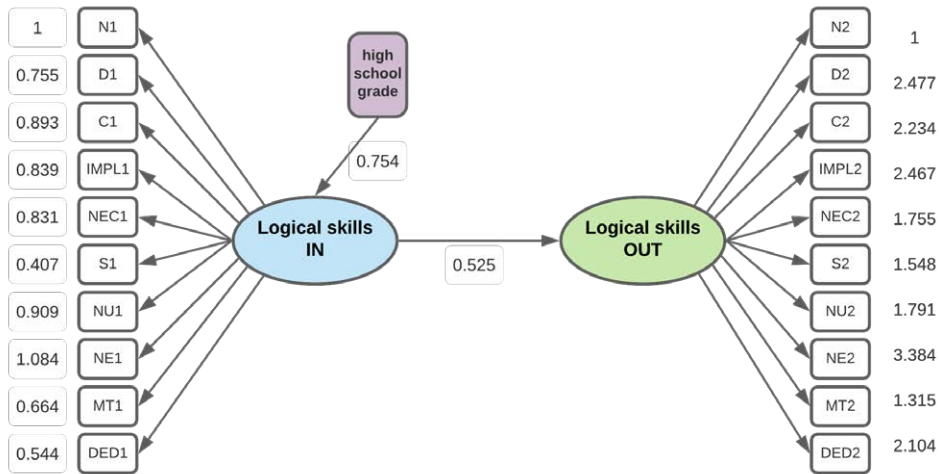
Figure 2: Final SEM.

Our results for the measurement part suggest that the course was indeed effective as the estimated items' coefficients are greater in magnitude after having attended the training course. (see Figure 1).

Regarding the regression part of the model, we found that the only significant determinant of the logical skill at entrance was the final grade obtained at the high-school. Also, we found that the only significant effect on the logical skill at exit was the logical skill at entrance. These results confirms that the short training course was indeed useful to sharp students' logical abilities and moreover it is consistent with the preliminary results obtained by Bertaccini et al. (2021).

Table 1: SEM results: measurement part, regression part, and covariances.

	Estimate	Std.Err	z-value	$P(>|z|)$	Std.lv	Std.all
Measurement:						
Dep: Ability IN						
N1	1.000				0.650	0.633
D1	0.755	0.257	2.936	0.003	0.490	0.483
C1	0.893	0.317	2.814	0.005	0.580	0.568
S1	0.407	0.277	1.471	0.141	0.265	0.264
IMPL1	0.839	0.495	1.696	0.090	0.545	0.535
NEC1	0.831	0.357	2.329	0.020	0.540	0.530
NU1	0.909	0.268	3.392	0.001	0.591	0.578
NE1	1.084	0.325	3.339	0.001	0.704	0.683
MT1	0.664	0.284	2.341	0.019	0.431	0.426
DED1	0.544	0.260	2.097	0.036	0.354	0.351
Dep: Ability OUT						
N2	1.000				0.298	0.298
D2	2.477	1.367	1.812	0.070	0.739	0.728
C2	2.234	1.275	1.752	0.080	0.666	0.659
S2	1.548	1.151	1.345	0.179	0.462	0.459
IMPL2	2.467	1.358	1.816	0.069	0.736	0.726
NEC2	1.755	1.002	1.751	0.080	0.523	0.520
NU2	1.791	1.253	1.429	0.153	0.534	0.530
NE2	3.384	1.925	1.758	0.079	1.009	0.983
MT2	1.315	0.918	1.433	0.152	0.392	0.391
DED2	2.104	1.167	1.802	0.071	0.627	0.621
Regression:						
Dep: Ability IN						
votomat	0.754	0.317	2.375	0.018	1.160	0.359
Dep: Ability OUT						
Ability IN	0.525	0.232	2.263	0.024	0.639	0.639
Covariances:						
NEC1;NEC2	0.498	0.143	3.480	0.001	0.498	0.671
C1;C2	0.412	0.125	3.296	0.001	0.412	0.644

4. Conclusions

In this paper, we took extended the previous work study by Bertaccini et al. (2021) to offer a more comprehensive and a unified framework to test the effectiveness of the training course for the development of the logical skills of students enrolling at the University of Florence. The effectiveness of the course is confirmed thus making advisable for the University of Florence to design an internal policy so that it may become a standard tool of training and evaluation.

References

Bartolucci, F., Bacci, S., and Gnaldi, M. (2019). *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Chapman and Hall/CRC.

Battauz, M. (2015). equateirt: An r package for irt test equating. *Journal of Statistical Software*, 68(1):1–22.

Bertaccini, B., Bruni, R., Crescenzi, F., and Donati, B. (2021). Measuring logical competences and soft skills when enrolling in a university degree course. In Bertaccini, B., Fabbris, L., and Petrucci, A., editors, *ASA 2021 Statistics and Information Systems for Policy Evaluation: Book of short papers of the opening conference*, volume 127. Firenze University Press.

Bollen, K. A. (1989). *Structural equations with latent variables*, volume 210. John Wiley & Sons.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Duncan, O. D. (2014). *Introduction to structural equation models*. Elsevier.

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling and more. version 0.5–12 (beta). *Journal of statistical software*, 48(2):1–36.

Clustering students according to their proficiency: a comparison between different approaches based on item response theory models

Rosa Fabbricatore, Francesco Palumbo

1. Introduction

Evaluating learners' competencies is a crucial concern in education, and home and classroom structured tests represent an effective assessment tool. Structured tests consist of sets of items that can refer to several abilities or more than one topic. Several statistical approaches allow evaluating students considering the items in a multidimensional way, accounting for their structure. According to the evaluation's ending aim, the assessment process assigns a final grade to each student or clusters students in homogeneous groups according to their level of mastery and ability. The latter represents a helpful tool for developing tailored recommendations and remediatiodddns for each group (Davino et al., 2020; Fabbricatore et al., 2021). At this aim, latent class models represent a reference.

In the item response theory (IRT) paradigm, the multidimensional latent class IRT models, releasing both the traditional constraints of unidimensionality and continuous nature of the latent trait, allow detecting sub-populations of homogeneous students according to their proficiency level also accounting for the multidimensional nature of their ability (Bartolucci et al., 2014). Moreover, the semi-parametric formulation leads to several advantages in practice: It avoids normality assumptions that may not hold and reduces the computation demanding.

However, when the interest is to accurately estimate the individual level of ability in addition to the clustering purpose, a two-step approach could be used.

In this vein, this study compares the results of the multidimensional latent class IRT models with those obtained by a two-step procedure, which consists of firstly modeling a set of unidimensional IRT models to estimate students' ability in each knowledge domain and then applying a clustering algorithm to classify students accordingly. Regarding the latter, parametric and non-parametric approaches were considered. In particular, the k-means clustering algorithm (MacQueen, 1967), the Gaussian mixture model-based clustering (McLachlan and Peel, 2000), and the archetypal analysis (Cutler and Breiman, 1994) were implemented.

The aim is to investigate similarities and differences in groups detection and students' classification. Indeed, describing students' profiles according to a set of reference groups can take many forms, depending on the adopted approach and estimation procedure.

2. Data and procedure

Data refer to the $N = 944$ subjects involved in the admission test for the degree course in psychology exploited in 2014 at the University of Naples Federico II.

The following five different domains represent the knowledge dimensions assessed by the admission test: Humanities (30 items), Reading (30 items), Mathematics (10 items), Science (10 items), and English (20 items). Correct answers receive one credit and are coded with 1, whereas blank and wrong answers receive no credit and are coded as 0.

Rosa Fabbricatore, University of Naples Federico II, Italy, rosa.fabbricatore@unina.it, 0000-0002-4056-4375
Francesco Palumbo, University of Naples Federico II, Italy, fpalumbo@unina.it

Firstly, we carried out the multidimensional latent class IRT model to cluster subjects into classes as homogeneous as possible according to their abilities, concurrently accounting for the multidimensional structure of the data. Secondly, we implemented three two-step procedures exploiting the k-means algorithm, the Gaussian mixture modeling, and the archetypal analysis, respectively. Finally, we compared the different approaches employing a graphic example and evaluated their agreement through the Adjusted Rand Index (ARI; Hubert and Arabie, 1985). The ARI is a commonly used measure to evaluate distances in clustering. It allows comparing a partition with another one on the same elements or with external criteria. Index computation is based on the number of pairs of elements that are allocated in the same (or different) cluster in both partitions (agreements) and the number of pairs of elements that are placed in the same cluster in one partition but in different clusters in the other (disagreements). The ARI values range from 0 (random partitioning) to 1 (partitions perfect agreement).

3. Statistical method

Methods we compared in this study exploit IRT models for students' ability estimation (see Bartolucci et al. (2019) for a review on the IRT models). In more detail, we considered the two-parameter logistic (2PL) IRT parametrization, where the parameters of guessing and ceiling are constrained to be equal to 0. Thus the probability of correct response depends only on the discrimination and difficulty item parameters and the student's ability. More formally, the probability that the subject s correctly answers the dichotomously-scored item i (with $i = 1, \ldots, I$) can be expressed as follows:

$$P(X_{si} = 1 | \theta_s, a_i, b_i) = \frac{e^{a_i(\theta_s - b_i)}}{1 + e^{a_i(\theta_s - b_i)}}, \qquad (1)$$

where X_{si} is the response of the subject s at the item i with realization $x_{si} \in [0, 1]$, $\theta_s \in R$ is the ability of the subject s, $a_i \in R$ is the item discrimination parameter, and $b_i \in R$ represents the item difficulty. It is worth noting that traditional IRT models are ground on three main assumptions: unidimensionality, monotonicity, and local independence. Moreover, the latent trait is described by a continuous normal probability distribution.

Within this theoretical paradigm, the multidimensional latent class IRT models represent a semi-parametric formulation of the traditional IRT models, allowing releasing both the constraints of unidimensionality and the continuous nature of the latent trait. This extension is particularly useful for detecting sub-populations of homogeneous students according to their ability level.

Since we defined the ability as a multidimensional latent trait, each subject is described by the ability vector $\mathbf{\Theta}_s = (\Theta_{s1}, \Theta_{s2}, \ldots, \Theta_{sD})'$ where D is the number of considered dimensions. Following the between-item multidimensional formulation, each item measures only one dimension, and thus items are divided into different subsets I_d with $d = 1, 2, \ldots, D$.

Moreover, according to the semi-parametric formulation, each latent trait have a discrete distribution with $\xi_1, \ldots, \xi_k$ support points defining k latent classes with weights $\pi_1, \ldots, \pi_k$. The main assumption is that subjects in the same latent class share common levels of the latent trait. The generic class weight π_c (with $c = 1, \ldots, k$) represents the probability of belonging to class c and can be expressed as $\pi_c = P(\Theta_s = \xi_c)$ with $\sum_{c=1}^{k} \pi_c = 1$ and $\pi_c \geq 0$.

Accordingly, the manifest distribution of the response vector $\mathbf{X} = (X_1, \ldots, X_I)'$ can be formalized as:

$$P(\mathbf{X} = \mathbf{x}) = \sum_{c=1}^{k} P(\mathbf{X} = \mathbf{x} | \mathbf{\Theta} = \xi_c) \pi_c = \sum_{c=1}^{k} \prod_{d=1}^{D} \prod_{i \in I_d} P(X_i = x_i | \Theta_d = \xi_{cd}) \pi_c, \qquad (2)$$

where $P(X_i = x_i | \Theta_d = \xi_{cd})$ It is herein specified according to the 2PL parameterization.

The number of classes k can be derived from theoretical assumptions or by comparing the model fit measures at different values of k. Each unit was assigned to the class that corresponds to the highest probability of belonging.

The estimation of the model parameters is usually based on the Maximum Marginal Likelihood (MML) approach. In the specific case of the latent class formulation, the Expectation-Maximization (EM) algorithm is used (Dempster et al., 1977). The estimation process is performed through the R packages `mirt` (Chalmers, 2012) and `MultiLCIRT` (Bartolucci et al., 2014) for the parametric and semi-parametric IRT formulation, respectively.

As stated before, the latent class IRT models allow removing parametric assumptions that may not hold and make the estimation process computationally demanding. Moreover, they are more flexible than the parametric formulation when the main aim is clustering individuals. However, this semi-parametric formulation provides a less accurate estimate of the individual level ability than the continuous one.

Regarding the clustering algorithms applied on the ability estimates, a very brief description below. The *k-means* produces a hard clustering changing the data partition at each step taking into account the Euclidean distance of each point from the cluster centers. It is one of the most used algorithms in cluster analysis mainly due to its ease of implementation and interpretation. Nevertheless, the k-means algorithm works well only when dealing with spherical clusters and no outliers are present in the data set. Firstly, accounting only for clusters' centroids is not suitable enough to properly detect subpopulations that also have covariance parameters significantly different. Secondly, centroids could be dragged by outliers.

Overcoming these issues, the *Gaussian mixture model* provides a model-based clustering allowing to detect differences between sub-populations that share the same (Gaussian) distribution but have one or more different vectors of parameters; thus, these models estimate a specific covariance matrix for each cluster and better manage the presence of outliers. On the other hand, they could entail the risk of overparameterization: increasing model complexity does not guarantee a better solution to the classification problem.

Compared to the methods mentioned above, the archetypal analysis allows more separate groups, detecting extreme representative observations that differ from each other as much as possible. Consequently, it approximates each point in a dataset as a convex combination of this set of extreme data points, called archetypes, lying on the convex hull of the data. Conversely, drawbacks reside in its computation costs, especially as the number of observations increases.

The corresponding R packages used to carry out the analyses were `stats`, `mclust` (Scrucca et al., 2016), and `archetypes` (Eugster, 2009).

4. Results

A set of multidimensional latent class IRT models with a different number of latent classes k were estimated. Basing on the Bayesian information criterion (BIC; Schwarz, 1978), we chose the model with $k = 3$ as the best one for describing our data. Looking at support points, we notice that latent classes are decreasing ordered according to the students' proficiency levels in all the considered domains (see Table 1). In particular, Class 1 encompasses students with poor performance in all the six domains; Class 2 includes students with low performance in Humanities, Math, Science and English, and high performance in Reading; Class 3 consists of students with a good performance in all the domains except for Humanities for which they achieved an average performance. Class weights indicate that Class 2 (moderate ability) is the largest one ($\pi_2 = 0.48$), followed by Class 1 (higher ability; $\pi_1 = 0.39$).

This result was compared with students' classifications obtained by a two-step procedure.

As stated before, the comparison involved different clustering algorithms that were carried out on the students' ability estimates provided by the set of unidimensional IRT models (see Table 1). It is worth noting that the number of classes was imposed equal to $k = 3$ in all the clustering procedure for comparison purposes.

Table 1: Standardized support points (Latent Class IRT), centroids (K-means), component means (Gaussian Mixture Model), archetypes (Archetypal Analysis), and class weights (π_c) for each clustering approaches.

		Latent Class IRT	K-means	Gaussian Mixture Model	Archetypal Analysis
Class 1	Humanities	-0.91	-0.52	-0.50	-0.54
	Reading	-0.17	0.23	0.57	0.01
	Math	-0.81	-0.36	-0.08	-0.87
	Science	-0.75	-0.27	0.11	-0.59
	English	-1.22	-0.68	-0.28	-1.22
Class 2	Humanities	-0.44	-0.38	-0.36	-1.01
	Reading	0.32	0.42	0.36	0.26
	Math	-0.24	-0.15	-0.36	0.22
	Science	-0.10	0.03	-0.09	-0.18
	English	-0.21	-0.05	-0.06	-0.17
Class 3	Humanities	-0.06	-0.24	-0.33	0.21
	Reading	0.76	0.58	0.47	0.92
	Math	0.35	0.15	0.15	0.47
	Science	0.51	0.28	0.18	0.79
	English	0.72	0.53	0.16	1.29
Class weight	π_1	*0.13*	*0.21*	*0.08*	*0.35*
	π_2	*0.48*	*0.42*	*0.43*	*0.24*
	π_3	*0.39*	*0.37*	*0.49*	*0.41*

The example reported in Figure 1, showing only two of the five ability dimensions for simplicity and lack of space, allows depicting differences in students allocation due to the considered clustering approaches. As can be guessed from the picture, the multidimensional latent class IRT model reached the strongest agreement in terms of classification when the k-means algorithm was implemented ($ARI = 0.53$). The confusion matrix showed that the main difference resided in a higher allocation rate in class 2 rather than in class 1 for the multidimensional latent class IRT model compared to the approach based on k-means. A weaker agreement was found with the archetypal analysis ($ARI = 0.39$), whereas the lowest one was reported with the parametric approach based on Gaussian mixture modelling ($ARI = 0.09$).

Notice that in addition to the ability estimates in Table 1, the considered clustering approaches also differ for the allocation procedure that strongly influences the level of agreement between the partitions and, consequently, the ARI.

(a) multidimensional latent class IRT

(b) k-means

(c) Gaussian mixture model

(d) archetypal analysis
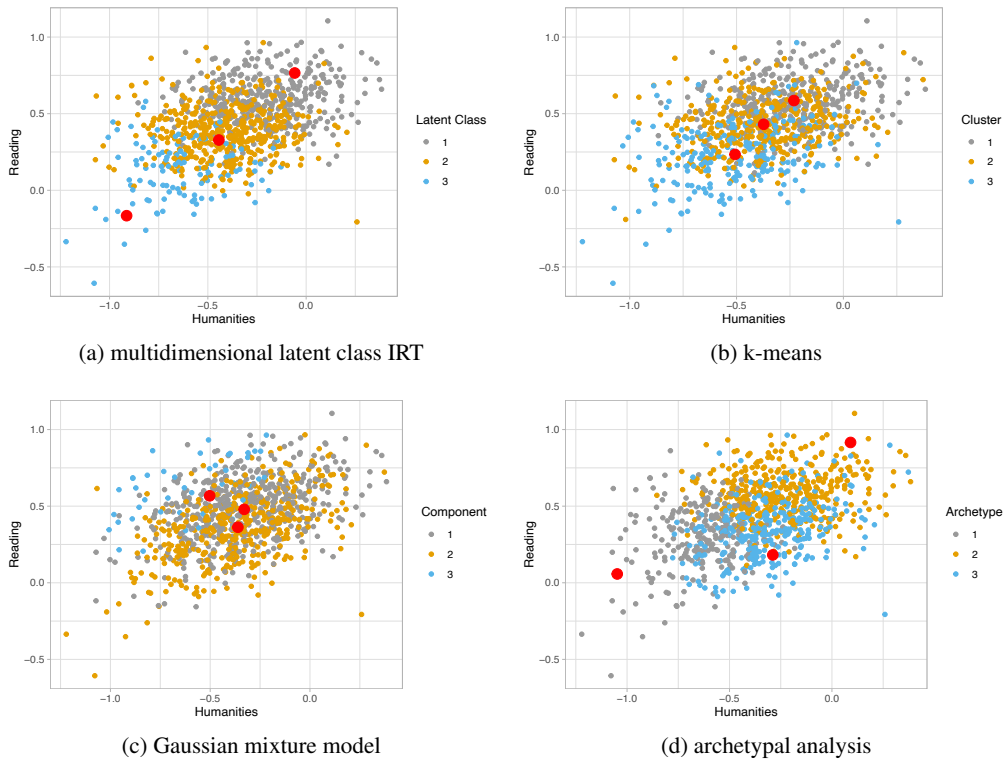
Figure 1: Students allocation based on different clustering approaches. X-axes and y-axes refer to students' ability estimation through the multidimensional IRT model in Humanities and Reading, respectively. According to the considered method, red points indicate: (a) standardized support points, (b) centroids, (c) component means, and (d) archetypes.

5. Conclusion

The study provides a useful insight in understanding dissimilarities between different approaches used for clustering purposes. Assuming as a matter of the fact that the adequacy of a method mainly depends on research goals and thus that there is not the best one in absolute terms, we compared different approaches illustrating which of the clustering algorithm we considered in the two-step procedure provides results more similar to those obtained by the multidimensional latent class IRT model.

The proposed comparison also invokes the difference between the parametric and semiparametric formulation of IRT models in practical applications.

Future research should investigate how the considered approaches work when a different data structure holds. Moreover, it would be interesting also to consider differences deriving from classical test theory rather than the IRT paradigm for the ability estimation.

References

Bartolucci, F., Bacci, S., Gnaldi, M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis*, **71**, pp. 971–985.

Bartolucci, F., Bacci, S., Gnaldi, M. (2019). *Statistical analysis of questionnaires: A unified*

approach based on R and Stata. Chapman and Hall/CRC, London, (UK).

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, **48**(6), pp. 1–29.

Cutler, A., Breiman, L. (1994). Archetypal analysis. *Technometrics*, **36**(4), pp. 338–347.

Davino, C., Fabbricatore, R., Pacella, D., Vistocco, D., Palumbo, F. (2020). Aleas: a tutoring system for teaching and assessing statistical knowledge, in *Proceedings of the Second Symposium on Psychology-Based Technologies*, eds. O. Gigliotta and M. Ponticorvo, CEUR Workshop Proceedings, 2730.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), pp. 1–22.

Eugster M.J.A., Leisch, F. (2009). From Spider-Man to Hero – Archetypal Analysis in R. *Journal of Statistical Software*, **30**(8), pp. 1–23.

Fabbricatore, R., Parola, A., Pepicelli, G., Palumbo, F. (2021). A latent class approach for advising in learning statistics: implementation in the ALEAS system, in *Proceedings of the First Workshop on Technology Enhanced Learning Environments for Blended Education - The Italian e-Learning Conference 2021*, eds. P. Limone and R. Di Fuccio, CEUR Workshop Proceedings, 2817.

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of classification*, **2**(1), pp. 193–218.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, eds. L. Lecam and J. Neyman, Cambridge University Press, Oakland, (CA), pp. 281–297.

McLachlan, G.J., Peel, D. (2000). *Finite mixture models*. Wiley-Interscience, New York, (NY).

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, **6**(2), pp. 461–464.

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, **8**(1), pp. 289–317.

Sustainable Innovation: worldwide trends in the scientific production through a bibliometric study

Rosanna Cataldo, Corrado Crocetta, Maria Gabriella Grassia, Paolo Mazzocchi, Antonella Rocca, Claudio Quintano

1. The Sustainable Innovation

The scientific production on the Innovation, especially on Sustainable Innovation, has grown in recent years. Research on sustainable innovation has expanded rapidly in order to understand how new technologies can make societies more sustainable.

Various expressions and definitions for sustainability and innovation have been reported in the literature. Sometimes the two concepts are combined and described with one term, *Sustainable Innovation*. Research on sustainable innovation has grown in popularity due to the need to incorporate sustainability within business practices (Boons and Lüdeke-Freund, 2013). Innovation that is seen not only as a tool to guarantee a competitive advantage for companies but also as a tool that provides environmental benefits and produces social well-being (Cillo et al., 2019).

Tello and Yoon (2008) define the Sustainable innovation as "the development of new products, processes, services and technologies that contribute to the development and well-being of human needs and institutions while respecting natural resources and regeneration capacities". Several studies have focused on sustainable innovation and they stated that sustainable innovation can be studied on the basis of three main perspectives: internal managerial, external relational and performance evaluation (Cillo et al., 2019).

The paper contributes to the literature on sustainable innovation by providing the worldwide trend in the scientific production over time through a research conducted on the metadata of Web of Science, the main database commonly used by researchers. A bibliometric analysis has been developed to analyse a total of 1,511 documents published between 2000 and 2021 in order to discover the research trends in this field and the main dimensions and words related to the term "Sustainable Innovation".

2. Methodology

A bibliometric analysis has been used to explore the evolution of research in the innovation field. Bibliometric analysis is a quantitative approach for the analysis of academic literature using bibliographies to provide the description, evaluation and monitoring of the published research (Garfield et al., 1964); (White and McCain, 1989).

The methodological aim is to analyze publications, citations and sources of information (Rodriguez - Soler et al., 2020). The scientific community has always used bibliometric methods as a tool for analysis. For this study, the Bibliometrix package (Aria and Cuccurullo, 2017), in the R programming language (https://www.r-project.org/) was used. This recent R-package

Rosanna Cataldo, University of Naples Federico II, Italy, rosanna.cataldo2@unina.it, 0000-0002-6324-8252
Corrado Crocetta, University of Foggia, Italy, corrado.crocetta@unifg.it, 0000-0001-9059-5092
Maria Gabriella Grassia, University of Naples Federico II, Italy, mariagabriella.grassia@unina.it, 0000-0002-7128-7323
Paolo Mazzocchi, University of Naples Parthenope, Italy, paolo.mazzocchi@uniparthenope.it, 0000-0002-6632-314X
Antonella Rocca, University of Naples Parthenope, Italy, antonella.rocca@uniparthenope.it, 0000-0001-8171-3149
Claudio Quintano, University of Naples Suor Orsola Benincasa, Italy, claudio.quintano@unisob.na.it, 0000-0001-8315-8476

provides a set of tools for quantitative research in bibliometrics and scientometrics, supporting scholars in all key phases of analysis from the data importing to the visualization of results.

3. Analysis

The data for this research project were collected in the Web of Science's database of the Institute for Scientific Information (ISI). Web of Science (WoS) is the world's most trusted independent global citation database. It is recognised as covering a broad range of relevant journals and peer-reviewed articles of high quality (Cataldo et al., 2019).

To collect the documents published on this topic field in the past 20 years, we queried the WoS database on July 7th, 2021. A total of 1,511 documents published between 2000 and 2021 (incl.) containing the topic "Sustainable Innovation" were retrieved. The majority (907; 60%) were research articles. The second most common type of documents was proceeding papers which constituted 32.43%. Details about documents were shown in Table 1. Those documents show an average citation per documents of 13.08 in the considered period and were written by 3,897 authors from 663 different sources, such as journals, books, etc. The author's keyords are 3,694, while the keywords plus are 2,341.

According to Garfield (1990), the keyword plus "provides search terms extracted from the titles of papers cited in each new article in the ISI database, is an independent supplement for title-words and author keywords". The collaboration index, that represents the mean number of authors per joint paperand is calculated as total authors of multi-authored articles/total multi-authored articles (Elango and Rajendran, 2012), is equal to 2.96. It implies the research team falls between 2 and 3 in the field of sustainable innovation.

Table 1: General Information

About Data	Timespan	2000-2021
	Sources (Journals, Books, etc)	663
	Documents	1511
	Annual Growth Rate	16.42 %
	Average citations per documents	13.08
Document Types	Articles	907
	Books	5
	Books reviews	11
	Editorial materials	37
	Proceedings papers	490
	Reviews	61
Document Contents	Author's Keywords	3694
	Keywords Plus	2341
Authors	Authors	3897
	Authors of single-authored documents	197
	Authors of multi-authored documents	3700
Authors Collaboration	Documents per Author	0.388
	Authors per Documents	2.58
	Collaboration Index	2.96

Figure 1 (a) presents the annual-trends of publications, indicating that sustainable innovation literature has been growing since 2007, peaking in 2018 with 333 documents published. Generally, almost 17% of annual growth rate has been observed in the production of research articles during the study period (see Table 1). Figure 1 (b) shows the annual number of citations. The works published in the first years of analysis have accumulated a lot of recognition. It is possible to note that the average of citations in 2002 was equal to 4.74, and a similar average is reached in 2017.

(a) Annual scientific production   (b) Average Citations per Year
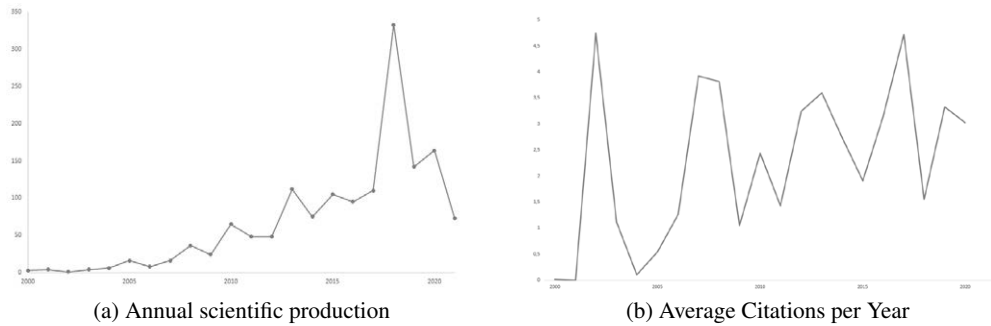
Figure 1: Scientific production (2000-2021), n=1511

Figure 2 shows the main ten sources of pubblication. The first source is a book with the title "A Creative Path to Sustainable Innovation" related to the Siam Physics Congress 2018 (SPC2018) with 190 documents published. From Figure 2 it is possible to note that the most relevant sources, based on the number of articles, are *Sustainability*, *Journal of Cleaner Production*, *Green Technologies for Sustainable & Innovation in Materials*, journals whose aims are to provide up-to-date information on new developments and trends in relation to this topics.

Figure 2: Main sources of pubblication

Figure 3 shows the number of articles produced by the authors of different countries and the rate of cooperation of each country's authors with other countries' authors (SCP: Single Country Publications; MCP: Multiple Country Publications).

Thailand produced a large number of papers in the analysis period, showing a rather low collaboration rate (MCP) with authors from other countries. This means that the Thai authors who write on this topic do not collaborate with foreign researchers. The USA, despite having the same number of documents as Thailand, has a higher MCP than Thailand. England is the nation with the highest rate of collaboration with foreign authors, followed by China and Netherlands. These links are highlighted in the Figure 4.

In the network the size of the circle of the country is related to the number of works published on the analyzed topic, the different colors of the countries and of the links represent the clusters that have been formed, as determined by the Louvain algorithm, while the strength of the collaboration is indicated by the thickness of the links (Crocetta et al., 2021). The networking analysis emphasizes the strong collaboration of the USA with China. USA collaborates with almost all the countries shown in the network, except for some such as Malaysia and Portugal.
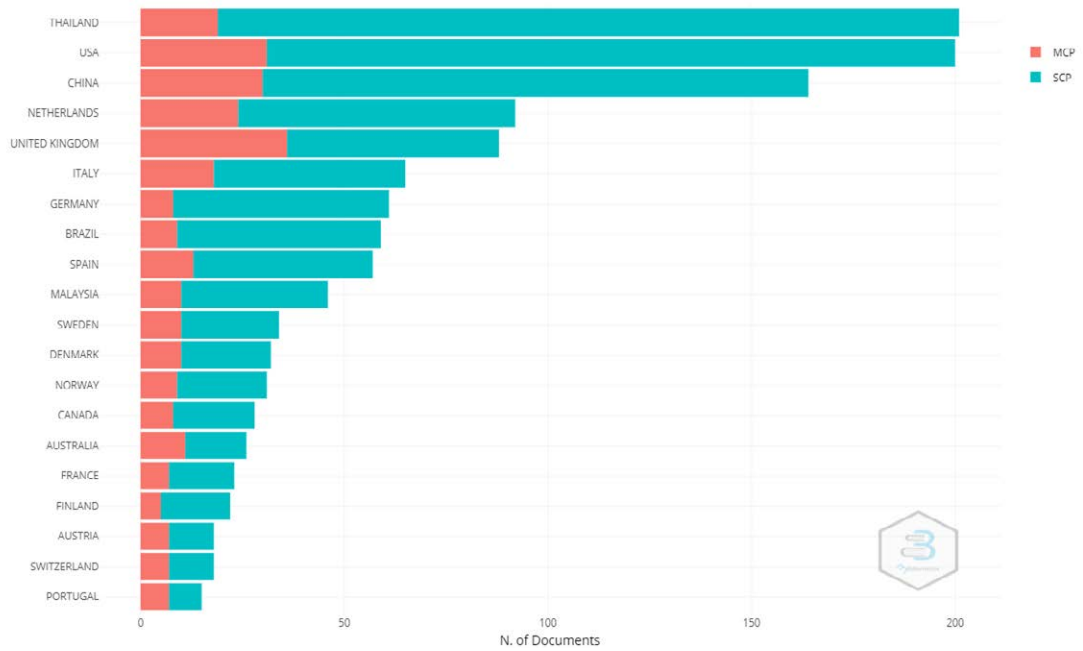
Figure 3: Corresponding Author Country
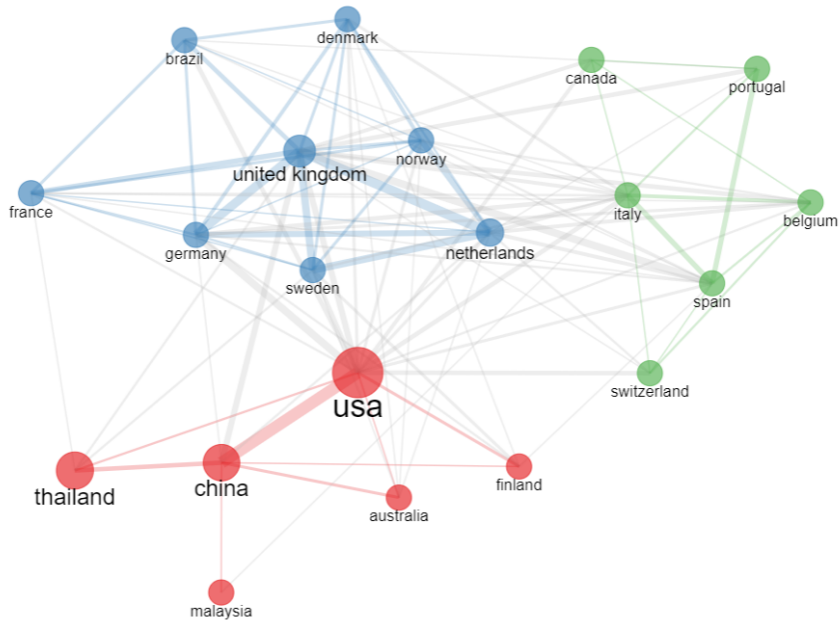
Figure 4: Collaboration Network

In the network we can see that there are only five connections from Thailand (to USA, China, Sweden, United Kingdom and France) and this reinforces what has been said about the low rate of collaboration.

The last Figure, Figure 5, is the thematic maps, an intuitive plot in which author's keywords

are viewed as themes, classified by different levels of density (which represents the development degree) and centrality (which represents the relevance degree) in the network of scientific keywords (Cataldo et al., 2019).
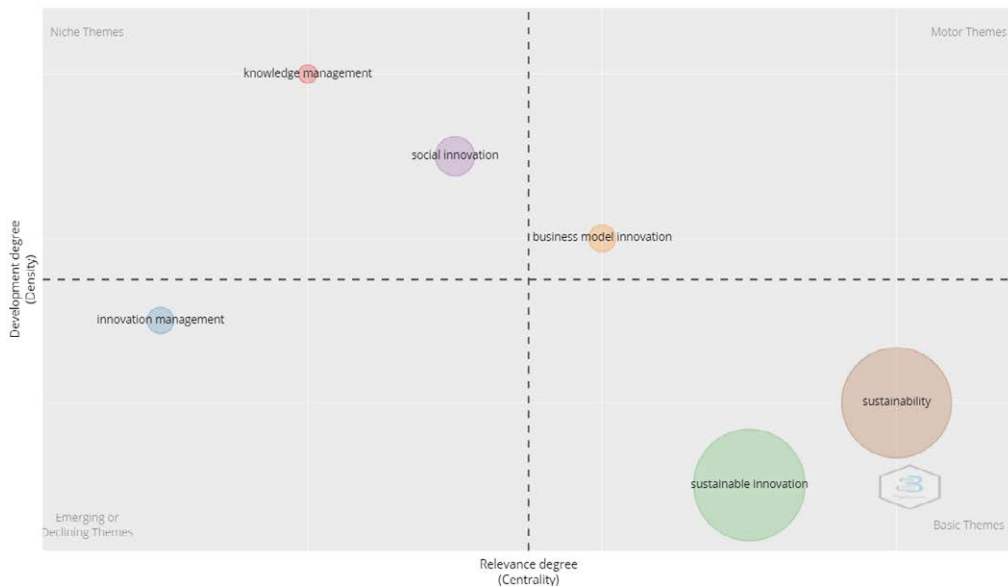
Figure 5: Thematic Map

The cluster named "business model innovation" represents the motor theme, topic that is developed and relevant to the research field. In this cluster there are keywords such as "innovation system", "barriers", "innovation ecosystems", "drivers", "green buildings". The themes such as "sustainability" and "sustainable innovation" represent the basic themes, topics that appear ubiquitously in different scientific works and can be considered a common synthesis of the content expressed in the literature. The cluster named "innovation management" is positioned as emerging or declining themes, because this cluster is formed by keywords that are weakly developed and marginal. This cluster includes keywords such as "big data", "environment", "research and development". Finally, the cluster "knowledge management" represents the isolated theme. It is formed by keywords such as "product innovation", "literature review", "organizational learning", "process innovation", all topics that are of limited importance for the research topic.

4. Final remarks

The main purpose of this paper was to review the literature related to the sustainable innovation. This study has tried to provide a comprehensive view of scientific papers between 2000 and the first six months of 2021 in this research field. In doing so, we identify 1,511 documents found relevant in the Web of Science database by using the keyword "sustainable innovation". The scientific production has grown very gradually over the years reaching a peak of 333 products in 2018, in the previous year there were only 110. This shows that until a few years ago the concept of sustainable innovation was not yet widespread in the scientific community.

This research has shown the Thailand, USA and China have been the most productive countries in this area. In particular, the main authors who write on this topic are from Thailand and

collaborate with each other, showing a very low collaboration rate with foreign researchers. British researchers, on the other hand, are those who collaborate with authors from different countries. The thematic map analysis has identified that cluster of "business model innovation" is the motor theme in this research field, while the cluster of "innovation management" has been emerging or declining theme. It must be said that the theme analyzed in this work is a fairly new and constantly evolving theme in literature. Therefore the results of this bibliometric analysis could be different in a few years. Furthermore, the analysis was carried out only with documents downloaded from the web of science, so it could be more global using other scientific databases. However, we hope the present study may assist researchers in investigation this theme in their researches.

References

Aria, M., and Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics*, Elsevier, **11**(4), pp 959–975

Boons, F., and Lüdeke-Freund, F. (2013). Business models for sustainable innovation: state-of-the-art and steps towards a research agenda. *Journal of Cleaner production*, Elsevier, **45**, pp 9–19.

Cataldo, R., Grassia, M.G., Lauro, C.L., Marino, M., and Voytsekhovska, V. (2019) A Bibliometric Study of the Global Research Activity in Sustainability and Its Dimensions. *International Conference on Data Science and Social Research*, Springer, pp 91-102

Cillo, V., Petruzzelli, A., Ardito, L., and Del Giudice, M. (2019). Understanding sustainable innovation: A systematic literature review. *Corporate Social Responsibility and Environmental Management*, Wiley Online Library, **26**(5), pp 1012–1025.

Crocetta, C., Antonucci, L., Grassia, M. G., Marino, M., and Cataldo, R. (2021). A bibliometric study of global research activity in relation to the use of partial least squares for policy evaluation. In *ASA 2021 Statistics and Information Systems for Policy Evaluation*, Firenze, **127**, pp. 49–54).

Elango, B., and Rajendran, P. (2012) Authorship trends and collaboration pattern in the marine sciences literature: a scientometric study. *International Journal of Information Dissemination and Technology*, Maharishi Markandeshwar University, **2**(3), pp 166–169.

Garfield, E. (1990). KeyWords Plus-ISI's breakthrough retrieval method. 1. Expanding your searching power on current-contents on diskette. *Current contents*, INST SCI INFORM INC 3501 MARKET ST, PHILADELPHIA, PA 19104 USA **32**, pp 5–9.

Garfield, E., Sher, I.H., and Torpie, R.J. (1964). The use of citation data in writing the history of science. *Institute for Scientific Information Inc Philadelphia PA*

Rodriguez-Soler, R., Uribe-Toril, J., and Valenciano, J. D. P. (2020). Worldwide trends in the scientific production on rural depopulation, a bibliometric analysis using bibliometrix R-tool. *Land Use Policy*, 97, 104787.

Tello, S.F., and Yoon, E. (2008). Examining drivers of sustainable innovation. *International Journal of Business Strategy* Citeseer, **8**(3), pp 164–169.

White, H.D., and McCain, K.W. (1989) Bibliometrics. *Annual review of information science and technology*, **24**, pp 119–186

Personal weaknesses recognized by high school students in the North-West of Italy

Luigi Bollani

1. Introduction

This study is part of a project aimed at supporting the weaknesses of young people who are experiencing, or are at risk of reaching, NEET condition (Not in Education, Employment or Training; aged between 15 and 29 or 34, according to different definitions).

At present the entire project consists of three phases. The first, concluded in 2020 with the volume "From Neet to Need" (see also Bollani, Rota, 2018), presents the European statistical situation of the Neet phenomenon and its sociological implications; it includes an empirical study, based on thorough interviews with people in Neet condition intended to collect their life stories (Merril, West, 2012); the purpose is to classify them as to their internal and external needs, in order to suggest adequate support for intervention. The second phase, to be completed in 2021 with a second book, is the object of this study: it searches for harbingers of Neet status in the difficulties encountered during high school (which in Italy is divided into an initial two-year set and a subsequent three-year set); in fact, in general it is more simple and effective to tackle minor problems by a preventive approach, starting from school age. Of course students as such are not Neets and the anonymity required for a survey dealing with very personal issues does not allow a longitudinal study to be carried out on the same individuals. However, the construction of a comparison base referring to a generic population, relative to the incidence of certain states of weakness (or some of their combinations), will allow for a subsequent comparison with the incidence of the same difficulties in school age for those who find themselves in a Neet situation. The third phase, started in 2021, is about identifying good practices in order to grow Neets into working adults; the people selected to accompany the Neets in this process are now fully trained, while the first groups of Neets will be involved by the end of the year. The research group was identified among members of the InCreaSe Association (Innovation Creativity Settings; www.increasegroup.org), which boasts transversal research skills. Numerous local authorities are involved in the project, which is supported by Compagnia di San Paolo Foundation.

As to the specific object of this paper, the first signs of the extreme discomfort caused by Neet condition were sought by investigating signs of weakness in high school students in the Piedmont, Valle d'Aosta and Liguria regions. A survey questionnaire was administered to students, an operator being present in the classrooms, shortly before the start of the pandemic. Collected results will therefore be more directly linked to school activities "in presence", even if discontinuities in teaching methods and characteristics caused by the pandemic might have brought in alterations that will need to be monitored over time.

2. Themes of investigation and subjects involved

School transmits knowledge accumulated and elaborated by society over the centuries, allowing younger generations to experience social, cultural and territorial belonging and to be citizens of the society with its rights and duties.

The proposed survey is embedded in the context of training institutions challenged to cope with rapid change and face uncertainty about the impact of offered training courses on personal

growth and access to work in the current context. The intention of the study is to highlight positive states of well-being, certainly present in a school context, but at the same time identify situations of discomfort among students which, if not adequately prevented, can develop in a harmful way. It was decided to address the issue from within the school itself, administrating questionnaires to students engaged in secondary education.

The topics of the survey, included in the questionnaire within specific sections, were:

• You and the school (role attributed to the school)
• The relationship with your school
• The relationship with your teachers
• The relationship with your schoolmates
• You and your future (for the three-year period only)
• The relationship with your family
• Personal and family data

Special attention was devoted to the field organization, with an operator present in each class during the survey in order to ensure correct presentation of the questionnaire and homogeneous assistance as well as sense of anonymity towards teachers during its compilation.

It should be mentioned that the Italian secondary education system is divided into an initial compulsory two-year period ("biennio" below) and a subsequent voluntary three-year period ("triennio"); it is also divided into a more academic (classical or scientific) school called "liceo", a technical school ("tecnico") and a vocational school ("professionale").

Students were selected following a multi-stage process, considering schools and classes as first and second level units and students as third level units. The first stage units, i.e. the schools, reflect different territorial situations, with several focuses in the three regions. In particular, for Piedmont distinctions were made according to size and location of the centres and among Turin city areas with different socio-economic patterns. The choice of second stage units, i.e. classes, allowed to reach, to a certain extent, an overall balance among different educational paths; finally, for third stage units, i.e. students, a complete survey of the students present in classrooms at the time of the survey was opted for. In this way, 14 schools (some classes for each) were surveyed, in line with the three Italian regions of interest. In Piedmont 10 schools are considered: three are in Turin city (see Torino1, 2 and 3 in table 1), two in the city belt (in the municipalities of Nichelino and Settimo Torinese), five in other areas of Piedmont (one in the municipality of Pinerolo, close to Turin, and four in different provinces). In Valle d'Aosta and Liguria, two schools were surveyed (see Aosta1 and 2 for Valle d'Aosta and two schools in the municipalities of Genova and Savona for Liguria). Globally, 931 students were surveyed, keeping a sufficient number for each type of educational path, as shown in table 1.

Table 1 - Number of students surveyed by education type and school area

	Biennio Liceo	Biennio Professionale	Biennio Tecnico	Subtotal for Biennio	Triennio Liceo	Triennio Professionale	Triennio Tecnico	Subtotal for Triennio	**Total**
Torino1	17			17	32			32	**49**
Torino2			20	20			24	24	**44**
Torino3		16		16		16		16	**32**
Nichelino	14		34	48	37		15	52	**100**
SettimoT.se		24	16	40		33	12	45	**85**
Pinerolo		18		18			16	16	**34**
Asti			40	40			18	18	**58**
Cuneo		43	25	68	37	19	23	79	**147**
Novara	45			45	24			24	**69**
Vercelli		15		15		39		39	**54**
Aosta1			32	32			35	35	**67**
Aosta2		37		37		32		32	**69**
Genova		17		17			27	27	**44**
Savona			58	58			21	21	**79**
Total	**76**	**170**	**225**	**471**	**130**	**139**	**191**	**460**	**931**

3. Growth of weakness as an accumulation process

A first research objective concerns how weakness is formed and how it can grow. One possibility that also seems to emerge from life stories is that there is an accumulation phenomenon. A perhaps tolerable weakness emerging for a particular aspect of life in school age may be aggravated by a second circumstance of discomfort, and so on, with the consequence of reaching a very intense state of discomfort due to a sum of causes, each of which is not particularly relevant.

In this perspective, the approach used for this paper consists in considering the key questions of each section of the questionnaire, which, each in view of a specific aspect, detected a condition of well-being or weakness of the answering student. The number of difficult situations encountered for each student was then added; this sum enabled to consider further degrees of individual difficulty.

In particular, ten questions were examined, two for each section of the questionnaire; they are presented below; labels used in figures and the percentage of students in a state of weakness (see ending "_yes") for each of them appear in brackets.

As regards the first part of the questionnaire, related to the role students attribute to the school and the relationship they have with it (first two sections), the selected questions were on:
- the importance of school in one's life (Life_yes: 4.40%)
- the school's ability to foster personal growth (Growth_yes: 13.43%)
- I thought about changing school (Change_yes: 47.91%)
- prolonged absences (Absences_yes: 30.08%)

An aspect fundamental in determining the categories of students most in difficulty certainly is the one concerning the relationship with classmates, which if problematic can lead to important episodes of exclusion and marginalization; thus the chosen questions were whether they:
- were teased or isolated by school friends (Isolated_yes: 24.92%)
- go out with classmates outside school hours (Outside_yes: 33.30%)

Other elements taken into consideration are the average grade and the relationship with their teachers; the selected questions concerned:
- whether they ask teachers questions when they do not understand (Questions_yes: 39.74%)
- which was their average grade in the last term (Grades_yes: 48.44%)

Finally, two questions were chosen from the section about the relationship with one's family, i.e. whether:
- students feel supported by their family (Support_yes: 16.33%)
- they have confidence with any adults (Confide_yes: 21.27%)

The answers to each of these ten questions were coded in terms of presence or absence of weakness; the resulting synthetic variable indicates the amount of weakness situations, which can vary from zero (absence of weakness) to ten (maximum weakness intensity) for each answering student. The frequency distributions of this variable, called "intensity of weakness", for all respondents and for those living in different territorial areas of interest for the survey, are presented in figure 1.

The shapes of the frequency distributions shown in the figure are very similar, although Turin belt and Liguria display the highest average levels of weakness intensity (3.0 and 3.1) and city of Turin and belt of Turin show the highest variability (standard deviation 2.0 and 1.9).

In all cases, at higher degrees of weakness frequencies progressively decrease. Referring to a more stable condition, intensity of weakness is also measured more synthetically as follows, assuming to set only three qualitative degrees of weakness: students with intensity of weakness 0 and 1 are categorized as students "without weakness" (or nearly so), those with scores 2 and 3 are considered to have "lower weakness", while students with a score higher than 3 fall into the "greater weakness" category. The subdivision into bands respects the subdivision into tertiles, considering the variable starting from the most disadvantaged in the general total column

(obtaining following frequencies: greater weakness 33.94%; lower weakness 39.53% and without weakness 26.53%).

Figure 1 – Frequency distributions of the intensity of weakness (0-10)
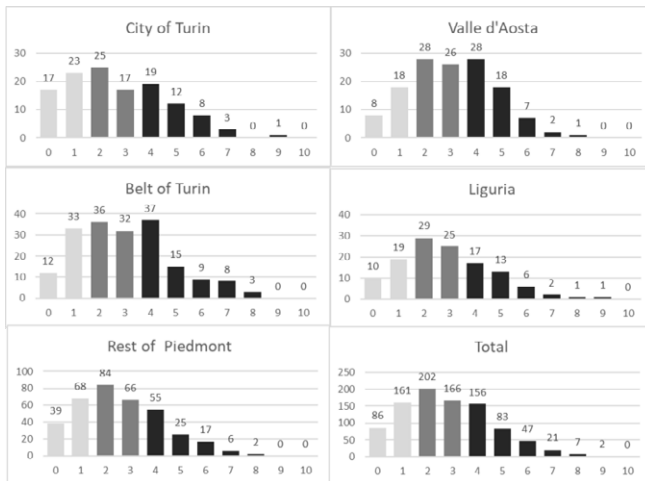
In order to check the usefulness of the considered variable empirically, as to the degree of weakness, a multiple correspondence analysis (MCA) was performed on the binary codified ten variables used for its construction; the intensity of weakness, as both ten level and three synthetic level variables, are maintained as supplementary. The MCA map is shown in Figure 2; software R and FactoMineR package were used for the analysis (Escofier, Pagès, 2008).

Figure 2 – Relationships among the ten main variables of the questionnaire (active variables are shown in the left box, while supplementary ones, relating to the intensity of weakness, are shown in the two boxes on the right)
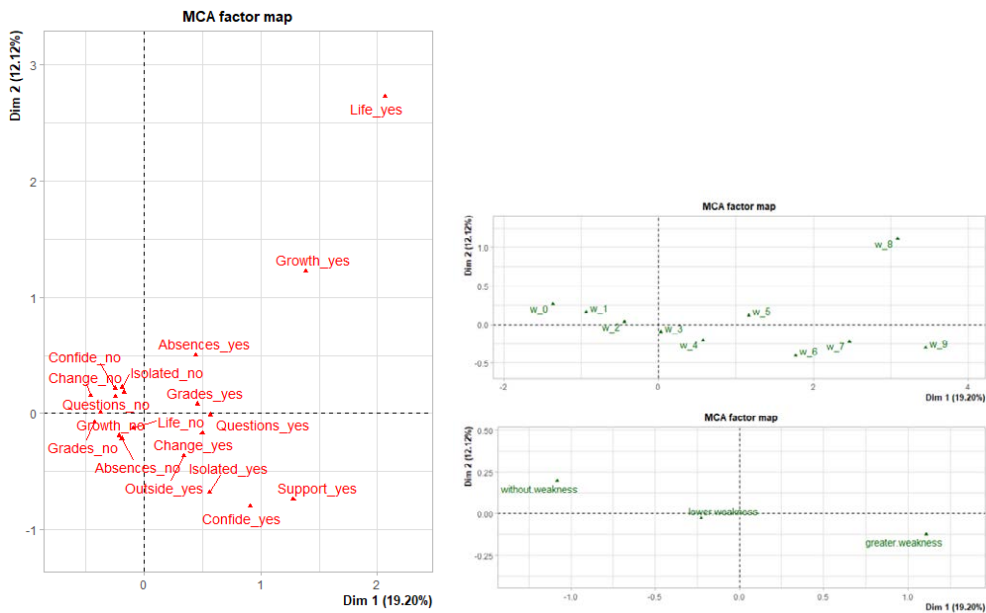
Note: the MCA, given its coding scheme, underestimates the percentage of inertia explained by the first factorial dimensions (Abdi, Valentin, 2007). In literature, some methods of re-evaluation of variances explained by first factorial dimensions are indicated to overcome this drawback (Benzécri, 1973; Greenacre, 2017). Using Greenacre's method (more parsimonious), we obtain a variance percentage of 74.65% for the first dimension and of 3.97% for the second (overall, therefore, the first factorial plan would explain the variance of 78.62%, which seems satisfactory; R software and in particular the ca package was used for this re-evaluation, as in Nenadic, Greenacre, 2007).

It can be observed that the intensity of weakness may be used as a good synthesis of the ten variables, as it displays quite homogeneous levels, growing from left to right of the graph and substantially following the first axis of the map. This is congruent with the presence of almost all "_no" (absence of weakness) statements for almost all variables on the left of the map, while the "_yes" (presence of weakness) statements are on the right.

It should also be noted that negative responses to the fact that school is useful for one's life and for personal growth, corresponding to a state of weakness visible in the map at the top right in an eccentric position (Life_yes and Growth_yes), are generally less recurrent than other types of weakness (4.40% for Life_yes and 13.43% for Growth_yes).

4. Living conditions that can facilitate weakness

Considering Figure 2 again, the most important reading direction is - as already shown - from the far left (absence of weakness) to the far right (greater weakness). However, a second useful piece of information can be found using the bottom-up dimension, especially in the quadrants to the right of the graph where the points are vertically more dispersed and can provide information on the type of discomfort experienced. In fact, staying on the right, towards the top there are the already known negative answers on the importance of school for one's life and growth and also situations of prolonged absence (Absences_yes): this could suggest a difficulty deriving from lack of motivation and escape. On the downside, however, there is a lack of support and attention from adults (Confide_yes; Support_yes), as well as isolation at school and outside peer groups (Isolated_yes, Outside_yes).

Figure 3 – Supplementary variables in the same plane as Figure 2.
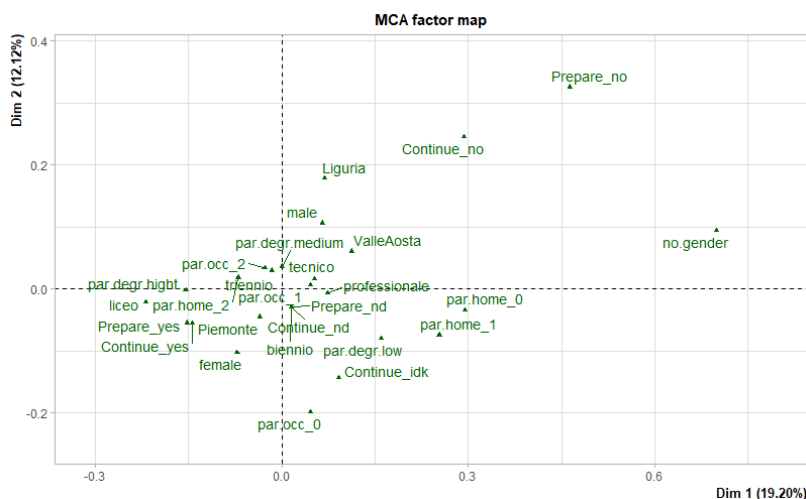
Figure 3 also shows, on the same factorial plane as Figure 2, some characteristics of the surveyed students.

On the left side of the map, where difficulty is generally low, we note the simultaneous greater

presence of both parents in the family unit of cohabitation (par. home_2), a high parental qualification (at least for one of the two parents; par. degr. high), "liceo" enrolment; for students of the second three-year period (who were asked more questions) also the conviction that school prepares for the future and they will to continue studying.

Compared to "liceo" students, those from other paths ("professionale", "tecnico") generally show more weakness. Females appear to be in a better position than males. Piedmontese schools seem to be in a better position than those of Valle d'Aosta or Liguria.

Looking at the top right of the map, students of the second three-year period, who do not think school prepares for the future (Prepare_no) and do not want to continue studying (Continue_no) are in the same position as those already discussed with lack of motivation and orientation to escape in Figure 2.

On the bottom right, which is the already discussed area with more students unsupported by adults and isolated by peers, Figure 3 adds the information concerning greater presence in this area of parents with low qualification (par. degr. low) and difficulty of presence of both parents in the family unit of cohabitation (par. home_0 or par. home_1).

5. Conclusions

In this study, which considers secondary education institutions from the inside by a survey conducted on students, a transversal look at the various sections of the survey questionnaire was maintained to highlight one of the key aspects of the research, aimed at observing situations of well-being, but also of progressive weakness of the students themselves.

The focus is on observing in the sense of describing and therefore favouring a reflection on causes determining weakness: a method to place surveyed students in a scale polarized between well-being on the one hand and extreme discomfort on the other was proposed; subsequently it was shown how some characteristics (to be explored as possible motivations), derived above all from the school and parenting context, can accompany the different situations of intensity of individual weakness.

Furthermore, in the last three-year school period, attention was devoted to how one's "feeling" within the school context can influence the desire for a future perspective for oneself and how it can influence choices.

References

Abdi, H.; Valentin, D. (2007). *Multiple correspondence analysis*, ed. N. Salkind, Encyclopedia of Measurement and Statistics, Sage Publications, Thousand Oaks, (CA), **95**, pp. 116-128.

AA.VV. (2020). *From Neet to Need. Il cortocircuito sociale dei giovani che non studiano e non lavorano*, eds. G. Lazzarini, L. Bollani, F. S. Rota, M. Santagati. Innovation Creativity Settings, Franco Angeli, Torino.

Benzécri, J.P. (1973). *L'Analyse des Données*. Dunod, Paris.

Bollani L., Rota F.S. (2018). Orientamenti per una comprensione ecosistemica dei NEET e conseguenti politiche di sostegno, *EyesReg*, **8**,2.

Escofier, B., Pagès, J. (2008). *Analyses Factorielles Simples et Multiples: Objectifs, Méthodes et Interprétation*; Dunod, Paris.

Greenacre, M. (2017). *Correspondence Analysis in Practice*; Chapman and Hall, New York.

Merril B., West L. (2012). *Metodi biografici per la ricerca sociale*, Apogeo.

Nenadic, O., Greenacre, M. (2007). Correspondence analysis in R, with two-and three-dimensional graphics: The ca package. *Journal of Statistical Software,* **20**(3), pp. 1-13.

Emergency remote teaching: an explorative tool

Emma Zavarrone, Maria Gabriella Grassia, Rocco Mazza, Alessia Forciniti

1. Introduction

The worldwide rapid spread and severity of the infectious disease caused by Coronavirus forced the WHO to declare a global state of pandemic emergency during March 2020, by leading the governments around the world to adopt policies that created the widest rift of education systems in human history. As 85% of worldwide countries, also Italy has temporarily closed each educational institution, by causing the disruption of tertiary education for 16.89% of the Italian learner's population.

To ensure the "pedagogic continuity", universities adopted the transitioning from traditional face-to-face to online learning (e.g., Tallent-Runnels *et al*., 2006; Sangrà *et al*., 2012; Todri *et al.*, 2021). In particular, the shift to fully remote teaching solutions as response to crisis is called by Hodges *et al.* (2020) as *emergency remote teaching (ERT)*. This paradigm shift created changes about the perception of the learning process (Lederman, 2020) and supposed significant didactic efforts in terms of digitalisation and interactive pedagogical approaches. The ERT main goal is not to re-design a long-term educational ecosystem, but to supply a rapid and temporary solution to a crisis condition (Appolloni *et al*., 2021) adopting a learning framework different from online ones.
This implicates venturing into uncharted territory with several logistical challenges and attitudinal modifications (Ribeiro, 2020), also in terms of teaching-learning assessment.

Thus, the evaluation of ERT on the quality of higher education becomes a sensitive issue. This paper raises the evaluation of the effectiveness of teaching delivery during the transition from a traditional model to the ERT one. The focus is to detect how ERT is perceived and how it can connect to students' performance and to the quality of education. The aspect of quality has been dealt out in terms of European Standards and Guidelines (ESG) adopted in 2005 by the Ministers of Higher Education of the countries participating in the Bologna Process (1999; Grano and Ricci, 2009). To ensure the quality of the tertiary education system, European countries have established monitoring agencies. In Italy, this agency is ANVUR, which received official accreditation by the European association for quality assurance in higher education (ENQA) in 2019. However, during the Coronavirus health emergency, ANVUR did not provide guidance to universities on how to manage distance learning and its evaluations, relying on the autonomy of universities which continued to adopt the traditional evaluation systems. In a higher education landscape dominated by quality assurance view for evaluating teaching quality and student satisfaction (Fabbris, 2007), the teaching-learning ERT solutions adopted could not be fit since it is affected by several new factors and the principles recalled by the Bologna Process may be not appropriate.

Thus, the paper focuses on an alternative simple tool for evaluating the quality of teaching-learning in ERT cases. Our research question has an explorative nature: we are interested in detecting empirical evidence about the learning assessment and engagement in higher education with focus on students' engagement and their success performance during ERT.
These dimensions have been represented in the ERT map inspired by perceptual maps of the consumer' theory (Whitlark & Smith, 2001; Gower *et al*., 2010). In our model, the ERT map has been realised by a data integration perspective which considers the university administrative and textual data in a multivariate scenario of methodologies. Textual information is represented by the student voice, since this provides essential information for Quality Assurance systems and for

monitoring and managing information for universities processes. It represents one of the central issues in the most recent version of the European Guidelines and Standards for Quality Assurance in the European Higher Education Area adopted at the Yerevan meeting in 2015 and in AVA 2017 (where AVA stands for Self-assessment, Evaluation, Accreditation; ANVUR, 2021).

In the following, section 2 introduces the theoretical framework; section 3 describes our model and data; section 4 shows the main findings; section 5 presents the future directions.

2. Theoretical framework

During the last two years, the ERT condition caused by the *SARS-CoV-2* pandemic allowed to enrich the literature by several contributions aimed to propose methods and techniques to evaluate the aspects of online teaching-learning in higher education.

Among approaches to evaluate engagement and performance, Bawa (2020) examined the effects of the pandemic related ERT on learners' grades, using an experimental design to investigate the shift of online learning. The analysis has been realised by comparing the same course content and assessment methods for an experimental group formed of students enrolled during 2019-2020 and a control group with students who attended college before the health crisis. The results showed better outcomes in the experimental group than in the control group, above all for highest range performance. Dost *et al.* (2020) investigated attendance and perceptions of medical students across 40 UK schools during May 2020. By means of a cross-sectional study conducted on a national level via an online survey on a 20-items questionnaire measured on Likert scales, the study examined the experiences of online teaching, perceived benefits and barriers and the reached outcomes. From the findings, it emerges that online teaching platforms allow students to digest information in their own time and at the same time to discuss them with peers and, showed to be effective in terms of achieving learning outcomes. Huang *et al.* (2020) analysed the students' engagement by adopting a mixed-methods design: from a quantitative descriptive approach to a qualitative visual method.
At the end of course, all students were given a format inspired by the Motivated Strategies of Learning Questionnaire evaluated on Likert scales about four components: task value, metacognitive self- regulation, effort regulation, and peer learning. In addition, the demographic information and the answers to open- ended questions were coded and grouped into themes by qualitative approach. The results demonstrated the engagement does not depend on learning experience but on extrinsic goal orientation.

Therefore, in a landscape of emerging difficulties to evaluate the student's performance and engagement in ERT contexts as the Coronavirus one, our work proposes a strategy of quantitative analysis aimed to show empirical evidence about the learning assessment and engagement in higher education.

3. Model and data

To assess the quality and ERT success our proposal is based on study of two dimensions: the students' engagement (SE) and success performance (SP) that represent the proxy variables used for constructing our model of analysis. SE has a textual nature comes from students' voice whilst SP uses the career data of students. We obtained our model (*Fig.*1) by integrating these sources of information: textual ones linked to the analysis of the strengths of the engagement and administrative ones related to the results of students' performance. We operated a multidimensional analysis to study the data from these different sources.
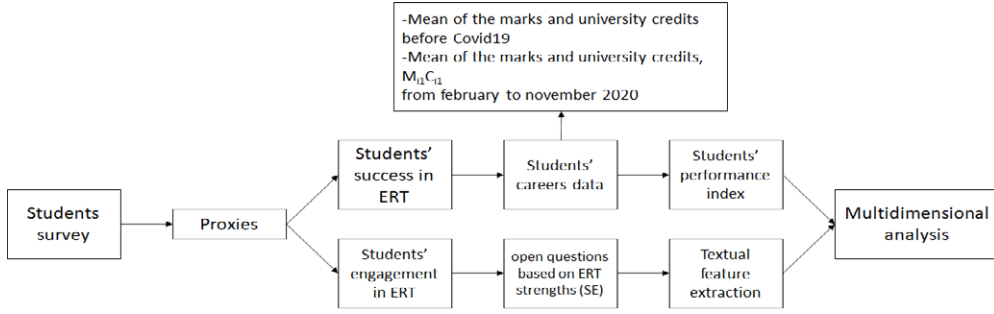
Fig.1: Model flowchart

The measurement of the SE was focused on answers referring to the strengths of the engagement which have been analysed by means of the textual approach. We applied bags of words scheme to transform unstructured texts data in a structured data matrix to analyse. We operated the following pre-treatment operations: 1. Texts normalization; 2. Lemmatization; 3. Stopwords deleting. We built a Document-Term Matrix (DTM) without low frequency words (frequency cut at 5 minimum term frequency and 2 at document frequency of the feature) and empty documents. The DTM has texts for each row (755) and words in columns (790).

The SP measurement is connected to the quantification of the students' results and for reaching this goal the administrative data were used. To construct the Success Performance Indicator (SPI), we considered the average of marks (*M*) and ECTS (European Credit Transfer and Accumulation System) or credits (*C*) achieved before and during the pandemic. More precisely, for measuring each variation of marks and ECTS we used (1):

$$\Delta M_i C_i = \frac{M_{i_1} C_{i_1} - M_{i_0} C_{i_0}}{M_{i_0} C_{i_0}} \qquad (1)$$

where M_{i0} and C_{i0} denote the average of marks and ECTS obtained until February 2020 for each *i-th* student, where $i=1, ... N$; M_{i1} and C_{i1} are the average of marks and ECTS achieved from February to November 2020.

Therefore, the SPI$_i$ was computed by (2):

$$SPI_i = \frac{\Delta M_i C_i}{\max(M, C)} \, 100 \qquad (2)$$

It considers each variation of average of marks and ECTS in relation to the maximum average of marks and ECTS (maximum of mark and ECTS number, 30x180=5400) reachable by 100.

To simplify the next steps, we recoded the SPI considering the quartile (Q) of SPI for each *i-th* student:

$$SPI_i = \begin{cases} 1 - \text{low if SPI}_i \leq Q_{1[SPI_i]} \\ 2 - \text{medium if } Q_{1[SPI_i]} < \text{SPI}_i \leq Q_{2[SPI_i]} \\ 3 - \text{high if SPI}_i > Q_{3[SPI_i]} \end{cases}$$

The SPI can be interpreted as: low performance when SPI$_i$ is lower or equal to the first quartile (SPI$_i \leq Q_1$); medium performance if SPI$_i$ is between the first and the second quartile ($Q_1 <$ SPI$_i \leq Q_2$); high performance in the case of SPI$_i$ is greater to the third quartile (SPI$_i > Q_3$).

Using the SPI$_i$ recoded in terms of quartile and the lexical dimension of SE, we created a contingency table to cross the proxy variables of our model. We inserted into this matrix the performance information at the student level in the rows and lexical keyword extracted from

DTM in the columns. This was possible because the DTM matrix (from SE study) and SP measurement matrix had the same rows. With a multidimensional data analysis approach, we operated a dimensional reduction through Correspondence Analysis (CA). A factorial plan or perceptual map, where we plotted our features, has been created. With this strategy, we can study the association between the two dimensions taken into account in our model. The intersection of these two proxy variables on the factorial plan shows on the horizontal axis SE while on the vertical axis SP. According to our model three principal theoretical areas can be imagined:

- low SE and SP identify the quality of ERT at first level denoting a situation of attention to the low level of engagement and performance;
- medium SE and SP denote the quality of ERT at second level denoting a situation of equilibrium between engagement and performance
- finally, high SE and SP represent the quality of ERT at third level denoting a situation of high engagement and performance.

The population is composed by the students enrolled in a three-year degree course at Iulm University of Milan (N=5000) during the academic year 2019-2020. The survey on ERT weaknesses and strengths had a response rate equal to 14% of the population. The investigated variables are related to the students' career: year of the course; type of high school; gender; average of the marks and ECTS obtained until February 2020, before the Coronavirus, and from February to November 2020, during ERT.

The female students are overrepresented. They respectively represent the 82.3% of the whole student population, the overall students enrolled in the first year are 44.3%. The Iulm University is composed by three Faculties: the 68% of the respondents studied at the Faculty of Communication and other respondents are equally split in other faculties.

4. Results

The recoded SPI barplot is symmetric and as shown in *Fig.2 (a)*.
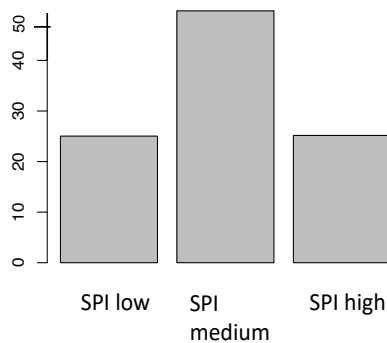
(a)   *(b)*

For the SE, 755 texts were parsed and tokenized from the corpus. The results are a set of strings containing words used in documents. Subsequently, we reduced language variability to avoid possible sources of noise and to improve the effectiveness of the next analytical steps. The step consisted in the normalization of words, spelling and brought back each inflected word in its canonical form. Finally, we pruned non-informative words and non-alphabetic characters from the texts. The vocabulary size consisted of 790 types. Subsequently, we reduced the dimensionality of

this matrix, we filtered sparse words (with a sparsity threshold of 2%). At the end of the process, each document was represented as a document-vector and the number of types was 190. The comparison wordcloud plot shows the most frequent terms for separate grouping level of SPI (*Fig. 2 (b)*). The wordcloud allows to capture some differences among the words related to the different level of SPI: terms like "riascoltare" "listen to the lesson again" characterized the low level of SPI, whilst terms with negative meaning in the SPI were in the high-level area.

As we affirm in Par.3, a contingency table was created, and we obtained the factorial plan through the CA (*Fig.3*) where the first dimension explains 54.4% and represents the ERT success while the second one explains 45.6% and denotes the SE. We can see three sections characterized by SPI levels splitting the map in the three horizontal levels: low, middle and high. On the contrary, the SE can be read easily from right to left, where we find an individual student engagement and the collective students' engagement respectively. At first glance, we discover that the individual engagement is at the high level of SPI. This puts the semantic dimension related to the individual experience close to the high and low performance factors, as far as an exploratory analysis is concerned. We want to highlight that the two polarities (high and low) referring to the performance are both in the same half plane. Obviously, both refer to two different ways of experiencing distance for students. The difference consists, for the high SPI, in the facilitated access to the technologies made available and the possibility of optimizing the time available for the study. The low performance is close to the properly teaching dimension and relative to the contents of the courses.
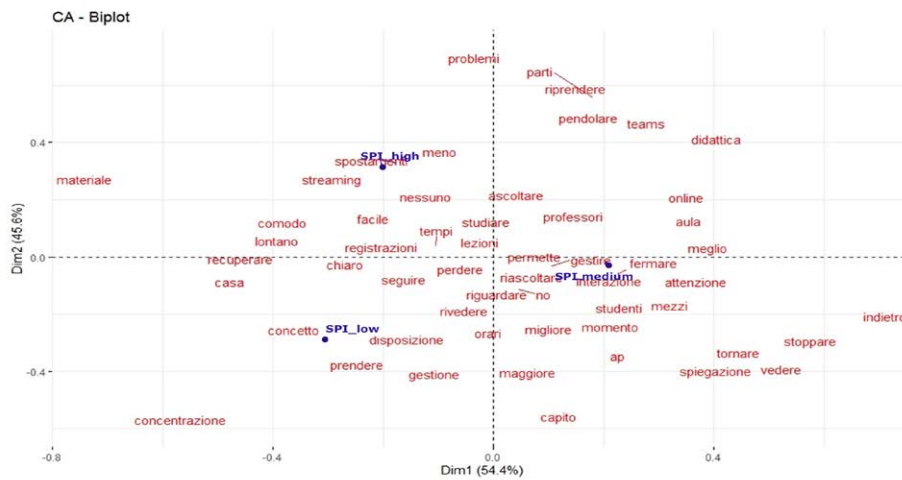
Fig.3: Factorial plan

5. Future directions

The work proposes the data integration approach to create the ERT map. The CA allows to explore this integration using the contingency table built on SPI and SE, proxies developed to detect the success of the performance and the level of student engagement respectively. Attention should be paid to the use of short texts, inspired by the dialogue on social media, which do not always allow- based on the Italian language- to extract the true underlying concept. For this reason, future developments are moving in two directions: creation of an Italian dictionary, specifically for evaluating ERT, and creation of indicators that can be used in an agile way for subsequent ERT evaluations. The indicators could also be useful for drop out screening and prevention by monitoring the level of collaborative engagement.

References

ANVUR (2021). Rilevazione Opinioni Studenti. https://www.anvur.it/attivita/ava/opinioni-studenti/. Last access: 12/07/2021.

Appolloni, A., Colasanti, N., Fantauzzi, C., Fiorani, G., Frondizi, R. (2021). Distance Learning as a Resilience Strategy during Covid-19: An Analysis of the Italian Context. *Sustainability* 2021, **13** (1388).

Bawa, P. (2020). Learning in the age of SARS-COV-2: A quantitative study of learners' performance in the age of emergency remote teaching. *Computers and Education Open.*

Cimea (2004). Il Processo di Bologna, Documenti Ufficiali. http://www.cimea.it/files/fileusers/7384_2004-Il%20Processo%20di%20Bologna.%20Documenti%20ufficiali.pdf. Last access: 15/04/2021.

Dost, S., Hossain, A., Shehab, M. *et al.* (2020). Perceptions of medical students towards online teaching during the COVID-19 pandemic: a national cross-sectional survey of 2721 UK medical students. *BMJ Open* 2020.

Fabbris, L. (2007*). Effectiveness of University Education in Italy. Employability, Competences, Human Capital.* Physica-Verlag Heidelberg.

Gower, J.C., Groenen, P.J.F., van de Velden, M., Vines, K. (2010). *Perceptual maps: the good, the bad and the ugly (ERS-2010-011-MKT). ERIM report series research in management Erasmus Research Institute of Management.* Erasmus Research Institute of Management.

Grano, C., Ricci, A. (2009). Il processo di Bologna. Lo Spazio Europeo dell'Istruzione Superiore. http://www.processodibologna.it/documenti/Doc/Pubblicazioni/Lo%20Spazio%20Europeo%20dellIstruzione%20Superiore. Last access: 15/04/2021.

Hodges, C., Moore, S., Lockee, B., Trust, T., Bond, A. (2020). The difference between emergency remote teaching and online learning. Educause Review. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning. Last access: 10/04/2021.

Huang, M., Shi, Y., Yang, X. (2020). Emergency remote teaching of English as a foreign language during COVID-19: Perspectives from a university in China. *International Journal of Educational Research and Innovation (IJERI),* **15**, pp. 400-418.

Lederman, D. (2020). Will Shift to Remote Teaching Be Boon or Bane for Online Learning?. *Inside Higher Ed.*

Ribeiro, R. (2020). How university faculty embraced the remote learning shift. *EdTech Magazine.*

Sangrà, A. Vlachopoulos, D., Cabrera, N. (2012). Building an inclusive definition for e-learning: an approach to its conceptual framework. *The International Review of research in Open and Distance Learning,* **13**(2), pp. 145-159.

Tallent-Runnels, M. K., Thomas, J. A., Lan, W. Y., Cooper, S., Ahern, T. C., Shaw, S. M., Liu, X. (2006). Teaching courses online: A review of the research. *Review of Educational Research,* **76**(1), pp. 93-135.

Todri, A., Papajorgji, P., Moskowitz, H., Scalera, F. (2021). Perceptions regarding Distance Learning in Higher Education, Smoothing the Transition. *Contemporary Education Technology,* **13**(1).

Whitlark, D. B., Smith, S. M. (2001). Using correspondence analysis to map relationships. *Marketing Research,* **13**(3), pp. 22-27.

Effects of an experimental online education support on lectures fruition and teaching effectiveness

Maria Cristiana Martini, Marco Furini, Giovanna Galli

1. Introduction

The fruition of university courses has significantly changed in the last decade, in consequence of the higher accessibility of technological devices: universities, as well as for-profit companies, started to propose video lectures, as a substitution or in support of traditional lessons, and massive online open courses (MOOCs) have gained more and more importance in the education processes. This tendency comes as an answer to a growing need for flexibility expressed by working students, life-long learning processes, students who have families and care burdens, those with some forms of disability or special needs that make it difficult to attend classes.

Many authors have investigated the effectiveness of video lectures, primarily in comparison with face-to-face classes, with mixing results: some found no significant differences between online and face-to-face courses (Lim et al., 2007; Neuhauser, 2002; Nemetz et al., 2017), while others suggested higher outcomes in online courses (Soffer and Nachmias, 2018; Burkhardt et al., 2008; Connolly et al., 2007; Lim et al., 2008). The Covid-19 pandemic has magnified and accelerated the surge of online teaching, in a way that makes the change hardly reversible. The ongoing debate on the effectiveness of video lectures in higher education is meant to last and intensify.

In this paper, we describe and discuss the implementation, the acceptability, and the effectiveness of an experimental service designed to capture, record, edit and stream video lectures; this system was introduced with the principal aim of supporting, and not substituting, in-class learning. In detail, Section 2 illustrates the experimental service and the main usage behaviours; Section3 presents the main results in terms of effectiveness and usage models, while some conclusions are drawn in Section 4.

2. ONELab: an experimental education support

ONELab is a system designed to capture, record, edit and stream video lectures, introduces by the Department of Communication and Economics of the University of Modena and Reggio Emilia in September 2017. Traditional face-to-face classes were regularly held, but ONELab was intended to ease the educational experience of those students who cannot attend classes regularly, and to provide an additional support to traditional students. Each classroom is equipped with a video camera pointed on the teacher's desk, an audio system to capture and amplify the teacher's voice, a screen to display the slideshow, and a live video production system to capture, mix, record and stream the video signals (i.e. teacher's video and slideshow) and the audio. After a minimal post-processing, the video lectures are loaded to the online platform and made available for students (see Furini et al., 2018; 2020 for more details).

In the first year of experimentation, from September 2017 to June 2018, 1,376 video lectures were produced, covering the 49 courses offered in the first year of the five bachelor's and master's degrees supplied by the Department, for a total of 2,064 hours. In the academic year 2018/19 these numbers doubled, and further increased in 2019/20, as the courses offered in the second and third year joined the experimentation.

Maria Cristiana Martini, University of Modena and Reggio Emilia, Italy, cmartini@unimore.it
Marco Furini, University of Modena and Reggio Emilia, Italy, marco.furini@unimore.it, 0000-0003-1094-6521
Giovanna Galli, University of Modena and Reggio Emilia, Italy, giovanna.galli@unimore.it

The analysis of the first year log files shows that students' reaction was enthusiastic, with an average of 8,323 video lectures played each month in the first year, from September 2017 to August 2018, and a peak of 14,483 views in January 2018, during the exams session. Overall, during the year students watched video lectures for 71,488 hours. The most watched lectures relate to technical subjects, such as mathematics and statistics, where students take advantage from the possibility to replay some difficult passages until they are clear.

Video lectures are watched mostly during the teaching semester, but a significant part of students resort to watch them when the semester is over, especially during the exams sessions. The usage analyses show that students watch video lessons mostly during the working hours, from Monday to Friday; however, 16% of the views happen during the week-ends, and 22% in the evening and during the night, suggesting that, when given the opportunity, students tend to customise the learning process to their needs and life-style.

Of the 1251 freshmen in the academic year 2017/18, only 319 (25.4%) never accessed the ONELab platform to watch video lessons during their first year, while 13.4% never accessed it, neither in the first or in the second year. Table 1 shows the percentage of non-users among different categories of students, separately for undergraduate and graduate students.

Table 1. Percentage of students who never accessed the ONELab video lectures during the first year and in the first two years, per students' characteristics.

	Graduate		Undergraduate		Total	
	First year	Two years	First year	Two years	First year	Two years
Males	8.4	1.1	30.4***	23.8***	25.5***	18.7***
Females	3.8	2.2	16.1***	13.1***	12.3***	9.7***
SLD	0.0	0.0	0.0***	0.0***	0.0***	0.0***
Non-SLD	5.4	1.8	22.9***	18.2***	18.8***	13.7***
Italian	4.9°	1.1**	22.4***	18.0**	17.5***	13.2***
EU	0.0	0.0**	0.0***	0.0**	0.0***	0.0***
Non-EU	22.2°	22.2**	27.3***	20.0**	25.6***	20.3***
Lyceum	5.5	1.2	18.6°	14.4°	13.8**	9.6**
Technical college	5.3	2.6	27.1°	21.7°	22.6**	17.8**
Vocational college	4.1	0.0	23.0	20.0	19.4	16.1
Other school	7.1	7.1	17.9	12.5	15.7	11.4
Dropouts	18.5°	14.8*	45.3***	44.6***	42.3***	41.2***
Non-dropouts	4.0°	0.4*	13.4***	6.9***	10.4***	4.8***
Total	5.4	1.8	22.6	17.9	17.9	13.4

Significance level: *** 99.9%, ** 99%; * 95%; ° 90%.

The use of ONELab is particularly popular among graduate students, while almost one undergraduate student out of four never watched any video lecture. Females are more conscientious than males, and look for every provided support to enhance their preparation, but the difference is statistically significant only among undergraduates. Students affected by Specific Learning Disorders are only 12, but none of them missed the new learning support, that allows for a certain degree of self-paced study, ensuring more control over their learning. On the other hand, undergraduate students coming from technical and vocational colleges are less organised in their study, and overlook video lectures to a greater extent, while non-EU foreigners miss this support both as undergraduate and graduate students.

The recourse to video lectures is extremely scarce among students who end up dropping out the university in the first year. To some extent, these students might have dropped out because they did not take advantage of ONELab to support their studies, but it might also be that some students decided to leave the university, or to transfer to a different degree, so early

that they did not have time to try the video lectures.

3. Effectiveness of the video lectures

The high percentage of usage is a first indirect indicator of effectiveness of the video lectures, but we aim at assessing the benefits of the video lectures in terms of learning outcomes, namely the number of acquired (European) credits and the final grades. We focus on students enrolled in 2017, and analyse data on their ONELab accesses and academic achievements during the academic years 2017/18 and 2018/19. We do not consider the third year, neither for three-year courses, because it corresponds to the Covid-19 pandemic outbreak, when face-to-face classes were totally replaced by video lectures in the second semester, which makes the situation incomparable.

Since early dropouts produce a low number of credits (or no credits at all) and a low level of access to the online platform, we remove them from the following analyses to exclude the existence of spurious relationships.

Table 2 shows that the average number of credits acquired by students who watched video lectures is largely bigger than for those who did not, both in the first and in the second year, and the difference is statistically significant. Students who accessed ONELab also performed better in terms of grades: the average grade of accessing students is higher than for non-accessing students, although the difference is only marginally significant in the second year. However, separate analyses carried out on undergraduate and master students demonstrate that students who accessed the video lectures show significantly better performances only in terms of acquired credits.

Table 2. Average number of acquired credits during the first year and average grade for ONELab users and non-users, per degree level.

		n	First year		Second year	
			Credits	Average grade	Credits	Average grade
Undergraduate students	ONELab users	590	33.8	23.9	36.9	23.5
	No ONELab users	137	17.5	23.4	17.0	23.4
	Total	*727*	*30.9*	*23.8*	*33.4*	*23.5*
	T test		*8.75*	*1.56*	*7.51*	*0.20*
	(p-value)		*(p < 0.001)*	*(p = 0.120)*	*(p < 0.001)*	*(p = 0.841)*
Graduate students	ONELab users	259	38.0	25.8	38.2	26.3
	No ONELab users	17	22.6	26.3	26.5	25.6
	Total	*276*	*37.3*	*25.8*	*37.6*	*26.3*
	T test		*3.24*	*-0.68*	*2.07*	*0.99*
	(p-value)		*(p = 0.001)*	*(p = 0.495)*	*(p = 0.039)*	*(p = 0.323)*
All students	ONELab users	849	35.2	24.5	37.3	24.5
	No ONELab users	154	18.0	23.0	18.0	23.8
	Total	*1003*	*32.8*	*24.5*	*34.6*	*24.4*
	T test		*10.08*	*2.66*	*8.38*	*1.957*
	(p-value)		*(p < 0.001)*	*(p = 0.008)*	*(p < 0.001)*	*(p = 0.051)*

The rough distinction between students who never watched video lectures and those who accessed the platform at least one time, although simplistic, has proven to be meaningful in explaining performance differences among students. We try to describe in more detail the different usage styles of those who accessed the platform at least one time in the two years through the following variables, separately measured on the first and the second year:

- *Total number of accesses*: this variable measures the general degree of usage during the first year. It varies between 0 and 864 for the first year, and between 0 and 1,885 in the

second year; the average is respectively 75.3 and 112.7;

- *Total number of different courses accessed*: this provides information on the students' choice of using the platform for one, some, or all of the scheduled courses. It varies between 0 and 25 in the first year, and between 0 and 28 in the second, and the average is 4.8 and 4.7 in the two occasions;
- *Number of courses accessed at least 5 times:* the wide range of observed values registered for the previous index suggests that some students have browsed in the platform clicking on more courses than the ones planned for their study course; this variable filters the courses that are only accessed on a random browse, and measures the number of courses accessed at least 5 times; both in the first and in the second year it varies between 0 and 13, and the average is 3;
- *Number of courses accessed at least 10 times:* this variable measures the number of courses accessed at least 10 times, signalling a larger commitment to the course; it varies between 0 and 9 in the first year, and between 0 and 11 in the second, and the average is 2.1 and 2.2;
- *Maximum number of accesses to a single course*: this index shows how many times each student played the video-lectures of the course he accessed most in the year; for the first year it ranges from 0 to 253, and the average is 27.3, while for the second year it ranges between 0 and 698 and the average is 44.1.

Some of these variables show unexpectedly high values (for example, students registering 1,885 total accesses, or students who played 698 times the video lectures of a single course), and the reason is twofold. First, every single access does not correspond to a complete play of the video-lecture; as reported by many students, "critical" passages, especially on some technical topics, have been repeatedly reloaded and re-played, and sometimes a lecture is erroneously played while looking for another one, or for a different part of the same recording. In addition, the platform was a novelty that probably raised curiosity among students, leading some to explore the resources far beyond the actual usage.

Based on the ten described variables, we perform an agglomerative hierarchical cluster analysis; the agglomeration criterion is the Ward's method that, at each step, merges the couple of units/clusters that leads to minimum increase in total within-cluster variance. The distance is the squared Euclidean. Given the different order of magnitude, all variables have been rescaled to the [0-1] range using min-max normalization.

This cluster analysis suggests the existence of four distinct groups; combining these clusters with the group of absolute non-users, we obtain the five profiles described in Table 3 (first year dropouts are excluded from the analysis):

1. Absolute non-users: They are 9% of all students; they never accessed the ONELab platform in the two academic years.
2. Episodic users: This group amounts to the 32.9% of all students; on average, they tried to play a few video-lectures from about 3 courses in the first year and only a couple on the second, played a single course about 5-10 times and almost never accessed more than 10 times to a single course.
3. Regular users: They represent 25% of students; on average, they accessed most of the courses planned in their study program, but chose 2-3 of them which were played more frequently, up to 40 times. In the second year their usage intensity declines, and they play less videos, from less courses, a smaller number of times; the experience during the first year helps them distinguish which courses are worth watching and re-watching and which are not.
4. Converted users: They are 23.6% of all students; during the first year, they show a scarce recourse to video lectures, larger than episodic users but far from regular users. Nevertheless, during the second year their usage intensity grows and exceeds regular

users. They are probably students who discovered the video lectures only late in the first year, or approached them first without enthusiasm but found out they were useful than expected for their preparation.

5. Zealous users: They amount to 9.5% of students, and they accessed the platform hundreds of times; they accessed more or less all of the courses provided in their study program, and they were assiduous on most of them, playing video-lectures from each single course up to 70 times in the first year, and even up to 120 in the second year.

Table 3. Average number of accesses, average number of courses accessed, average number of courses accessed at least 5 times, average number of courses accessed at least 10 times, and maximum number of accesses to a single course, for each group

	Accesses		Courses accessed		Courses accessed 5 times or more		Courses accessed 10 times or more		Maximum n° of accesses to a single course	
	year1	year2	year1	year2	year1	year2	year1	year2	year1	year2
Absolute non-users	0	0	0	0	0	0	0	0	0	0
Episodic users	20.7	16.3	3.1	1.9	1.2	0.8	0.6	0.4	11.2	10.5
Regular users	135.3	93.2	7.6	5.8	5.8	3.5	4.1	2.4	46.3	42.9
Converted users	47.6	190.3	4.9	7.4	2.6	4.7	1.4	3.6	22.4	77.9
Zealous users	260.9	413.9	8.3	9.8	6.9	7.8	5.9	6.7	74.0	121.7
Total	*76.6*	*112.7*	*4.9*	*4.7*	*3.1*	*3.0*	*2.1*	*2.2*	*27.5*	*44.1*

For each group of students, the learning performances are reported in Table 4. The level of performance increases with the frequency of usage of the ONELab services. Regarding the number of credits, all the group means are statistically different at least at a 95% significance level, except for Converted users, that are not significantly different from Regular users in the first year, and from Zealous users in the second year. The average grade shows only slight differences, nevertheless consistent with a better performance for regular and zealous users. Differences between graduate and undergraduate students are not noticeable.

Table 4. Average number of acquired credits and average grade, in the first and second year, for each group, per degree level.

	First year		Second year	
	Credits	Average grade	Credits	Average grade
Non-users	12.1	23.4	9.6	23.4
Episodic users	25.6	23.8	24.0	23.8
Regular users	39.7	25.1	36.7	25.3
Converted users	38.9	24.5	49.2	24.2
Zealous users	44.2	25.2	53.8	24.6
Total	*32.8*	*24.5*	*34.6*	*24.4*

4. Conclusions

In this paper, we analysed the effectiveness of an experimental platform to provide university students with remote access to video lectures to support traditional face-to-face classes. Results show higher learning outcomes for students who regularly watched the video lectures, primarily

in terms of the number of acquired credits. This is consistent with the conclusions drawn in Cagliero et al. (2017), who report higher student's success rates following the introduction of an analogous system to provide video-recorded lessons to complement in-class learning. In our experience, the beneficial is particularly pronounced for undergraduate students, although they show a more limited recourse to the platform than graduate students.

However, a more careful analysis of the principal beneficiaries of the implemented service casts a shadow on the capacity of the system to smooth learning ability differences and recover those students who have a hard time keeping pace with their studies and exams. Video lectures, in fact, are mainly watched by conscientious students, i.e. females, students coming from "lyceum" high school, and graduate students, who aim at improving their learning through additional educational material, while critical students are those who access the platform less. This suggests that the information about the new service should be conveyed to students in a more careful and focused way, addressing especially to students at risk of being left behind and dropping out. In this sense, given the strong connection between dropouts and video lectures (non) usage, monitoring and analysing the access data might help to detect critical students, and try to prevent them from dropping out.

Finally, a negative consequence of the introduction of this service was a dramatic decrease in the number of students attending classes, much before the university classrooms were emptied by the pandemic crisis. When face-to-face classes will return to normality after more than one year of online teaching, the problem is likely to become even more compelling, forcing teachers and pedagogues to rethink face-to-face classes in a more interactive and engaging format.

References

Burkhardt, J.M., Kinnie, J., Cournoyer, C.M. (2008). Information literacy successes compared: Online vs. face-to-face. *Journal of Library Administration,* **48**(3-4), pp. 379-389.

Cagliero, L., Farinetti, L., Mezzalama, M., Venuto, E., Baralis, E. (2017). Educational video services in universities: a systematic effectiveness analysis, in *2017 IEEE Frontiers in Education Conference (FIE 2017),* Indianapolis, Indiana (USA), October 18-21, 2017, pp. 1-9.

Connolly, T.M., MacArthur, E., Stansfiled, M., McLesslan, E. (2007). A quasi-experimental study of three online learning courses in computing. *Computers & Education,* **54**(4), pp. 345-359.

Furini, M., Galli, G., Martini, M.C. (2018). ONELab: Online education with minimal human supervision, in *ACM Proceedings of the International Conference on Smart Objects and Technologies for Social Good (Goodtechs '18),* November 2018.

Furini, M., Galli, G., Martini, M.C. (2020). An online education system to produce and distribute video lectures. *Mobile Networks and Applications,* **25**, pp. 969-976.

Lim, D.H., Morris, M.L., Kupritz, V.W. (2007). Online vs. blended learning: Differences in instructional outcomes and learner satisfaction. *Journal of Asynchronous Learning Networks,* **11**(2), pp. 27-42.

Lim, J., Kim, M., Chen, S.S., Ryder, C.E. (2008). An empirical investigation of student achievement and satisfaction in different learning environments. *Journal of Instrucional Psychology,* **35**(2), pp. 113-119.

Nemetz, P.L., Eager, W.M., Limpaphayom, W. (2017). Comparative effectiveness and student choice for online and face-to-face classwork. *Journal of Education for Business,* **92**(5), pp. 210-219.

Neuhauser, C. (2002). Learning style and effectiveness of online and face-to-face instruction. *The American Journal of Distance Learning,* **16**(2), pp. 99-113.

Soffer, T., Nachmias, R. (1918). Effectiveness of learning in online academic courses compared with face-to-face courses in higher education. *Journal of Computer Assisted Learning,* **34**, pp. 534-543.

DECISION MAKING

Measuring the effectiveness of COVID-19 containment policies in Italian regions: are we doing enough?

Demetrio Panarello, Giorgio Tassinari

1. Introduction

The Coronavirus disease 2019 (COVID-19), caused by the SARS-CoV-2 virus, was first identified in Wuhan, China, in December 2019. The disease quickly spread to the rest of the world. The earliest cases of Italian citizens infected by the virus were detected on the 21st of February 2020 (Romagnani et al., 2020). Italy was sent into a severe lockdown on the 10th of March (Verma et al., 2020) and emerged from it on the 4th of May, slowly starting to reopen its economic activities (Buonomo and Della Marca, 2020).

While the lockdown conveyed a message of danger, the reopening might have led citizens to perceive that the threat had come to an end (Reinders Folmer et al., 2020). Indeed, during the lockdown, inhabitants were obliged to confine themselves under severe penalties; after that, the issue was confidently put into citizens' hands, who were now able to choose how much they were willing to cooperate. An effective response to the pandemic relies heavily on citizens' compliance with the restrictive measures put in place to halt its spread (Sobol et al., 2020), ultimately reducing the number of deaths.

With this paper, we aim at giving an insight into how Italian citizens' compliance with the restrictions – measured through longitudinal data on sanctions and movement trends – has affected the number of deaths over time. Moreover, we investigate what would have happened if, in the event of insufficient compliance on the part of citizens, heavier restrictions were put in place. In so doing, we provide an estimate of how many human lives could have been spared as a result of stricter public health regulations.

2. Data and Methods

Our data come from several sources of information. For each considered variable and each of the 107 Italian provinces, we collected 260 daily observations, pertaining to the period running from the 24th of February to the 9th of November 2020.

First, we collected the daily distribution of COVID-19 positive cases, performed swabs, and recorded deaths in the country's 19 regions and 2 Autonomous provinces, provided by the Italian Civil Protection (Dipartimento della Protezione Civile, 2020). To each province, we associated the corresponding regional values. Country-level daily swabs (in thousands), positive cases (in hundreds), and deaths are plotted in Figure 1. The number of swabs, which was remarkably low at the beginning of the pandemic, shows a major increase in the second half of the considered period. At this point, the deaths line starts keeping pace with the swabs one, so that the number of deaths becomes close to 1 per 100 positive cases.

Furtherly, we made use of the Containment and Health Index, developed by the University of Oxford's Blavatnik School of Government (Hale et al., 2020), tracing the government response to the pandemic outbreak over time. It is a composite index made up of 12 country-level indicators on closings of schools and universities, closing of workplaces, cancelling of public events, restrictions on private gatherings, closing of public transport, stay-at-home requirements, restrictions on internal movements, restrictions on international travel, presence of public information campaigns, testing policy, contact tracing, and facial coverings policy.

Moreover, we gathered the number of daily controls and fines imposed on citizens due to

Demetrio Panarello, University of Bologna, Italy, demetrio.panarello@unibo.it, 0000-0003-1667-1936
Giorgio Tassinari, University of Bologna, Italy, giorgio.tassinari@unibo.it, 0000-0002-5161-7989

disrespecting the COVID-19-related restrictive measures, made available by the Italian Ministry of the Interior (Ministero dell'Interno, 2020) at the national level. We can calculate the sanction rate as the ratio between the number of fines and the number of people who were controlled on a given day; the compliance rate is the one's complement to this rate, which represents a proxy of citizens' degree of adhesion and consent to the measures aimed at containing the Coronavirus spread.

Additionally, we employ Google's Community Mobility Reports, capturing movement trends across various locations at the province level (Google LLC, 2020). We include five categories of places: retail stores and recreation sites, grocery stores and pharmacies, parks, transit stations, and workplaces. The data consist in daily per cent variations in the number of visitors compared to a pre-pandemic baseline.

Finally, we include some variables describing the demographic characteristics of the Italian provinces, taken from the Italian National Institute of Statistics (Istat): activity rate, population density, and ratio of over-65s to the total population.
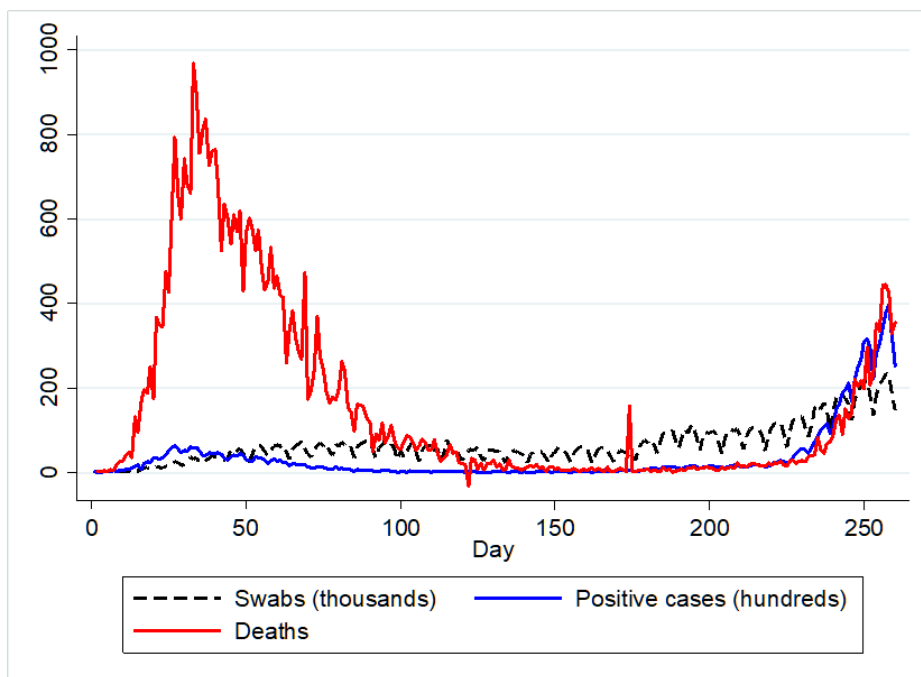
Figure 1 – Swabs, positive cases and deaths over time (Italy, 24 February – 9 November 2020).

We estimate Negative Binomial regressions of the regional deaths count on regional positive cases, regional swabs, Containment and Health Index, Compliance rate, Google Mobility data (for retail and recreation, grocery and pharmacy, parks, transit stations, and workplaces), activity rate, population density, and percentage of over-65s to the total population.

Indeed, as we employ a count variable as dependent, the correct investigation approach is given by regression models based on the Negative Binomial distribution (Chan et al., 2021). The time-varying variables are employed with a 17-day lag from the dependent variable, as we add the median time from the onset of symptoms to death, which was estimated in 12 days in Italy (Gruppo della Sorveglianza COVID-19, 2020), to the mean incubation period (i.e., the time between the contact with a positive individual and the onset of symptoms) of approximately 5 days (Linton et al., 2020). Our specifications employ the robust estimator of variance and do not include fixed effects.

We run the first model on the complete sample. Then, as the schools' reopening on 14

September 2020 is said to have been the primary cause of the resurgence of the pandemic in Italy (Sebastiani and Palù, 2020), we estimate the same model on two subsamples: until the 13th of September and since the 14th of September, which is marked as the beginning of the "second wave" of the pandemic.

Finally, we add up 10 points to the Containment and Health Index daily since the 1st of September, to investigate what would have happened if stricter public health regulations were put in place two weeks before schools restarted. Through this, we provide an estimate of how many human lives could have been spared in the period from 14 September to 30 October 2020 in the Italian regions affected by the highest lethality rate.

3. Results

The results of our estimations are shown in Table 1. As regards the number of deaths, analysed through Negative Binomial regression models, most variables are highly significant and show the expected signs. Our results confirm that the lockdown policies have had a beneficial impact on the pandemic, having been able to reduce the number of deaths caused by COVID-19. Moreover, the number of deaths exhibits a negative relationship with the Compliance rate.

Table 1 – Results from Negative Binomial regressions of regional deaths.

	Overall	Until 13th Sep	Since 14th Sep
	Coefficient (Robust S.E.)	Coefficient (Robust S.E.)	Coefficient (Robust S.E.)
Regional positive cases (lag 17)	0.002^{***}	0.002^{***}	0.001^{***}
	(0.0001)	(0.0001)	(0.0001)
Regional swabs (lag 17)	0.000^{***}	0.000^{***}	0.000^{***}
	(0.0000)	(0.0000)	(0.0000)
Containment and Health Index (lag 17)	-0.022^{***}	-0.011^{***}	-0.122^{***}
	(0.0023)	(0.0027)	(0.0054)
Compliance rate (lag 17)	-0.345^{***}	-0.314^{***}	-1.283^{***}
	(0.0170)	(0.0233)	(0.0710)
Google Mobility: Retail and recreation (lag 17)	-0.028^{***}	-0.028^{***}	-0.011^{***}
	(0.0022)	(0.0028)	(0.0032)
Google Mobility: Grocery and pharmacy (lag 17)	0.023^{***}	0.021^{***}	0.014^{***}
	(0.0010)	(0.0011)	(0.0023)
Google Mobility: Parks (lag 17)	-0.001^{*}	0.002^{***}	-0.003^{***}
	(0.0004)	(0.0005)	(0.0005)
Google Mobility: Transit stations (lag 17)	-0.007^{***}	-0.018^{***}	-0.003^{**}
	(0.0012)	(0.0018)	(0.0011)
Google Mobility: Workplaces (lag 17)	0.007^{***}	0.007^{***}	0.013^{***}
	(0.0012)	(0.0015)	(0.0016)
Activity rate	0.038^{***}	0.070^{***}	-0.012^{***}
	(0.0027)	(0.0030)	(0.0029)
Density (pop. per sq. km)	0.000	0.000	0.000^{***}
	(0.0000)	(0.0000)	(0.0000)
Percentage of over-65s to total population	0.063^{***}	0.093^{***}	0.017^{**}
	(0.0059)	(0.0074)	(0.0068)
Intercept	32.166^{***}	25.240^{***}	136.106^{***}
	(1.6180)	(2.2958)	(7.3005)
Log-transformed over-dispersion parameter (ln α)	0.178^{***}	0.105^{**}	-0.364^{***}
	(0.0290)	(0.0449)	(0.0283)
Observations	21641	16741	4900
McFadden's pseudo R^2	0.163	0.187	0.144
Log-pseudolikelihood	-54050.46	-39330.62	-13235.40

*Notes: *, ** and *** stand for $p < 0.10$, $p < 0.05$ and $p < 0.01$.*

We replicated the analysis by dividing the sample into two subperiods: the first one until the 13th of September and the second one since the 14th of September. The results roughly confirm those from the analysis carried out for the whole period, demonstrating the goodness of the model. Nevertheless, some regressors change their sign from one period to the other: mobility towards parks is positive in the first period, but negative in the second one, and the same goes for activity rate. Moreover, the magnitude of some coefficients changes considerably. In particular, the coefficient for Compliance rate in the second period is over four times that of the first period; additionally, the coefficient for Containment and Health Index shows an increase of about 11 times. This means that the importance of the restrictive measures and of citizens' accord on their abidance has greatly increased since the end of the summer, also because the stringency level of the adopted measures has critically declined, which was preparatory to the formation of the "second wave" of the pandemic. Finally, the share of population aged 65 or more always shows a positive sign, which reflects the known situation of higher lethality characterising the elderly population (Rinaldi and Paradisi, 2020). However, in the second period, its coefficient is about one fifth that of the first period: indeed, this shows that the demographic dynamics of the pandemic have changed compared to the beginning and that the elderly have become more cautious in the second phase of the pandemic.

Trying to sum up our achieved outcomes, the restrictions represented by the Containment and Health Index appear essential to contain the pandemic until the vaccination campaign has produced the so-called herd immunity. However, these restrictions are not sufficient when they are not accompanied by citizens' consent, which translates into adherence to the mobility restrictions, observed through the reduction in Google mobility indices: indeed, it is not realistic to think that repressive actions are enough to enforce compliance with the new mobility rules.

Finally, we add up 10 points to the Containment and Health Index since the 1st of September, providing a prediction of the deaths count from 14 September to 30 October 2020 in the six Italian regions affected by the highest overall lethality rate, in the hypothesis of higher stringency put in place starting from two weeks before the reopening of schools. These simple estimates do not consider the variations in compliance and mobility which could result from a hypothetical change in stringency. The results are summarised in Table 2 and plotted in Figure 2.

Apart from Valle d'Aosta, which experienced a low number of deaths due to its small population size, the predictions show that a significant number of losses could be averted by introducing more restrictions in good time before schools restarted. In particular, Lombardia – the region in which the outbreak started – could have saved 429 lives just between 14 September and 30 October, compared to the 563 deaths faced in the same period (-76.20%).

Table 2 – Count of deaths and cases, population, Case Fatality Rate, Lethality Rate, deaths averted.

Region	Deaths 14th Sep – 30th Oct	Cases 14th Sep – 30th Oct	Population	CFR 14th Sep – 30th Oct	Lethality Rate 24th Feb – 30th Oct	Deaths averted 14th Sep – 30th Oct	Deaths averted (per cent)
Lombardia	563	83486	10103969	0.67%	0.173%	429	76.20%
Valle d'Aosta	20	1876	125501	1.07%	0.132%	-68	-
Liguria	170	15688	1543127	1.08%	0.113%	129	75.88%
Emilia-Romagna	143	20293	4467118	0.70%	0.103%	86	60.14%
Piemonte	202	33996	4341375	0.59%	0.100%	158	78.22%
P.A. Trento	32	3247	542739	0.99%	0.081%	3	9.38%
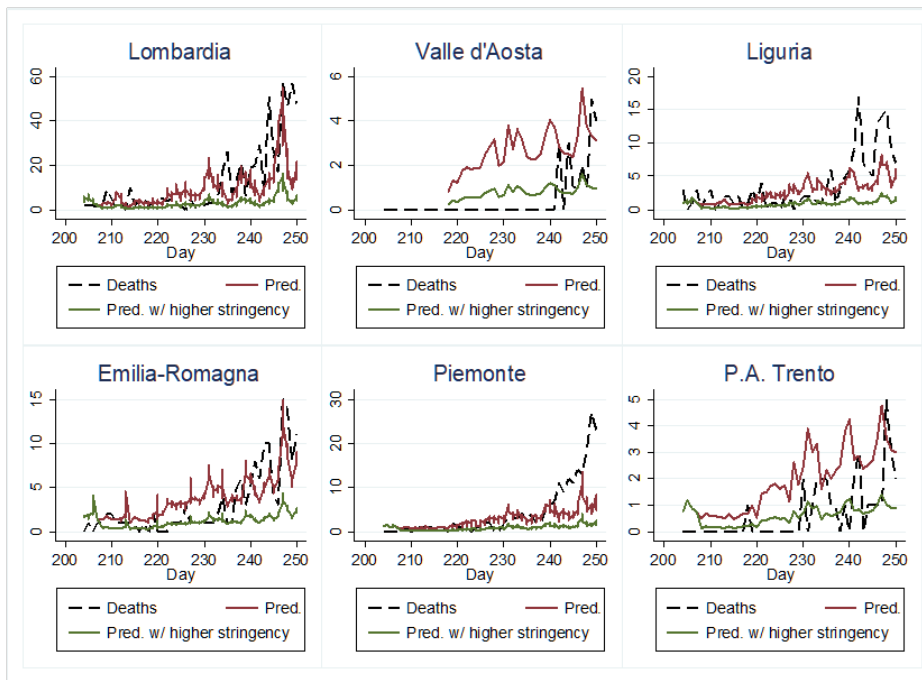
Figure 2 – Deaths over time, real prediction, prediction in the case of higher stringency level.

4. Conclusive remarks

We should be aware that mitigating the spread of infections is a cooperative process: hence, all policymakers (State and Regional authorities) should manage communication to motivate the citizens and avoid contradictory behaviours that confuse the population. Indeed, it is necessary to act to address people's behaviours, as the defeat of COVID-19 begins in people's minds.

But it is not just a psychological and political communication problem. The role played by the closure of workplaces, except for essential activities, should also be borne in mind. In the period that began on the 14th of September, the contribution of workplace-related mobility to the deaths count has almost doubled, which leads us to question whether in the second phase of the pandemic there has been some hesitation in taking more incisive measures, such as the partial closure of productive activities.

As we have seen, we would have been saved hundreds of deaths if more restrictions were promptly introduced before schools' reopening. With no additional interventions, the number of lost lives will eventually become much greater than that suffered in the very first period of the pandemic (Vollmer et al., 2020). Moreover, it should be remarked that timeliness in introducing restrictive measures is essential to reduce their required duration (Chang et al., 2020).

References

Buonomo, B., Della Marca, R. (2020). Effects of information-induced behavioural changes during the COVID-19 lockdowns: the case of Italy. *Royal Society Open Science*, **7**(10): 201635.

Chan, S., Chu, J., Zhang, Y., Nadarajah, S. (2021). Count regression models for COVID-19. *Physica A: Statistical Mechanics and its Applications*, **563**: 125460.

Chang, S.L., Harding, N., Zachreson, C., Cliff, O.M., Prokopenko, M. (2020). Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature Communications*,

11(1), pp. 1-13.

Dipartimento della Protezione Civile (2020). COVID-19 Italia - Monitoraggio situazione. Available at: https://github.com/pcm-dpc/COVID-19.

Google LLC (2020). Google COVID-19 Community Mobility Reports. Available at: https://www.google.com/covid19/mobility.

Gruppo della Sorveglianza COVID-19 (2020). Caratteristiche dei pazienti deceduti positivi all'infezione da SARS-CoV-2 in Italia: Dati al 2 dicembre 2020. Available at: https://www.epicentro.iss.it/coronavirus/bollettino/Report-COVID-2019_2_dicembre.pdf.

Hale, T., Boby, T., Angrist, N., Cameron-Blake, E., Hallas, L., Kira, B., Majumdar, S., Petherick, A., Phillips, T., Tatlow, H., Webster, S. (2020). Variation in Government Responses to COVID-19. Blavatnik School of Government, Working Paper, Version 9.0. Available at: https://www.bsg.ox.ac.uk/research/publications/variation-government-responses-covid-19.

Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.M., ..., Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of Clinical Medicine*, **9**(2): 538.

Ministero dell'Interno (2020). Coronavirus, i dati dei servizi di controllo. Available at: https://www.interno.gov.it/it/coronavirus-i-dati-dei-servizi-controllo.

Reinders Folmer, C., Kuiper, M.E., Olthuis, E., Kooistra, E.B., de Bruijn, A.L., Brownlee, M., ..., van Rooij, B. (2020). Maintaining Compliance When the Virus Returns: Understanding Adherence to Social Distancing Measures in the Netherlands in July 2020. Amsterdam Law School Research Paper No. 2020-53. Available at SSRN: https://doi.org/10.2139/ssrn.3682546.

Rinaldi, G., Paradisi, M. (2020). An empirical estimate of the infection fatality rate of COVID-19 from the first Italian outbreak. medRxiv. Available at: https://doi.org/10.1101/2020.04.18.20070912.

Romagnani, P., Gnone, G., Guzzi, F., Negrini, S., Guastalla, A., Annunziato, F., ..., De Palma, R. (2020). The COVID-19 infection: lessons from the Italian experience. *Journal of Public Health Policy*, **41**, pp. 238-244.

Sebastiani, G., Palù, G. (2020). COVID-19 and School Activities in Italy. *Viruses*, **12**(11): 1339.

Sobol, M., Blachnio, A., Przepiórka, A. (2020). Time of pandemic: Temporal perspectives related to compliance with public health regulations concerning the COVID-19 pandemic. *Social Science & Medicine*, **265**: 113408.

Verma, B.K., Verma, M., Verma, V.K., Abdullah, R.B., Nath, D.C., Khan, H.T., ..., Verma, V. (2020). Global lockdown: An effective safeguard in responding to the threat of COVID-19. *Journal of Evaluation in Clinical Practice*, **26**(6), pp. 1592-1598.

Vollmer, M.A., Mishra, S., Unwin, H.J.T., Gandy, A., Mellan, T.A., Bradley, V., ..., Bhatt, S. (2020). Report 20: Using mobility to estimate the transmission intensity of COVID-19 in Italy: a subnational analysis with future scenarios. medRxiv. Available at: https://doi.org/10.1101/2020.05.05.20089359.

Motivation of basketball players: a random-effects logit model for the probability of winning

Silvia Bacci, Tijan Juraj Cvetković

1. Introduction

Professional sports are getting more competitive as athletes strive to improve their sports performance and sport organizations employ various coaches in order to help athletes in achieving this aim. In an environment where athletes are physically dominant and have high skill mastery, psychological factors can make a difference to prevail over other athletes. For this reason, sport psychology (Perry, 2015) plays an important role in preparing an athlete from a mental perspective, just as a coach prepares from a physical perspective. In sports, motivation is a key factor of success, hence sport organisations, decision makers, sport psychologists, and players themselves must address it constantly in order to keep it and perform at the highest levels.

Psychology offers various theories and theoretical models to explain the motivational process, its benefits, and how to create a motivational climate. In this contribution we considered McClelland's Need achievement theory (McClelland, 1961) and the Nicholls' Achievement goal theory (Nicholls, 1984). These theories have something in common: goal setting, the incentive value of success, and the probability of success. Estimating the probability of success is difficult, subjective and, often, inaccurate. An error in any step of the motivational process may lead to a mistake in the role assignment, performance, and goal setting.

This paper aims at estimating the probability of success and, consequently, at making clear the motivational process such that a team or an athlete can be easily assigned to a certain role, can enhance their performance, and can set a goal as in deciding what segments of a sport must be improved. The estimation of the success probability relies on detecting the variables that affect the probability of winning in a statistically significant way. As these variables differ according to various sports, in this paper we focus on basketball, in particular the U.S. National Basketball Association (NBA). The study is based on the analysis of the traditional box scores of the regular season games played in the seasons 2016-17, 2017-18, 2018-19, and 2020-21. Because of the hierarchical structure of data at issue, with multiple observations for each team, a random intercept logit model was formulated and estimated.

The remaining part of the paper is organized as follows. The theoretical background concerning the motivational process from a psychological point of view is illustrated in Section 2, data are described in Section 3, and the main results related with the random intercept logit model are shown in Section 4. Finally, some remarks conclude the paper.

2. Motivation

Need achievement theory (McClelland, 1961) is a theory that explains what a person goes through when he/she decides to adopt a certain behaviour. McClelland considered the one's implemented behaviour as the result of a combination between personality traits and situational, resultant, and emotional factors, as illustrated in Figure 1. In detail, there are two main personality traits driving the behaviour along alternative paths: "need to achieve" and "need to avoid failure". The need to achieve is characterized by a drive to successfully compete with the standards of excellence, whereas the need to avoid failure distinguishes for a negative motivation oriented to avoid failure and criticism. These factors link with situational factors, including the probability of success and the incentive value of success. A person weights his/her probability to success and what he/she stands to gain from it. This interaction is crucial as its

resultant leads to either approach success or avoid failure. Moreover, emotional factors influence whether we focus on pride or shame. As result of these two main paths, the implemented behaviour consists, respectively, in seeking a challenge and enhanced performance or avoiding challenges, less effort and risk.

Achievement goal theory (Nicholls, 1984) is an orthogonal theory based on the persons approach to a task. Nicholls emphasizes the journey to the goal rather than the results of the goal itself. In the achievement goal theory a person focuses on skill mastery, a self-comparative perspective in which beating a previous personal result is success, or on the ego in which success is determined by comparison with others.
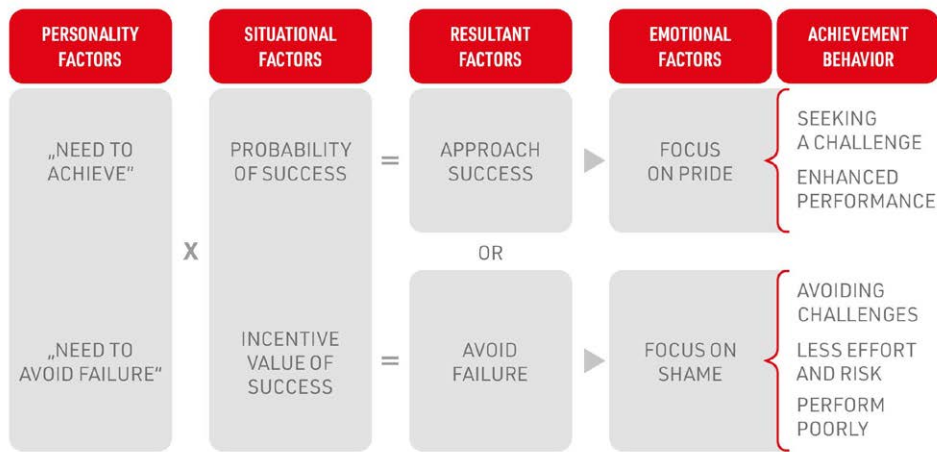
Figure 1 - Illustration of McClelland's (1961) Need Achievement Theory (our elaboration).

3. Data

The data used in the model was collected on websites NBA.com and BasketballReference.com. The dataset was constructed using the traditional box score statistics of the NBA for each game played in the seasons 2016-17, 2017-18, 2018-19 and 2020-21. The traditional box score contains information about: opposing teams, final outcome of the match (in terms of winning and losing), duration of the game (in minutes), total points scored, field goals made, field goals attempted, field goal shooting percentage, 3 point field goals made, 3 point field goals attempted, 3 point field goal percentage, number of free throws made, number of free throws attempted, percentage of free throws, offensive rebounds, defensive rebounds, total rebounds, assists, number of stolen balls, number of lost balls, number of blocks, personal fouls.

Data was arranged with a record per game (i.e., two teams in a single row) and variables were rescaled with extreme values omitted to avoid singularities. Due to changes in the leagues structure, omitted games include games played after the implementation of the Play-in tournament in the 2020-21 season and the 2019-2020 season in order to avoid variance due to circumstances. The resulting dataset is composed of 4,770 games played by 30 teams.

4. Random intercept logit model

To properly address the multilevel data structure, consisting in multiple observations per team, the probability of winning is modelled through a random-intercept logit model, where teams are the upper-level units and games are the lower-level units. The dependent variable of the model is a binary one equal to 1 if the team won the game and 0 otherwise.

As concerns the independent variables, in order to estimate the probability of success for a team we considered the *differences of the per game statistics*, which are an average of the variables based on previous games, *and the opponent per game statistics*, which are averages of other teams performances against the team of interest.

For the sake of clarity, we illustrate how to build the independent variables and to estimate the probability of success of a game played between the Utah Jazz (Team A) and the Sacramento Kings (Team B). Let us consider the following variables:
-   FGM: field goals made
-   FGA: field goals attempted
-   3PM: 3-point field goals made
-   3PA: 3-point field goals attempted
-   FT%: free throw percentage
-   DREB: defensive rebounds
-   REB: total rebounds
-   AST: assists
-   STL: steals
-   PF: personal fouls

The values of per game statistics representing the offensive performance of Team A and Team B, respectively, are displayed in Table 1, whereas the opponent per game statistics, representing the defensive performance of Team A and Team B, are reported in Table 2.

Table 1 - Per game statistics

Variable	Team A	Team B
FGM	41.3	42.6
FGA	88.1	88.6
3PM	16.7	12.2
3PA	43.1	33.4
FT%	79.3	74.3
DREB	37.6	32
AST	23.6	25.5
STL	6.5	7.5
TOV	14.2	13.4
PF	18.6	19.4

Table 2 - Opponent per game statistics

Variable	Team A	Team B
FGM	40.9	43.6
FGA	91.4	89.4
3PM	10.9	12.4
3PA	31.8	32.6
FT%	76.8	78.7
DREB	32.8	34.6
AST	22.3	25.3
STL	7.7	7.6
TOV	11.5	13.7
PF	19	18.7

Cross averaging Team A per game statistics and Team B opponent per game statistics, and vice-versa, results in a set of variables that take into account both offense and defense of the teams of interest (Table 3). This set of variables considers how the Utah Jazz (Team A) attack will vary against the Sacramento Kings (Team B) defence.

Table 3 - Cross averages of Team A per game and Team B opponent per game statistics, and vice-versa

Variable	Team A	Team B	Difference
FGM	42.4	41.8	0.7
FGA	88.8	90	-1.2
3PM	14.6	11.5	3
3PA	37.8	32.6	5.3
FT%	79	75.6	3.4
DREB	36.1	32.4	3.7
AST	24.4	23.9	0.5
STL	7.1	7.6	-0.5
TOV	13.9	12.4	1.5
PF	18.7	19.2	-0.6

Differences in the last column of Table 3 are then used as independent variables in the random-intercept logit model.

Estimates of the fixed effects of the model are shown in Table 4 (letter "d" before the variable names stays for "difference") and the related correlation matrix in Table 5. The selected model fits data in a very satisfactorily way, being the conditional and marginal pseudo-R^2 equal to 95.5% (Nakagawa and Schielzeth, 2013).

Table 4 – Random-intercept logit model: estimates of fixed effects (significance level 5%)

Variable	Estimate	Std. Error	Z-value	p-value
(Intr)	0.000	0.063	0.002	0.999
dFGM	0.745	0.026	27.876	<0.0001
dFGA	-0.126	0.009	-14.474	<0.0001
d3PM	0.624	0.023	26.572	<0.0001
d3PA	-0.053	0.009	-6.114	<0.0001
dFT%	7.538	0.442	17.034	<0.0001
dDREB	0.391	0.018	21.930	<0.0001
dAST	-0.022	0.011	-2.110	0.035
dSTL	0.145	0.020	7.390	<0.0001
dTOV	-0.421	0.023	-17.9	<0.0001
dPF	-0.444	0.016	-26.724	<0.0001

We note that, in addition to variables displayed in Table 2, we investigated other possible determinants that, however, did not result statistically significant. In particular, no significant effect resulted for the difference in field goal percentage, 3-point field goal percentage, free throws made, free throws attempted, offensive rebounds, rebounds, blocks, and for the game season (dummies were added to the model for seasons 2016-2017, 2017-2018, 2018-2019 versus season 2020-2021).

Table 5 – Correlation matrix for significant independent variables.

	dFGM	dFGA	d3PM	d3PA	dFT%	dDREB	dAST	dSTL	dTOV
dFGM									
dFGA	-0.488								
d3PM	0.449	-0.077							
d3PA	0.071	-0.238	-0.587						
dFT%	0.503	-0.173	0.316	-0.086					
dDREB	0.074	-0.006	0.396	-0.038	0.000				
dAST	-0.328	0.102	-0.158	-0.082	-0.060	-0.021			
dSTL	0.122	-0.041	0.142	-0.050	0.139	0.020	0.017		
dTOV	-0.129	0.285	-0.344	0.056	-0.016	-0.717	0.010	0.430	
dPF	-0.696	0.137	-0.512	0.082	-0.439	-0.149	0.110	-0.180	0.050

5. Conclusions

By analysing the traditional box scores of regular season games of the National Basketball Association (NBA) played in the seasons 2016-17, 2017-18, 2018-19 and 2020-21 we found several variables influencing the probability of winning, such as field goals made, field goals attempted, 3-point field goals made, 3-point field goals attempted, free throw percentage, number of defensive rebounds, number of assists, number of steals, number of turnovers and number of personal fouls.

Knowing the effect of these variables on the probability of winning helps a sport organization to improve the motivation of its athletes and to adopt a team-oriented approach to games. By objectively defining the probability of success and knowing what aspects of the game to focus on, the team decision makers can make changes accordingly. For instance, the roles assignment within a team can be improved assembling a team of players that are individually specialized in the significant categories and can consistently obtain values favouring the probability of winning. Moreover, goal setting such as keeping the opposing team under a certain number of made field goals or any other category, or rather prioritizing certain categories to maximise the probability of winning, can be easily identified benefiting both the team as a whole, by improving its chances, and the single athletes, by making him/her more proficient in a single task.

For the future research, we intend to investigate the role of an additional independent variable aimed at considering how injuries of key players affect the probability of winning. The role of team key players will be determined by analyzing their win share statistics. In particular, it will be interesting to assess the effect on the probability of winning of the number of injured or missing key players (none, one, more than one) in a game.

References

"NBA Win Shares - Basketball-Reference.com". Basketball-Reference.com. Retrieved 2016-03-25.

McClelland, D. C. (1961). *The achieving society*. Van Nostrand, Princeton, NJ.

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, pp. 133-142.

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review,* **91**(3), pp. 328-346.

Perry, J. (2015). *Sport psychology: A complete introduction.* Hodder & Stoughton, UK, pp. 105-123.

Reducing inconsistency in AHP by combining Delphi and Nudge theory and network analysis of the judgements: an application to future scenarios

Simone Di Zio

1. Introduction

The Delphi is a widely used method for collecting data from panels of experts (Dalkey and Helmer, 1963) and its key characteristics are: anonymity, interaction, controlled feedback, and statistical aggregation of responses (Rowe and Wright, 1999), while the main goal is reaching a consensus among the panel members on the issue dealt with (Linstone and Turoff, 2011). Another well-known and widely spread method in the context of decision making is the Analytic Hierarchy Process (AHP), a Multi-Criteria Decision-Making (MCDM) method designed to solve problems containing multiple conflicting criteria (Pirdashti et al., 2011). Developed by Thomas Saaty (Saaty, 1980), it has many advisable properties, such as the combination of subjective aspects, the chance of integrating objective and subjective data, and a way to combine individual and group priorities.

As far as we know, no study takes advantage of the Delphi features for reducing the inconsistency in the AHP matrices, a known problem but practically inevitable, given that it is mostly the product of cognitive biases (Bonaccorsi et al., 2020). In case of high inconsistency, generally experts are asked to evaluate again the AHP matrices, but no expert likes to give again judgements because the first ones are inconsistent, which basically means wrong. Furthermore, even if they accept, there are no guarantees that the new judgements are less inconsistent. Our proposal is to exploit the Nudge theory, which proposes suggestions to influence the behaviour of groups involved in a decision-making process (Thaler and Sunstein, 2008). A Nudge is known as a "gentle push" to make better choices which, in our context, means more consistent evaluations. In this paper we propose a new method that exploits a combination of the Delphi method and the Nudge theory to reduce the inconsistency of the AHP matrices. The method has several advantages. In addition to reducing inconsistency, it allows the collection of textual material (expert comments), a valuable data in any decision-making context. A function of the inconsistency is used as stopping criterion of the Delphi rounds. Given the Delphi logic, the participants know from the beginning that they will be reconsulted, therefore they do not feel scrutinized or pressured and they are never told that their judgments are inconsistent. This, at least in principle, ensure freer and more sincere participation and a more willing attitude to evaluate again the judgments. Finally, since at each Delphi round only the matrices with the highest inconsistency values are sent back to the experts, round after round the length of the questionnaire diminishes, and this help in reducing the dropout.

In the next sections we provide an overview of the AHP method, while section 3 shows how Nudge theory can help in reducing the inconsistency of the AHP matrices. Sections 4 presents a case study and finally the paper ends with some concluding remarks.

2. The inconsistency of the AHP matrices

The Analytic Hierarchy Process (AHP) is a general theory of measurement, useful to derive ratio scales for multi-criteria decision problems, suitable when the decision problem is complex and ill-structured. The decision factors are organized in a hierarchical structure where *criteria* and *alternatives* are compared pairwise using the Saaty scale (Saaty 1980). The goal is to find a set of weights ($W_1, W_2, ...$) for each level of the hierarchy (called *local weights*) and, from these, a vector

Simone Di Zio, Gabriele d'Annunzio University, Italy, s.dizio@unich.it, 0000-0002-9139-1451

of *global weights* $(GW_1, GW_2, \dots, GW_N)$ representing a rating of the alternatives in achieving the decision problem (N denotes the number of alternatives). The AHP can be adapted to group decisions (group AHP), and there are two families of methods for the combination of the individual preferences (Ossadnik et al., 2016), known as the Aggregation of Individual Judgements - AIJ - and the Aggregation of Individual Priorities - AIP - (Wu et al, 2008). Anyhow, no technique considers the variability in the distribution of responses, so that AIJ and AIP approaches do not take account of the degree of consensus/dissensus among participants, a fundamental issue in a group decision setting (Pirdashti et al., 2011). A voting procedure can overcome these limitations (Lai et al. 2002), but a majority vote is a "winner-take-all" system, where the opinions of the losers are completely disregarded (Di Zio and Maretti, 2014). This is why some scholars have proposed an integration of the AHP and the Delphi (Tavana et al. 1993; Di Zio and Maretti, 2014) which gives the Delphi the task of structuring a convergence towards a single solution shared by all.

Given a_{ij}, the pairwise judgement of alternatives i and j, for a perfect consistent matrix we should have $W_i/W_j = a_{ij}$ ($\forall i, j$) and $a_{ij} = a_{ih} \cdot a_{hj}$ ($\forall i, j, h$) but human judgements are never perfectly consistent and in practical applications the equalities do not occur. Inconsistency in expert judgments has been observed in many fields and, for lack of space, we refer the reader to the vast specialized literature. In short, inconsistency is practically inevitable, because it is the product of cognitive biases (Bonaccorsi et al., 2020) and/or problem complexity. Consequently, there is a need to check the consistency through the calculation of a consistency index. The Consistency Ratio (CR), is the most common index used to check for consistency (Brunelli, 2018), calculated as $CR = CI/RI$, where $CI = (\lambda_{max} - n)/(n - 1)$, λ_{max} is the maximum eigenvalue of the matrix and RI (the random index) is the average of the CIs calculated over many random square matrices, reciprocal and positive. As a rule of thumb, introduced by Saaty (1980), if $CR \leq 0.1$ the judgements of a matrix can be considered consistent, otherwise the matrices must be reviewed by the expert (Liao, 2010), as many times until to have $CR \leq 0.1$. The critical point is going back and stress the expert telling him/her that he/she made a wrong evaluation that needs to be revised.

All that being said, the reduction of the inconsistency in the AHP method remains an open issue, and here we propose a new approach which involves asking the experts for new evaluations according to the Delphi logic, in a structured and iterative procedure that, by means of *nudges*, gently push them towards more consistent solutions.

3. Reducing the inconsistency by combining the Delphi and the Nudge theory

Although it still has open issues (Pill, 1971) - such as how to choose the experts, how many experts to include in the panel or how to measure the expertise - the Delphi is a method that offers undoubted advantages in the context of group decisions. In the Delphi-AHP the experts are consulted more than once, and starting from the second round, for each AHP matrix, we propose to give a nudge as feedback. By using a "nudge approach" we obtain both a reduction of the inconsistency of the AHP matrices and the elimination of the problem of choosing an aggregation method. After the first round (time t_1) we get $R + 1$ matrices for each expert. With $\mathbf{A}_m^{1,t_1}$ we denote the 3D array containing the $N \times N$ pairwise comparison matrices according to the first criterion, at time t_1, where $m = m_1, m_2, \dots, m_M$ denotes the expert and M the cardinality of the panel. Since each participant give $s = N(N - 1)/2$ judgements, we have s vectors of size M. For the first criterion, the vector of the generic pairwise comparison (i, j) is $\left[a_{ij,m_1}^{1,t_1}, a_{ij,m_2}^{1,t_1}, \dots, a_{ij,m_M}^{1,t_1} \right]$. To synthetize these judgements, we use the median (other syntheses are possible) and as a result, we obtain a matrix representing the judgments of the whole panel after the first round, say $\mathbf{A}_{med}^{1,t_1}$. On this matrix we calculate the consistency ratio CR_{med}^{1,t_1} and by using the 17 values of the Saaty scale $(1/9, 1/8, \dots, 8, 9)$ we replace the first element of $\mathbf{A}_{med}^{1,t_1}$, namely $a_{12,med}^{1,t_1}$ (as well as its symmetric $a_{21,med}^{1,t_1} = 1/a_{12,med}^{1,t_1}$) obtaining 17 different matrices. On each resulting matrix we calculate the

consistency ratio among which we find the smallest - say $\text{CR}_{12,best}^{1,t_1}$. This value is the result of a specific value of the Saaty scale, say $\text{V}_{12,best}^{1,t_1}$. This figure represents the theoretical assessment which, for the cell (1,2), gives the best consistency of the matrix, given all the other values. By repeating the same search for the upper triangular of the matrix, we obtain s different "best CRs" among which we find the smallest, that represents the "best of the bests": $\text{CR}_*^{1,t_1} = min\{\text{CR}_{ij,best}^{1,t_1} : i = 1, ..., n; j > i\}$. We denote the position of this value with (i_{best}, j_{best}) and its corresponding value of the Saaty scale with SA_{nudge} (actually our nudge), that is the judgement that most improves the consistency of the matrix.

In the second round of consultation, the panel is invited to judge each a_{ij}^r (and c_{lm}) inside the proposed interquartile range - $IQR_{i,j} = \left[Q1_{ij}^{t_1}; Q3_{ij}^{t_1}\right]$ - where $Q1_{ij}^{t_1}$ and $Q3_{ij}^{t_1}$ are, respectively, the first and third quartile in the distribution of judgements in the cell i, j (Di Zio and Maretti, 2014). The quantity $\Delta = IQR_{i_{best}, j_{best}}/2$ is used to create a symmetric interval around SA_{nudge}. For the judgement in position (i_{best}, j_{best}), in the second round, instead of the IQR, the interval proposed to the panel is $\left[SA_{nudge} - \Delta; SA_{nudge} + \Delta\right]$. Therefore, among the s proposed intervals $s - 1$ are $IQRs$ but one is a Nudge which gently pushes the respondents towards a more consistent matrix. The same process applies to all the matrices of the hierarchy and the procedure is repeated iteratively in the following rounds. If the consensus triggers, there will be a progressive reduction of the consistency ratios: $\text{CR}_{med}^{r,t_1} \geq \text{CR}_{med}^{r,t_2} \geq \text{CR}_{med}^{r,t_3} \geq \cdots$.

This method has several advantages. The aggregation of judgements is managed optimally, by considering the degree of consensus, and this reduces, at least in principle, the dropout rate. Simultaneously, we reduce the inconsistency of the matrices in a gentle way, because there are no pressures on the participants. The experts do not perceive any kind of "mistake message" and are softly driven to revise their judgements. A right nudge, in the AHP context, "pushes gently" the participants to more consistent judgements. So, the method stimulates consensus and reduces inconsistency at the same time.

The rule to stop the Delphi iterations is twofold. To make the benefits of the Delphi at least two rounds must be performed, therefore the first stopping criterion is $i \geq 2$ (here i denotes the rounds). During the rounds we have, for each matrix, a sequence of Consistency Ratios $CR^{r,t_1}, CR^{r,t_2}, ..., CR^{r,t_i}$ (here $r = 1,2, ..., R + 1$ and we removed the subscript med to simplify) and the second stopping criterion is that at least one CR in the sequence is less or equal than 0.1. After the round t_2, for the matrix r, we have four possible cases. 1) $CR^{r,t_1} > 0.1$ and $CR^{r,t_2} \leq 0.1$; the Delphi for the matrix r stops and as result we take the matrix coming from the second round: $\mathbf{A}_{med}^{r,t_2}$. 2) $CR^{r,t_1} \leq 0.1$ and $CR^{r,t_2} > 0.1$; the Delphi stops, but the matrix we take is $\mathbf{A}_{med}^{r,t_1}$. 3) $CR^{r,t_1} \leq 0.1$ and $CR^{r,t_2} \leq 0.1$; the Delphi stops, and we choose between $\mathbf{A}_{med}^{r,t_1}$ and $\mathbf{A}_{med}^{r,t_2}$ the matrix with the lowest inconsistency. 4) $CR^{r,t_1} > 0.1$ and $CR^{r,t_2} > 0.1$; in this last case only the condition $i \geq 2$ holds, therefore the Delphi continues. This double-stopping criterion is appropriate because after a reduction of the inconsistency, there is no guarantee that continuing the rounds the CR decreases monotonically. If after k rounds, for a matrix no index in the sequence is less than 0.1, we suggest the following solution: take the round z such that CR^{r,t_z} is the minimum and hold the matrix used for the calculation of the intervals for the round $z + 1$. This matrix, by definition, has $CR_*^{r,t_z} < CR^{r,t_z}$. Since the above algorithm applies to each matrix of the hierarchy, it may happen that for one matrix only two rounds are necessary while for another matrix we can have three or more rounds. The advantage is that the length of the AHP questionnaires reduces during the rounds. The result of the method is a vector of global weights, with lower levels of inconsistency in the pairwise matrices than the classic AHP.

4. Application on four future scenarios with network analysis

We applied the proposed method in the evaluation of four future scenarios on the genetic modification experiments. It is called CRISPR the new technology that allows splicing of DNA molecules, and in the future, it could allow human selection of characteristics of children, including escaping of many diseases. The ethics of this technology is obviously questionable. Starting from these considerations, Theodore J. Gordon (one of the fathers of the Delphi method) sketched four brief future scenarios on CRISPR technology. For lack of space, we do not report the scenarios in full but only their titles: Scenario A. *Genetic tech self-regulation*; Scenario B. *Genetic tech external control*; Scenario C. *Genetic tech uncontrolled*; Scenario D. *Genetic tech downside*. Each scenario represents an alternative of the AHP hierarchy.

Following Gordon and Glenn (2018), the main factors measuring the usefulness of a scenario, and which here constitutes the criteria of the AHP, are the following: *Plausibility*: the paths to the futures must be seen as feasible and may not be viewed as impossible. *Consistency*: the paths to the futures and the resulting images must not be mutually contradictory. *Simplicity*: a good scenario describes paths to the future scenario in a way that is easily understood. Therefore, we had 4 alternatives (the scenarios) and 3 criteria, and we wanted to find a ranking of importance of the scenarios according to these criteria. The survey was performed on Alchemer (https://app.alchemer.com) where each pairwise comparison was built on a radio button that reproduces the whole Saaty scale. This avoided that the respondents fill in the matrices, in general a complicated task for non AHP-experts.

The panel consisted of 26 experts, recruited around the world, diversified according to age, gender, expertise and employment, and having skills both in the field of futures studies and genetics. For each round they gave 21 pairwise judgements and voluntary comments. For each round and for each matrix (A^1, A^2, A^3, C) we obtained the consistency ratios reported in Table 1.

Table 1. Consistency Ratios along three Delphi rounds

Round	Plausibility	Consistency	Simplicity	Criteria
t_1	0.0936	0.0035	**0.0195**	0.1591
t_2	0.0514	0.0087	0.0225	**0.0056**
t_3	**0.0366**	**0.0018**	0.0295	0.1850

For the calculation of the local and global weights we take the matrix resulting from the last round for *plausibility* ($CR^{1,t_3} = 0.0366$) and *consistency* ($CR^{2,t_3} = 0.0018$). For the criterion *simplicity* the best value derives from the first round ($CR^{3,t_1} = 0.0195$) and for the comparison of the three criteria we take the matrix coming from the second round ($CR^{C,t_2} = 0.0056$). In all cases the values are very good, being all well below the 0.1 threshold. The result consists of a vector of global weights, which quantifies the relative importance of each future scenario. The best scenario, according to the panel of experts, is scenario B, Genetic tech external control ($GW_B = 0.52$). It follows scenario A, Genetic tech self-regulation ($GW_A = 0.28$), and scenario D, Genetic tech downside ($GW_D = 0.10$). The last is scenario C, Genetic tech uncontrolled ($GW_C = 0.09$). About the local weights of the criteria, the experts considered *plausibility* as the most important criterion ($W_{Pla} = 0.47$). Following we have *consistency* ($W_{Con} = 0.43$) and *simplicity* ($W_{Sim} = 0.10$).

After that, we explored the network structure of the scenarios and criteria. A network refers to a structure representing a group of objects and relationships between them, and its mathematical representation is a graph, which consists of nodes and edges. Since each scenario/criterion is linked to the others through a preference ratio, it is useful to represent the results of the AHP through weighted direct graphs, in which the nodes are the scenarios/criteria and the edges are proportional to the geometric mean (or median) of the judgments provided by the experts. By considering the matrices with the lowest CRs (Table 1, bold digits) we obtained the four digraphs of Figure 1.

From each graph emerges, with a single glance, the whole structure of the preferences expressed by the panel in comparing the future scenarios under each criterion and the criteria, as well as the

structure of relationships between the scenarios/criteria, with evident advantages over the representation through matrices. Also, we can build a network for each expert and, even more interesting, we can consider each expert as a layer in a multiplex network, that is a network in which the same set of nodes are connected via more than one type of links (Kyu-Min et. al, 2015). Besides, we can consider each criterion as a layer, to study the interactions between scenarios and criteria, or even each Delphi round as a layer, to explore, within each criterion, the interactions between scenarios and rounds. In short, there are many possibilities to represent and analyse the outputs of a Delphi-AHP through the Network Analysis. So, we can study whether scenarios behave similarly across experts, across Delphi rounds or across criteria. Hence, this is not only a way of visualizing the results but a statistical tool for modelling the Delphi-AHP data in a way that to highlight the structure of relationships between experts, scenarios, criteria and Delphi rounds.

Figure 1. Network representation of the results (nodes sizes are proportional to the closeness)
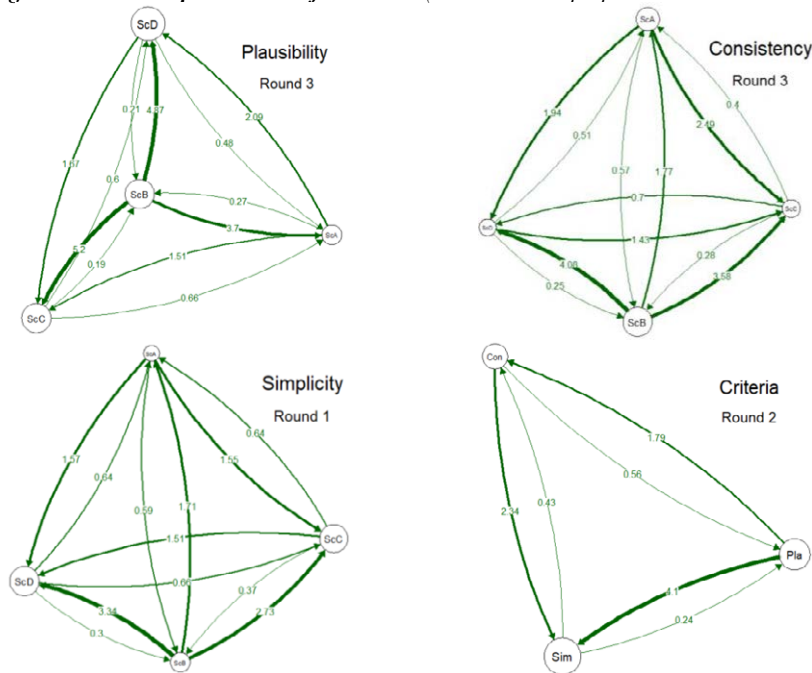
To give a taste of the measures that can be computed, we calculated the closeness (CL) of the networks of the Figure 1 (where nodes sizes are proportional to CL), which gives information on how close a scenario (or a criterion) to all the others is. Plausibility network: $CL_A = 0.549$, $CL_B = 0.800$, $CL_C = 0.799$, $CL_D = 0.925$. Consistency network: $CL_A = 0.627$, $CL_B = 0.779$, $CL_C = 0.474$, $CL_D = 0.457$. Simplicity network: $CL_A = 0.335$, $CL_B = 0.422$, $CL_C = 0.597$, $CL_D = 0.626$. Criteria network: $CL_{PLA} = 1.246$, $CL_{CON} = 1.014$, $CL_{SIM} = 1.488$. Scenario D is the "closest" to the others under plausibility and simplicity criteria, while under consistency the scenario with major closeness is B, and simplicity is the criterion with the higher closeness.

5. Concluding remarks

We have introduced a new method to use the Delphi method to nudge responses of participants toward better consistency in the AHP pairwise comparison matrices. The network analysis helps to depict the structure of interactions between alternatives and criteria of the AHP hierarchy. We applied the method for the evaluation of four future scenarios, dealing the management of genetic modification technologies. The study confirmed quite well the research hypothesis, since the inconsistency in all the AHP matrices remained under, or dropped below, the threshold of 0.1.

Although the rounds of the Delphi must stop after the second round, in the application we performed three rounds for all the matrices, to explore all the potentialities of the method. The method removes the problem of choosing an aggregation method of the individual judgements, because the Delphi produces a convergence toward a synthesis of the evaluations which includes all points of view, even the extremes or the minority ones. By using a multiplex network approach, the structure of relationships between experts, scenarios, criteria and Delphi rounds can be studied.

As future developments we can think of the graph representation as a tool to be included in the Delphi questionnaires, which help to visualize in real time the answers that each participant gives. Also, when considering each expert as a layer of a multiplex network, the similarity measures between layers could be exploited to explore new measures of consensus in the Delphi method and new ways of aggregating the individual judgements could be also studied.

References

Bonaccorsi, A., Apreda R., Fantoni, G. (2020). Expert biases in technology foresight. Why they are a problem and how to mitigate them. *Technological Forecasting & Social Change*, **151**, pp. 1-17.

Brunelli, N. (2018). A survey of inconsistency indices for pairwise comparisons. *International Journal of General Systems*, **47**(8), pp. 751-771.

Dalkey, N.C., Helmer, O. (1963). An Experimental Application of Delphi Method to the Use of Experts. *Management Science*, **9**, pp. 458-467.

Di Zio, S., Maretti, M. (2014). Acceptability of energy sources using an integration of the Delphi method and the analytic hierarchy process. *Quality & Quantity*, **47**(6), pp. 2973–2991.

Gordon, T.J., Glenn, J.C. (2018). Interactive Scenarios, in *Innovative Research Methodologies in Management, Volume II: Futures, Biometrics and Neuroscience Research*, eds. L. Moutinho and M. Sokele, Palgrave Macmillan, London (UK).

Kyu-Min, L., Byungjoon, M., Kwang-Il, G. (2015). Towards real-world complexity: an introduction to multiplex networks. European Physical Journal B, **88**(2), pp. 1-20.

Lai, V.S., Wong, B.K., Cheung, W. (2002). Group decision making in a multiple criteria environment: A case using the AHP in software selection. *European Journal Operational Res.*, **137**, pp. 134-144.

Liao, C.N., (2010). Supplier selection project using an integrated Delphi, AHP and Taguchi loss function. *ProbStat Forum*, **3**, pp. 118-134.

Linstone, H.A., Turoff M. (2011). Delphi: A brief look backward and forward. *Technological Forecasting & Social Change*, **78**(9), pp. 1712-1719.

Ossadnik, W., Schinke, S., Kaspar, R.H. (2016). Group Aggregation Techniques for Analytic Hierarchy Process and Analytic Network Process: A Comparative Analysis. *Group Dec. and Negotiation*, **25**, pp. 421-457.

Pill, J. (1971). The Delphi method: Substance, context, a critique and an annotated bibliography. *Socio-Economic Planning Sciences*, **5**(1), pp. 57-71.

Pirdashti, M., Omidi M., Pirdashti H., Hassim M.H. (2011). An AHP-Delphi Multi-Criteria Decision Making Model with Application to Environmental Decision-Making. *Iranian Journal of Chemical Engineering*, **8**(2), pp. 3-17.

Rowe, G. and Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis, *International Journal of Forecasting*, **15**(4), pp. 353-375.

Saaty, T.L. (1980). *The Analytic Hierarchy Process*. McGraw-Hill, New York (NY).

Tavana, M., Kennedy, D.T., Rappaport, J., Ugras, Y.J. (1993). An AHP-Delphi Group Decision Support System applied to conflict resolution in hiring decisions. *Journal of Management Systems,* **5**(1), pp. 49-74.

Thaler, R. and Sunstein, C.R. (2008). *Nudge. Improving decisions about health, wealth and happiness*. Yale University Press, New Haven (CT).

Wu, W.H., Chiang, C.T., Lin, C.T. (2008). Comparing the aggregation methods in the analytic hierarchy process when uniform distribution. *Transactions on Business and Economics*, **3**(5), pp. 82-87.

Mapping and factoring the 2007 ATECO categories in regard to specialised human capital

Luigi Fabbris, Paolo Feltrin

1. Introduction

The paper describes an exercise of classification of the five-digit categories of the 2007 Ateco classification system of economic activities (https://www.istat.it/en/methods-and-tools/classifications). The purpose of the work is to highlight the categories showing top levels of human capital (HC) in order to pinpoint the categories that are likely to lead the Italian economic growth in the near future.

An attempt to measure the effects of HC concentration in a territory was realised by Moretti (2012) in the United States. He showed that innovation can attract in a territory many other jobs and form the basis of a global knowledge economy (see also Etzkowitz and Leydesdorff, 1997). In Italy, attempts to link HC to territorial clusters was studied, among others, by Colombo and Delmastro (2002) with reference to technology incubators, Liberati et al. (2013) to science and technology parks, and Bertamino et al. (2014) to technological districts. All Italian studies conclude that locating firms within a specialised territory does not influence significantly business R&D. This may be due to the different demographic density of the United States and Italy. In this work, we ignored the location of economic activities and scouted the complementary hypothesis that HC impacts certain activities more than others and this can lead to an enduring development of the activities.

Our exercise was realised through a multivariate mapping of the two-digit Ateco categories on the basis of HC indicators and a factorisation of the indicators so to understand if and how higher competence can be considered a connection trait of certain economic categories.

The results of our analysis could be useful, among else, to evaluate possible relations between academic education and economic growth in Italy. This possible relationship correlates also with some strategies of the Italian PNRR (National Plan for Recovery and Resilience; https://www.governo.it/sites/governo.it/files/PNRR.pdf) and may help forecasting its possible outcomes.

The rest of the paper is organised as follows. Section 2 shortly describes the indicators created for defining the HC of Ateco categories and the methodology adopted for the exercise; Section 3 presents the main results of the data analysis; and Section 4 discusses the results of the statistical analysis with reference to the mainstream literature and then concludes.

2. Data and methodology

The indicators of HC associated to the Ateco categories are the following.
1. Per cent frequency of workers with a university degree (from now on also "higher HC") out of total Italian workers at year T.
2. Per cent ratio between the relative frequency of "higher HC" at years T and $T-1$.
3. Per cent frequency of workers in intellectual or scientific jobs (ISCO-02, from now on also "higher jobs") out of total workers at year T.
4. Per cent ratio between the relative frequency of ISCO02 workers at years T and $T-1$.
5. Per cent frequency of self-employed workers with a university degree ("higher HC-SE) out of total self-employed at year T.

Luigi Fabbris, University of Padua, Italy, luigi.fabbris@unipd.it, 0000-0001-8657-8361
Paolo Feltrin, University of Trieste, Italy, paolo.feltrin@gmail.com, 0000-0003-2801-5151

6. Per cent ratio between the relative frequency of higher HC-SE at years *T* and *T-1*.
7. Per cent frequency of self-employed workers in intellectual or scientific jobs (ISCO-2) out of total self-employed at year *T*.
8. Per cent ratio between the relative frequency of ISCO02-SE at years *T* and *T-1*.
9. Per cent frequency of self-employed workers out of total workers at year *T*.
10. Per cent ratio between the relative frequency of self-employed at years *T* and *T-1*.

To obtain more stable estimates, year *T* data were averaged over years 2018 and 2019 and *T-1* data were averaged over years 2011 and 2012. The Ateco categories that changed from 2011 to 2019 or were null at either year were merged or excluded from the analyses. We ended up with 84 Ateco categories. The Covid-19 pandemic particularly threatened employment; that is why, in this work, we considered anomalous, and then ignored the 2020 data.

The idea in the background of our choice of indicators was that a leading economic category is one that is qualified by a high frequency of college-educated workers and parallels that of people working in higher jobs. This frequency is evaluated for both all Italian workers and the self-employed. While the relevance of higher education as a distinctive trait of leading economic activities recurs in the mainstream literature (Autor et al., 2003 and Moretti, 2012, though the latter argues that excellent exceptions are numerous), that of self-employment as a qualitative symptom derives from studies on the future of work (European Commission, 2013; OECD, 2019), which forecast a growing relevance of self-employment for job creation or job restructuring in the next decades.

The relational analysis of the indicators was based on a Varimax-rotated principal-component factor analysis (Browne, 2001). The analysis aimed to elicit the multiple relationships between indicators and define a mapping system of categories inclusive of all intercorrelated indicators.

The R Studio package was used to compute the estimates.

3. Results

The statistical analysis of the basic indicators showed that, in Italy, both the percentage of workers with a college (from now on also "higher education") degree and that of workers in an intellectual or scientific job (from now on also "higher job") are important and show an increasing trend (Table 1). In particular, the proportion of the employed with a higher education degree was 23.3% in 2018/19 and had an amazing increase since 2011/12: +29.5% (basis 2011/12=100).

Even the employed in a higher job represented a relevant and increasing quota of Italian workers: in 2018 and 2019, the quota of workers in an intellectual or scientific job was 14.8%, with a notable increase (+14.8%) from to the basic year. This may be due to the diffusion of technological innovation also in many traditional sectors, which, in turn, activated an additional demand for highly qualified jobs.

The proportion of self-employed was relevant (22.8%, average value of the years 2018 and 2019), at least in comparison with other European countries, but in diminution (-8.3%) from 2011/12. The stream-lining concerned the categories of para-subordinates and of self-employed in craft, commerce and agriculture: in fact, most movers from these categories either retired or became employees. Instead, the number of employers and freelances increased during the examined time span (Fabbris and Feltrin, 2021). Our data show that the increase concerned both the self-employed with a college degree and those in a higher job and this increase was larger than that that involved employees. It is interesting that all indicators of level are positively skewed, and this allows pinpointing the Ateco categories with the higher levels of the examined indicators.

So, we applied factor analysis twice, once to examine the relationships among the level at 2018/19 and its variation from 2011/12 of three basic indicators (per cent of workers with a higher education degree, in a higher job, and self-employed) and another including also the level and variation of two qualified categories (workers with a college education and workers

in an intellectual or scientific job) among the self-employed. The latter, which was named 10-variable analysis, was an attempt to involve in the analysis the interactions between self-employment and the other two basic indicators.

Table 1. Mean values of HC indicators at years 2011/12 and 2018/19, Italy

Indicator	Mean 2011/12	Mean 2018/19	Variation 2018/19 vs. 2011/12
Per cent workers with a college degree	18.0	23.3	29.5
Per cent workers in an intellectual or scientific job	12.9	14.8	14.8
Per cent self-employed workers	24.9	22.8	-8.3
Per cent self-employed with college degree	20.0	26.6	33.2
Per cent self-employed in an intellectual or scientific job	16.2	19.3	18.9

Table 2. Correlation coefficients between human capital indicators (Italy, 2018/19) (significance levels in the upper triangle: ***<1‰; **<1%; *<5%; °<10%)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	=		***		***	*	***		*	
X_2	-0.039	=		***	**		*			°
X_3	0.893	-0.157	=		***		***		**	
X_4	0.008	0.501	-0.087	=						
X_5	0.842	-0.297	0.802	-0.174	=		***		**	
X_6	-0.231	0.053	-0.169	-0.155	-0.054	=				
X_7	0.828	-0.243	0.912	-0.178	0.883	-0.130	=		**	
X_8	-0.099	0.084	-0.073	-0.002	-0.027	0.133	-0.038	=		
X_9	0.275	-0.096	0.346	-0.104	0.309	0.076	0.349	-0.104	=	*
X_{10}	-0.034	-0.196	0.021	-0.175	0.083	0.178	0.019	-0.110	0.283	=

Table 3. Two-factor Varimax-rotated configuration with 6 and 10 HC indicators, 2018 and 2019, Italy.

	6 indicators		10 indicators	
	f_1	f_2	f_1	f_2
X_1	0.84	-0.45	0.94	0.15
X_2	-0.38	-0.69	-0.20	0.68
X_3	0.90	-0.32	0.95	0.01
X_4	-0.31	-0.73	-0.09	0.75
X_5	=	=	0.91	-0.15
X_6	=	=	-0.24	-0.40
X_7	=	=	0.95	-0.09
X_8	=	=	-0.10	0.09
X_9	0.59	0.09	0.40	-0.35
X_{10}	0.25	0.51	0.00	-0.62
Eigenvalue	*2.17*	*1.56*	*3.80*	*1.75*

The correlation coefficients between the indicators, presented in Table 2, showed that:
- o The indicators of the 2018/19 levels highly correlated to each other: the correlation among X1, X3, X5 and X7 is exceptionally high, since all correlation coefficients were above 0.80. Just X9 – the self-employment rate – though positively intercorrelated, is below this high level. This means that education and innovation cross-fertilise both among employees and the self-employed within certain economic activities and stay close to the bottom in others.
- o Correlations between change indicators are weaker and follow different patterns: the

correlation between the rate of variation of high education and high jobs is positive both among the complex of Italian workers (0.50) and the self-employed (0.13), while it is not significant or negative for all the other analysed variations.

Both the 6-indicator and the 10-indicator factor analyses (Table 3) showed that two factors are enough to represent the between-indicator correlations. In fact, the first two factors explained, respectively, 62% and 55.5% of the global variance. Though, the higher complexity of the 10-indicator analysis led us to privilege it for our analysis. The two-factor solution (Figure 1) showed that:

- o There is a strong positive inter-correlation between the indicators X1, X3, X5 and X7 and a mild one with X9. The five indicators describe the level of qualified workers in 2018/19, that is why we can call the first factor "high skill levels" and higher scores pinpoint the activities with a higher density of very skilled jobs.
- o There is a positive relation between the second factor and indicators X2 and X4 of time change and a negative one with X6, again of time change. The other two variation indicators (X8 and X6) do not fit neither this factor, nor the previous one. The second factor, which includes variables X2, X4 and X10 (the latter with a negative sign) can be called "positive trend of high skill activities".

Figure 1. Map of the Ateco categories on the surface defined by the first (abscissa) and second factor (ordinate) obtained with a Varimax rotated 10-indicator factor analysis, Italy (numeric codes refer to two-digit Ateco classification; X-arrows represent the indicators)
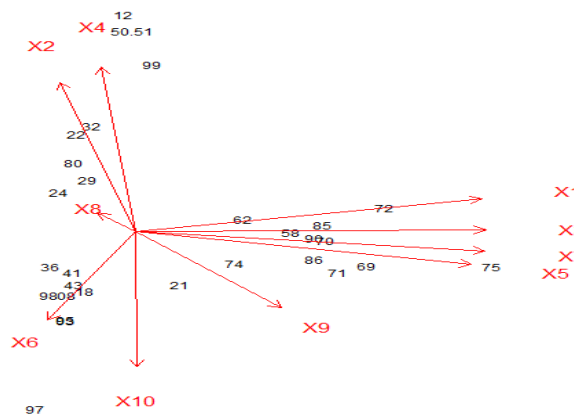
The Ateco categories represented in Figure 1 show that the categories leading the intensity scale of human capital as measured by higher education, higher jobs and entrepreneurial spirit were: 75 (veterinary services), 72 (scientific research and development), 70 (business management and consultation), 71 (studies of architecture, engineering and other technical services), 86 (health services), 69 (legal and accounting offices), 85 (education) and 90 (creative, artistic and entertainment activities). In all these categories graduates exceeded 50% of total workers and, with the exception of category 90, exceeded 60% rate of workers possessing a higher education degree. Also, the categories number 58 (editorial activities), 62 (software production) and 74 (other professional, scientific or technical activities) scored positively on this main factor. All the quoted categories but number 62 (software production) showed also a positive trend at the end of the examined period. As expected, high skill jobs are associated to innovative sectors and refer to both employees and the self-employed.

The category scoring negatively on the first factor but showing a steep qualitative increase during the considered period is number 97 (family and community assistance), meaning that the personnel to assist families for housework and/or people with impairments were less educated than average in 2011 but are notably increasing their education and skills in the last years.

4. Discussion and conclusion

In this work, we aimed to highlight the Ateco categories showing higher and/or increasing levels of human capital at 2018 and 2019. The indicators of HC intensity referred to both college-educated skills, workers' employment in higher jobs, and entrepreneurship. Taken together, the indicators aimed at representing both the rate of superior knowledge required by jobs at certain economic activities and the innovation necessary to improve products and processes, as well as the entrepreneurial spirit that should accompany knowledge and innovation as drivers of business opportunities. We examined both the level of indicators and their dynamic perspective. Variation was taken with reference to 2011 and 2012 as baseline years.

Our exercise is similar to that of defining what economists, referring to industrial clusters, call "the Marshallian trinity of information exchange, specialized suppliers, and a pool of labor with specialized skills" (Krugman, 2017). Of course, in our case, proximity does not refer to territory but to similar economic activities: paraphrasing Becattini (1990), we asked ourselves if there were economic categories sharing a system of values, views, language, expectations and behaviours, combined to an entrepreneurial culture and knowledge, that shape the productive atmosphere and drive the development of the firms and the workers in them. In particular, people working for themselves should know what hitherto and in perspective would be managed for them.

We have found that there are categories leading the trinomial: knowledge-innovation-entrepreneurial spirit. Some of them could be given for granted, such as medical, veterinary, education and R&D activities that are mainly related with top jobs. Legal, accounting, architecture, engineering and other highly technical activities require superior education and training and are often realised in a self-employed environment, either solo or in small offices.

What may be a novelty in this knowledge-oriented group of activities is that of business and management consultation and that of creative artistic and entertainment industry. Business consultant and managers are relevant to the development of both local and global businesses and work in competition to each other at national and international levels. To consult and manage firms you need not only a specific knowledge but also culture and a personality adequate to make strategic decisions. Specific education and training and the capacity to identify themselves with entrepreneurs are essential components of the professional personality of these workers.

The relevance of creative industries as a driver of local development is underlined in Moretti (2012). In Italy, these sectors refer to the so-called "four Fs" of made in Italy (fashion, food, factory automation, furniture and design), as well as tourism, leisure and information diffusion. The peculiarity of this industry is that technical activities are of the creative and cultural type.

All this raises an education issue. With reference to universities, the issue implies decisions ranging from educational strategies to practicalities, such as thinking in terms of building transferable skills, and in particular developing attitudes and skills for running a business. Educating students to start an own business and developing their business competencies could support the graduates in their job finding and raise the productive capacity of the whole economic system. Also, the matching between economic activities involving graduates and higher education paths could help universities to pinpoint their productive stakeholders and imagine future courses.

Our hypotheses of a cogent relationship between the intensity of higher education required by certain jobs and that of higher jobs of a given economic activity was confirmed. We did not find overt relations between these two variables and self-employment. The correlation between knowledge intensity and self-employment frequency was positive but the variation between the two variables was negative, namely, while the intensity of knowledge employed by firms grows in time, the number of self-employed diminishes. This does not imply that self-employment requires lower education, but the opposite. In fact, the groups of self-employed diminishing in recent years are craft, commerce and agriculture self-employed workers, who are, on average, less educated than the other self-employed.

So, while the number of self-employed diminishes, the knowledge required to them and to the

category of business owners, with or without other employees, increases. Education and training help the self-employed to become self-reliant. He may become better at finding customers, or at least at handling his personal finances. He may even understand the world of business better if he has all had to run one, however small, at one time or another, and this introduces a variable we did not consider in our exercise, the age of worker. This could be matter for future work.

Even though the self-employed work showed time trends diverging from those of knowledge and innovation, entrepreneurship remains a relevant pillar of our argument. We support the idea that creating the conditions to foster self-employment is socially and economically relevant. A better understanding of how the self-employed organise their work and harness the benefits of knowledge and innovation while managing their job activities can offer insights to policymakers, employers and employees on the changing work domain. Though, the Covid-19 calamity may prevent some workers to run an own business, even for long time.

Our exercise has a main limit: the two-digit Ateco classification. This classification is raw; the analysis of one more digit classification might inform more than we did. Even this exercise showed problematic while computing variations because of low frequencies in some classes. This means that a more implied analysis should be realised *cum grano salis*. Finally, we did not assume a criterion variable and the presented analyses concerned just the non-hierarchical relationships among the selected indicators. Even this could be a further issue for future work.

References

Autor, D.H., Levy, F., Murnane, R.J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, **118(4)**, pp. 1279-1334.

Becattini, G. (1990). The Marshallian industrial district as a socio-economic concept, in F. Pyke, G. Becattini and W. Sengenberger (eds). *Industrial Districts and Inter-Firm Cooperation in Italy* (pp. 37-51). Geneva: IILS.

Bertamino, F., Bronzini, R, De Maggio, M., Revelli, D. (2014). Local policies for innovation: The case of technological districts in Italy (https://www.bancaditalia.it/pubblicazioni/altri-atti-convegni/2014-innovazione-italia/Bertamino-Bronzini-DeMaggio-Revelli.pdf).

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research,* **36(1)**, pp. 111–150.

Colombo M., Delmastro M. (2002). How effective are technology incubators? Evidence from Italy, *Research Policy*, **31**, pp. 1103-1122.

Etzkowitz H., Leydesdorff L. (1997). *Universities and the Global Knowledge Economy*. London: Pinter.

European Commission (2013). *Entrepreneurship 2020 Action Plan. Reigniting the Entrepreneurial Spirit in Europe*. Brussels: DG Internal Market, Industry, Entrepreneurship and SMEs.

Fabbris, L., Feltrin, P. (2021). Measuring the movement between employment and self-employment: a survey proposal. In: B. Bertaccini, L. Fabbris and A. Petrucci (eds.) *ASA 2021 Statistics and Information Systems for Policy Evaluation* (pp. 65-70), Firenze University Press, Florence (https://fupress.com/capitoli/measuring-the-movement-between-employment-and-self-employment-a-survey-proposal/6065).

Krugman, P. (2017). The gambler's ruin of small cities (Wonkish) (https://www.nytimes.com/2017/12/30/opinion/the-gamblers-ruin-of-small-cities-wonkish.html).

Liberati, D., Marinucci, M., Tanzi, G.M. (2013). Science and technology parks in Italy: Main features and analysis of their effects on hosted firms (https://www.bancaditalia.it/pubblicazioni/altri-atti-convegni/2014-innovazione-italia/Liberati-Marinucci-Tanzi.pdf).

Moretti, E. (2012). *The New Geography of Jobs*. Boston, New York: Houghton Mifflin Harcourt.

OECD (2019). *The Future of Work. OECD Employment Outlook 2019. Highlights* (https://oe.cd/employment-outlook).

Modelling the spatio-temporal dynamic of traffic flows with gravity models and mobile phone data

Maurizio Carpita, Rodolfo Metulini

1. Introduction

The analysis of origin-destination traffic flows may be useful in many contexts of application and have been commonly studied through the *Gravity Model* (Tinbergen, 1962). The popularity of Tinbergen's log-linear specification of the Gravity Model is due to its good performance in modelling international trade flows and to the strong theoretical foundations provided in papers such as Anderson (1979) and Anderson & Van Wincoop (2003). At the macro-level, this model states that the volume of trade between any two countries is proportional to the product of their gross domestic products (GDP) and a distance deterrence function, where distance is broadly construed to include all factors that might create trade resistance. The Gravity Model equation can be straightforward translated to micro-level flow data, such as, for example, passenger flows, simply by substituting trade flows with the total number of passenger flows from two cities, a measure of dimension of the city of origin and of the city of destination (such as their population) instead of GDP, and the geographical (or network) distance among the two cities in place of trade resistance.

Using data on the flow of mobile phone signals of TIM (*Telecom Italia Mobile*) users among different *census areas* (ACE of ISTAT, the *Italian National Statistical Institute*), recorded on hourly basis for six months, in this preliminary study we model such a flows in the *Mandolossa* to predict flows' intensity during flood episodes in the context of smart cities emergency management plans. Traffic flows data can be integrated to mobile phones densities and used to develop dynamic exposure to flood risk maps, as proposed in Balistrocchi et al. (2020). From a prevention perspective, this could make the identification of preferential traffic flows possible, thus evidencing potential risks during inundation onsets or emergency situations.

Whereas, as explained above, for the classical Gravity Model a traditional *static* mass explanatory variable is represented by GDP or by residential population (Kepaptsoglou et al., 2010) also thanks to the availabiity of a time series of data, we propose to use a most accurate set of explanatory variables in order to better account for the *dynamic* over the time. First, we employ a time-varying mass variable represented by the density of TIM users by area and by time period, which has been estimated from mobile phone data using the method proposed by Metulini & Carpita (2021) and adopted by Balistrocchi et al. (2020) to derive crowding maps for flood exposure. Second, a proper set of time effects is included. We show that the joint use of these two novel sets of explanatory features allow us to obtain a better linear fit of the Gravity Model and a better traffic flow prediction for the flood risk evaluation.

2. The mobile phone flows and the other datasets

The TIM mobile phone flows used in this study has been provided by Olivetti (*www.olivetti. com/en/iot-big-data*) and FasterNet (*www.fasternet.it*), for the development of the MoSoRe

Maurizio Carpita, University of Brescia, Italy, maurizio.carpita@unibs.it, 0000-0001-7998-5102
Rodolfo Metulini, University of Salerno, Italy, rmetulini@unisa.it, 0000-0002-9575-5136

Project 2020-2022 co-founded by Lombardy Region (*bit.ly/2Xh2Nfr*), and has been used at the DMS StatLab of the University of Brescia (*dms-statlab.unibs.it*).

The original data flows are square origin-destination (OD) matrices of dimension $N \times N$, where $N = 235$ represents the number of *census areas* or ACE (*Aree di Censimento*, using the standard definition of ISTAT) in the Province of Brescia, available at each hour's interval for six months from September 2020 to February 2021, so the length of the time series is 24×181 ($T = 4,344$). Furthermore, ISTAT provided the shape files for SCE (*Sezioni di Censimento*), with additional information about the belonging to their ACE and its area (*www.istat.it/it/archivio/104317*).

We restrict our attention to a particular subset of OD matrices, as the core of the analysis regards the area of the *Mandolossa*, which has been identified with 4 ACE (Brescia Mandolossa, Cellatica, Gussago and Rodengo Saiano) intersecting with the identified flooding-risk area (return period of 10 years), as reported in the left chart of Figure 1. We choose other 38 neighboring ACE aggregated as represented in the map, which fulfil the criteria of having a minimum (considering the four ACE of the Mandolossa) outflow of 10 in both three sample days chosen randomly. The total flows counted between the 4 Mandolossa's ACE and the 38 selected neighboring ACE counts for about the 84% of the total outflows from the Mandolossa's ACE.
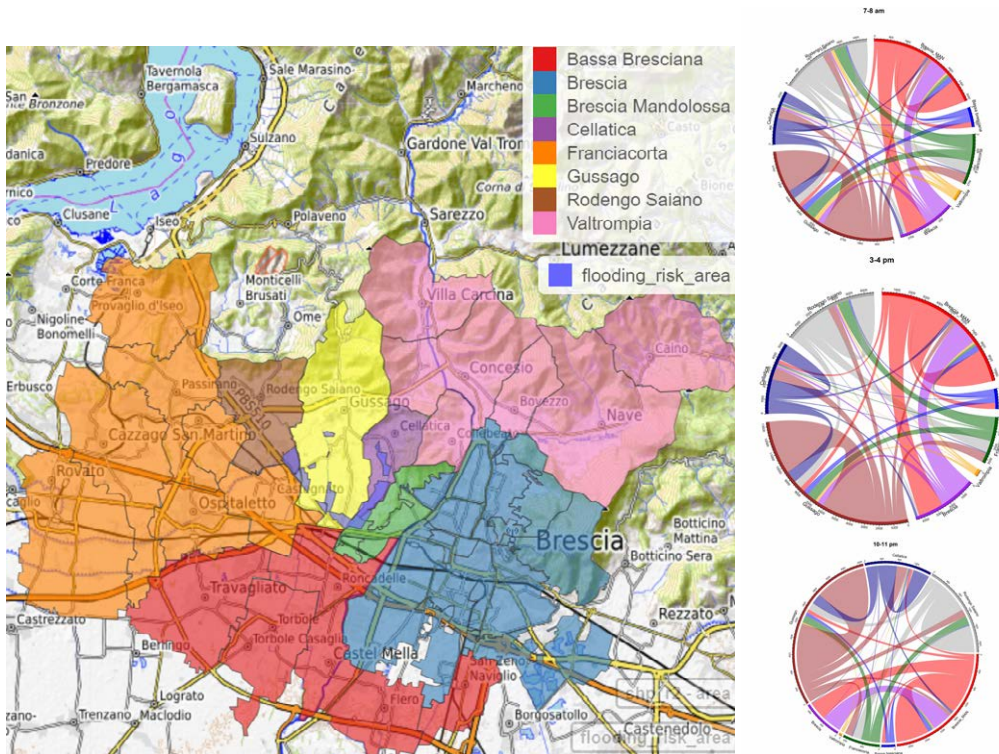
Figure 1: Map of flooding risk area, ACE in Mandolossa and neighboring (by macro-area) (left). Kriskograms of TIM flows between the eight macro-areas (right).

The three kriskograms in Figure 1 show flow between the 8 macro-areas of interest at three different hours (7-8 am, 3-4 pm, 9-10 pm): the diameter of the circles (proportional to the total flow) highlights that flows increase from morning to afternoon and decrease from afternoon to evening, and that for the four ACE of the Mandolossa flows are internal flows are very high. As show in Section 4, these evidences have suggested to introduce in the Gravity Model a parabolic

effect for *hour* and a dummy effect for three internal flows.

About Gravity Model's variables, we collect ISTAT data on residential *population* in each SCE (and, by aggregation, each ACE) at January 1st, 2016, and on the *distance* in km between the centroids of the 4 Mandolossa's ACE and the other 38 neighboring ACE. Furthermore, to extend the classical Gravity Model we have used the *mobile phone density* of TIM users, computed for each hour and ACE of interest, which can be interpreted as the average number of mobile phones simultaneously connected to the TIM network in that area in that time interval (Carpita & Simonetto, 2014). These data are created by Metulini & Carpita (2021) and used in Balistrocchi et al. (2020) for the analysis of the Mandolossa in the period 2014-2016. As the mobile phone densities for 2020 and 2021 are not yet available, we have used as proxy the data in the same month, hour and day of the week of 2015 (from September to December) and 2016 (for January and February).

3. The Gravity Model and its extension

The classical Gravity Model states that *flows* from origin i to destination j (F_{ij}) are proportional to *masses* of both origin and destination (M_i and M_j) and inversely proportional to *distance* between them (D_{ij}), where G and γ are positive constants:

$$F_{ij} \propto G \cdot \frac{M_i \cdot M_j}{D_{ij}^{\gamma}} \qquad (1)$$

Assuming masses as functions of *Populations* (P_i and P_j), the Gravity Model can be linearised using the logarithmic transformation of (1) and specified as a multiple linear regression model with a temporal dependence subscript t (in our case the hour), with random errors ϵ_{ijt} (LeSage & Pace, 2009):

$$\log(F_{ijt}) = \alpha + \beta_1 \cdot \log(P_i) + \beta_2 \cdot \log(P_j) - \gamma \cdot \log(D_{ij}) + \epsilon_{ijt} \qquad (2)$$

Model (2) can be extended introducing as other explanatory variables the dynamic masses (dependent from t) *mobile phone densities* (MP_{it} and MP_{jt}), the fixed effect for *Internal flows* (IF_{ij}) and a vector of pure *Time effects* ($\mathbf{TE}_t$), with parameters α, β_1, β_2, γ, δ_1, δ_2, ω and $\boldsymbol{\lambda}$ that must be estimated:

$$
\begin{aligned}
\log(F_{ijt}) &= \alpha + \beta_1 \cdot \log(P_i) + \beta_2 \cdot \log(P_j) - \gamma \cdot \log(D_{ij}) + \\
&\quad \delta_1 \cdot \log(MP_{it}) + \delta_2 \cdot \log(MP_{jt}) + \omega \cdot IF_{ij} + \boldsymbol{\lambda}^T \mathbf{TE}_t + v_{ijt} \qquad (3)
\end{aligned}
$$

It must be considered that this traditional log-linear specification of the Gravity Model along with Ordinary Least Squares (OLS) estimation method can be inappropriate when bilateral flows are frequently zero. Many studies estimate the log-linear model on samples of observations using the truncated OLS approach but, by disregarding pairs of observations that do not have a positive flows with each other can generate biased estimates (Helpman et al., 2008). Silva & Tenreyro (2006) have shown that log-linearisation of the Gravity Model leads to inconsistent estimates in the presence of heteroscedasticity in flows levels, and propose a Poisson specification along with the Poisson Pseudo Maximum Likelihood (PPML) estimator. However, when just interested on the flows between areas with positive flows, as in our explorative case study, it is possible to rely on OLS without any loss in estimation efficiency.

4. Application and preliminary results

The parameters of the classical Gravity Model (2) and its extension (3) presented in Section 3 have been estimated using the standard OLS method using data described in Section 2. For

this preliminary study, a sample of flows of 6 hours (7,10,13,15,18,21) and 4 days of the week (Monday, Wednesday, Thursday and Saturday) for the six months from September 2020 to February 2021 has been extracted from the 4 Mandolossa's ACE and the 38 neighboring ACE. Then, this sample of 6,912 observations has been randomly partitioned in *training set* (6,000 observations) used for estimation and *test set* (912 observations) used to evaluate prediction performance.

To assess the goodness of fit of the four models considered in this preliminary analysis, Residual standard error and adjusted R^2 have been used, whereas the AIC (Akaike's information criterion) for the training set and the correlation between observed and predicted flows ($Cor(Y,\hat{Y})$) for the test set have been used to assess prediction performance. The F tests of significance for the parameters of the considered (*full*) model and for the model included (*nested*) in the considered model are reported too.

Table 1 shows preliminary results for the four Gravity Models described in the previous section. MOD1, the classical Gravity Model in formula (1) with only *Population* and *Distance* as explanatory variables, has statistical significance (t and F tests have zero p-value), but rather low goodness of fit (adjusted R^2 is 34.5%) and prediction performance (for the test set, correlation between observed Y and estimated $\hat{Y}$ flows is 0.595); as expected, the estimated effects on the *Flows* are positive for *Population* and negative for *Distance*. MOD2, that includes *Mobile phone density* as explanatory variables, has statistical significance (F test reject the nested model MOD1), but doesn't improve substantially the fit (adjusted R^2 is 34.9%) and has the same prediction performance of MOD1 (but AIC is a little lower and $Cor(Y,\hat{Y})$ for the test set is 0.594). When the dummy for the three *Internal flows* is added to the model (see the end of Section 2), results noticeably improve: for MOD3 the F test reject the nested model MOD2, the fit gets better (adjusted R^2 is 53.1%) and prediction performance increases (AIC decreases a lot and $Cor(Y,\hat{Y})$ for the test set is 0.741); note that the presence in the model of *Internal flows* strongly reduce the effect of *Distance* (from -0.186 to -0.06) and slightly increase the effects of the two *Mobile phone density* on *Flows*. Finally, the introduction of the temporal effects as in MOD4 further improves the results: the F test reject the nested model MOD3, adjusted R^2 is 62.7%, AIC decreases further and $Cor(Y,\hat{Y})$ for the test set is 0.808. *Hour* has the expected significant and parabolic effect on *Flows* (increasing from morning to afternoon and decreasing from afternoon to evening), *Day of the week* has a significant and negative effect for Saturday and *Month* has significant and negative seasonal effects, i.e. flows are lower in Autumn and Winter with respect to September: this rather unexpected effect may have been caused by the limitations caused by the COVID19 pandemic that began in October 2020. Note that introducing the time effects doesn't change substantially the parameter estimates for the other regressors respect to MOD3.

5. Concluding remarks

Using data on the flow of mobile phone signals of TIM users among different ISTAT census areas the classical Gravity Model and some its extensions have been preliminarily adopted to study dynamic of such flows over the time in the *Mandolossa*, an area at the western outskirts of Brescia in northern Italy, with the final aim of predicting the traffic flow during flood episodes.

In addition to the usual population and distance regressors, the joint use as explanatory variables in the model of time-varying mass variable represented by the density of TIM users by area and by time period and a proper set of temporal effects allow us to obtain a better linear fitting with respect to the classical Gravity Model, and a better traffic flow prediction for the flood risk evaluation. These preliminary results are promising, but some in-depth analyses have yet to be carried out. As explained at the end of Section 3, it will be important to evaluate

Table 1: Preliminary results of four Gravity Models for the Mandolossa flows

Regressors	MOD1	MOD2	MOD3	MOD4
Population origin	1.023***	0.891***	0.730***	0.694***
Population destination	1.027***	0.851***	0.706***	0.671***
Distance in km	−0.186***	−0.186***	−0.060***	−0.060***
Mobile phone density origin		0.155**	0.231***	0.279***
Mobile phone density destination		0.196***	0.261***	0.301***
Internal flows			1.576***	1.578***
Hour				0.559***
Hour^2				−0.019***
Day of the week (reference: *Monday*)				
Wednesday				0.074*
Thursday				0.061·
Saturday				−0.269***
Month (reference: *September*)				
October				−0.088*
November				−0.273***
December				−0.350***
January				−0.291***
February				−0.250***
Constant	−18.992***	−19.156***	−18.306***	−21.804***
Residual standard error	1.138	1.134	0.963	0.858
Degrees of freedom	5,996	5,994	5,993	5,983
Adjusted R^2	0.345	0.349	0.531	0.627
F test full model	1,053***	644***	1,133***	632***
F test nested model		19.752***	2,329***	156***
AIC training set (6,000 obs.)	18,583	18,549	16,581	15,211
Cor(Y,Ŷ) test set (912 obs.)	0.595	0.594	0.741	0.808

Notes: For all the models, the variables flows, population, distance and mobile phone densities are in logarithms. Parameter estimates have been obtained using the standard OLS method.
Significance codes for t and F tests: . $p < 0.1$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

the possibilities offered by the most appropriate estimation methods; moreover, the actual predictive capacity of the model for the purposes of the MoSoRe Project will have to be further investigated.

Finally, we are also evaluating to introduce in the Gravity Model other non-standard explanatory variables, related to to the number and the type of streets, the number of offices, restaurants or cinemas, which may be retrieved from `OpenStreetMap`, would allow to better characterize the areas of interest and further improve the model performance.

Future use of 5G and GPS technologies will facilitate the real-time assessments of the spatial distribution of people: with an early-warining system, alternative safe pathways could be identified and communicated to exposed people in order to facilitate their evacuation.

Acknowledgments

References

Anderson, J.E. (1979). A theoretical foundation for the gravity equation. *The American Economic Review,* **69**(1), pp. 106–116.

Anderson, J. E., Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review,* **93**(1), pp. 170–192.

Balistrocchi, M., Metulini, R, Carpita, M., Ranzi, R. (2020). Dynamic maps of human exposure to floods based on mobile phone data *Natural Hazards and Earth System Sciences,* **20**, pp. 3485–3500.

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

Helpman, E., Melitz, M., Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics,* **123**(2), pp. 441–487.

Kepaptsoglou, K., Karlaftis, M.G., Tsamboulas, D. (2010). The gravity model specification for modeling international trade flows and free trade agreement effects: a 10-year review of empirical studies *The Open Economics Journal,* **3**(1), pp. 1–13.

LeSage J., Pace R.K. (2009). Introduction to Spatial Econometrics. *Chapman & Hall/CRC, New York (NY)*

Metulini, R, Carpita, M. (2021). A spatio-temporal indicator for city users based on mobile phone signals and administrative data *Social Indicators Research,* **156**(2-3), pp. 761–781.

Silva, J.S., Tereyro, S. (2006). The log of gravity. *The Review of Economics and Statistics,* **88**(4), pp. 641–658.

Tinbergen, J. (1962). Shaping the world economy; suggestions for an international economic policy *Twentieth Century Fund, New York (NY).*

The effectiveness of marketing tools in a consumer goods market in Italy during the Great Recession (2010-2015)

Giorgio Tassinari, Demetrio Panarello

1. Introduction

During the Great Recession of 2008-2015, household consumption decreased in all EU countries (Streeck, 2016; Tooze, 2018). In Italy, for instance, at constant 2015 prices, the value of private national consumption decreased by 4.6% (Istat, 2021). It must be borne in mind that the considered period is very diverse. The Great Recession 2008-2015, as is well known, is W-shaped, with a first lower turning point in 2009 (financial crisis that spread from the USA to all high-income countries) and a second lower turning point in 2013 (sovereign debt crisis in EU countries). Therefore, this cyclical profile sees, within the considered period, the alternation of depressive and moderately expansive phases.

Faced with this situation, undertakings active in the markets for consumer goods with a high purchase intensity mostly reacted by means of strategies based on price reductions and promotions; conversely, most enterprises reduced their promotional advertising investments (Freo et al., 2020). Companies' marketing strategies in times of economic recession have been the subject of in-depth studies (Deleersnyder et al., 2009; Van Heerde et al., 2013). Most of the evidence presented in the literature confirms that the adopted marketing strategies vary throughout the different phases of the economic cycle (Lamey et al., 2012; Van Heerde et al., 2013). In the wake of recession, households reduce consumer spending, for instance by switching from national brands to private labels brands; at the same time, companies react by changing the marketing mix, reducing regular prices, making greater use of promotions, and cutting advertising investments. Van Heerde et al. (2013) shows that price elasticity increases during the downward phases of the economic cycle, whereas advertising elasticity increases during the expansionary phases. Besides, other studies find that increasing or maintaining advertising investments has a positive impact on brand performance during recessions (Deleersnyder et al., 2009; Kashmiri and Mahajan, 2014).

The subject of our analysis is the Italian market of tea-based beverages in the period 2010-2015, of which the marketing tools' effectiveness and the competitive structure are examined, in order to ascertain the intensity and extent of price-based strategies compared to those that leverage advertising investments.

Based on the literature, we expect the price elasticity of each brand to be greater than the elasticity to advertising. Since we are dealing with a stationary market, we employ a market share model, making use of the methodology described by Cooper and Nakanishi (1988). This approach allows us not only to measure the impact of marketing mix on each brand's market share but also to identify the competitive structure.

2. Data and preliminary analyses

The present study analyses the competitive situation and the effectiveness of price maneuvers and advertising investments in terms of increasing market shares.

We make use of monthly observations obtained by aggregating IRI Infoscan weekly surveys concerning Italian hypermarkets and supermarkets in the period from November 2010 to October 2015. For each brand, the sales in value and volume, the price per liter, the possible presence of price promotions, and the weighted distribution are known. The advertising carried out by each brand, sourced from Nielsen, is expressed in terms of Gross Rating Points referring to all mass

communication channels.

The brands on the market are numerous and heterogeneous. Price differences are considerable between one brand and another, ranging from 0.56 to 1.65 euros per liter. In order not to saturate the information capacity of the model in relation to the available data, requiring the estimation of an excessive number of parameters, we separately consider five different brands, so that they adequately represent the heterogeneity of competitors on the market, both in terms of price and market share. Altogether, such five brands cover about three quarters of the category's volume sales.

Since tea-based beverages are characterized by a low emotional involvement, the differentiation between brands is mostly based on tangible attributes and the product's price is seen as a mirror of its intrinsic quality. This is also reflected in the ratio of market share to share of voice, which is high for lower-priced brands with a higher market share. The choice of the product is made directly at the point of sale and is based on habits and routines. The process of purchasing tea-based beverages, a "luxury" good, follows the do-learn-feel scheme, whereby consumers know and are able to evaluate the product only after making the purchase; additional purchases may only happen after they learn about the product's actual quality and feel satisfied. Therefore, for products of this kind, the advertising investment is primarily aimed at ensuring that consumers recognize the product and are induced to carry out an initial test; then, only after repeated purchases, the goal turns into strengthening users' loyalty.

As could be expected, the advertising investments made by the different brands during the analyzed period show a very marked variability, in relation to the different time intervals in which the companies carried out their advertising campaigns. Before proceeding to the estimation of the attraction model, we verify that the five brands' market shares do not present unit roots (stochastic non-stationarity), by performing a Dickey-Fuller test augmented by means of seasonal dummies and deterministic trend (Table 1) on the log-centered market shares.

Table 1 – Dickey-Fuller unit root test.

Brand	Test	p-value
Ferrero	-3.88	0.019
San Benedetto	-4.76	0.002
Nestlé	-3.90	0.018
PepsiCo	-3.70	0.030
Coca-Cola	-4.95	0.001

Note: The null hypothesis is the presence of unit root.

The trend of volume sales has a purely seasonal pattern. The total volume of sales in the category is not markedly affected by the economic crisis. We remark the existence of a consumer segment, not insignificant in its size, which buys the category even in the winter months. Category-level advertising investments follow the seasonal pattern: each year, they begin in the spring months, reach their peak at the beginning of the summer, and then gradually decrease until almost zero in winter.

The average price of the five considered brands increases over the period. Every year, the price decreases in the spring-summer months, due to more frequent price promotions, and increases in autumn and winter, with the average maximum price of € 1.06 recorded in November 2014. In general, prices are gradually increasing: in the considered period, the price went from € 0.89 to € 0.99 per liter, with an increase of more than 11%, against an increase in the general level of consumer prices in the same period of 7.5% and, for the general category of non-alcoholic beverages, of 8% (Istat, 2021). Overall, the fairly modest decrease in volume sales is therefore offset by the increase in unit prices. These characteristics make, in our opinion, the study of this category particularly interesting and peculiar.

3. Attraction model estimation

For the estimation, the method of ordinary least squares individually applied to each equation was used in the first place; this procedure is equivalent to the Zellner estimator when the same regressors appear in each equation of the system (Cooper and Nakanishi, 1988). Prices and advertising investments of each brand were employed as independent variables in each equation, in addition to the weighted distribution, while the dependent variables are the log-centered market shares. Both the MCI and MNL models (Cooper and Nakanishi, 1988) are theoretically plausible; therefore, it is necessary to resort to empirical criteria for their selection. We opted for the use of the MNL model, as advertising investments are absent for many brands and in different periods (flight strategy).

We proceeded as follows: a) formal seasonality test; b) estimation of the complete MNL model by means of the OLS method after performing the data log-centering; c) analysis of the presence of residual autocorrelation; d) estimation of the final model through the Zellner's SUR method (Cooper and Nakanishi, 1988).

For what concerns seasonality, we performed a formal test through an OLS regression on seasonal dummies, not reported for the sake of brevity. In the equations for each brand, according to the previous results, the seasonal dummies and the deterministic trend are included.

The choice between the use of either a static or a dynamic model was solved by performing a Durbin-Watson autocorrelation test (Table 2) on the residuals from the OLS estimates, the results of which led us to prefer a dynamic formulation of the error correction type, as in three equations out of five we come to reject the hypothesis that residuals are white noise (at a 1% significance level).

Table 2 – Durbin-Watson autocorrelation test.

Brand	Test	p-value
Ferrero	1.93	0.072
San Benedetto	1.35	9.356e-005
Nestlé	1.67	0.007
PepsiCo	1.23	1.182e-005
Coca-Cola	1.75	0.015

The presence of seasonality in market shares is confirmed by the OLS estimates. From brand to brand, the seasonality pattern presents different shapes and the deterministic trend shows different slopes.

The presence of non-significant coefficients in the OLS estimates led us to estimate the system of equations through the SUR method by setting the values of the barely significant parameters to zero. The equations also include the seasonal dummies, the deterministic trend, and each brand's log-centered market share delayed by one lag, in order to consider the dynamic aspect highlighted by the Durbin-Watson test. The results are shown in Table 3. The R^2 coefficient weighted for the entire system is equal to 0.947.

The parameters concerning the influence of a brand's price on its market share present a negative sign, while those relating to competitors' prices are generally positive. Most coefficients regarding advertising investments are not significant, while the distribution confirms itself as an important marketing tool.

To verify the appropriateness of the restrictions imposed on the parameters of the set of equations, we made use of the F test (not reported for the sake of brevity), which confirms that we can rely on the restricted model.

Brands/Variables	Ferrero	San Benedetto	Nestlé	PepsiCo	Coca-Cola
Constant	2.1515***	6.9337***	-0.6655	0.9299***	-9.3094
Price – Ferrero	-0.1323*	-4.6418	0.3431**	-0.2959*	-
Price – San Benedetto	-	2.1982***	0.9738***	-	0.8641***
Price – Nestlé	0.7355***	0.6966***	-2.9534***	1.2991***	-
Price – PepsiCo	1.0016***	0.8764***	-2.1558	-1.9880***	-
Price – Coca-Cola	-	0.3737***	0.6692***	0.7689***	-1.6465***
Group – Ferrero	-	-	-	-3.4374	-
Group – San Benedetto	-1.954e-05*	8.6315	-	-	-
Group – Nestlé	2.2499**	-	-	-	-
Group – PepsiCo	-	3.3450	-	-	-8.9940e-0**
Group – Coca-Cola	-	-0.0002***	-	0.0003**	-
Distrib. – Ferrero	-	-0.0512**	-9.6300e-13	7.4800e-13	0.0615**
Distrib. – San Benedetto	-0.0296***	-	-	1.6750	0.0245**
Distrib. – Nestlé	0.0116***	-0.0103***	0.0127***	-	-0.0164***
Distrib. – PepsiCo	-	0.0024*	-0.0021	-	-
Distrib. – Coca-Cola	-0.0082***	-0.0050***	-	-0.0102***	0.0247***
Time	-	0.0073***	-0.0023	-0.0135***	0.0093**
Spring	-	0.0699***	-0.0191	-	-
Summer	-	0.0646***	0.0965***	-0.1005***	-0.0709
Autumn	-	0.0457***	-0.0084**	0.0253	-
MS(-1)	0.1932***	0.0606	0.0685	0.1039***	0.0959*
R2corr	0.96	0.97	0.93	0.89	0.86

Notes: p-value < 0.01 = ***; p-value < 0.05 = **; p-value < 0.10 = *.

4. Elasticity of shares with respect to marketing tools and basic market shares

The estimated parameters of the restricted model allow us to determine the cross-elasticity coefficients, according to the following formula (Cooper and Nakanishi, 1988):

$$\varepsilon_{S_i ;x_{kj}} = (\beta^*_{kij} - \sum_{h=1}^{m} S_h \beta^*_{khj})X_{kjt}$$

where $\varepsilon_{S_i ;x_{kj}}$ is the elasticity of the market share of brand i with respect to the k marketing tool of brand j; β_{khj} are the estimated coefficients; and S_h are the average market shares.

Examined by rows, the elasticity matrices provide information on the effects of marketing variables (own and competitors') on the share of a brand, while by columns they indicate the effects produced by a specific brand's marketing tool on its own share and on that of competitors; in essence, they provide useful information on the competitive situation in the examined market. In the following tables, we present the elasticities of market shares to the various brands' prices, advertising investments and weighted distribution.

Table 4 − Elasticity to prices (authors' elaboration, Italy, Nov. 2010 – Oct. 2015).

Market share/Price	Ferrero	San Benedetto	Nestlé	PepsiCo	Coca-Cola
Ferrero	-0.1947	0.4587	0.3930	0.3401	-0.1719
San Benedetto	0.0228	-1.2112	0.3673	0.2402	0.1132
Nestlé	0.5869	1.1985	-2.0424	-0.4578	0.3387
PepsiCo	-0.4637	0.4587	0.7650	-2.0414	0.4147
Coca-Cola	0.0228	1.1151	-0.0926	-0.4578	-1.4280

By observing the elasticities to the average prices in the whole period (Table 4), it is immediately clear that those relating to each brand's own price are all negative. Therefore, a price increase manifests itself in a more or less sharp decrease in a brand's own market share. Moreover, a modest proportion of cross-price elasticities shows a different sign than expected.

Observing the table by columns, we find the elasticities of market shares with respect to the price of the examined brand: it is easy to notice that they differ much, both from one brand to another (which shows that price variations have effects of varying strength depending on the brand) and within the same column. Indeed, this is a clear sign of the existence of strong competitive asymmetries in the analyzed market.

Let us now examine the elasticities by brand, which present some noteworthy characteristics. Most of the values in the matrix are below one. Higher values mean more price competition between brands. Analyzing the values by rows, it emerges that Ferrero, the highest-priced brand, is characterized by a low direct elasticity in absolute value and is relatively isolated from the price maneuvers of other brands, thus confirming its role as the market leader.

Moving on to the column relating to San Benedetto, which is the most important follower, we can see that its price has a relevant impact on the other medium- and low-priced brands.

Briefly, the examination of the cross-price elasticity matrix, while confirming the importance of competitive asymmetries, underlines that the two main brands are rather isolated from each other as regards the effect of prices on market shares.

Table 5 − Elasticity to advertising investments (authors' elaboration, Italy, Nov. 2010 – Oct. 2015).

Market share/Investments	Ferrero	San Benedetto	Nestlé	PepsiCo	Coca-Cola
Ferrero	0.0001	-0.0046	0.0020	-0.0002	0.0010
San Benedetto	0.0001	0.0039	-0.0012	0.0004	-0.0025
Nestlé	0.0001	0.0013	-0.0012	-0.0002	0.0010
PepsiCo	-0.0022	0.0013	-0.0012	-0.0002	0.0051
Coca-Cola	0.0001	0.0013	-0.0012	-0.0017	0.0010

The coefficients regarding the elasticity of market shares to advertising investments (Table 5) are all close to zero. Therefore, such investments almost never seem to produce any noteworthy variation in either own or competitors' market shares. This is very much in tune with the coefficients relating to advertising investments resulting from the MNL models estimated through the SUR method, which were mostly non-significant.

Table 6 − Elasticity to the weighted distribution (authors' elaboration, Italy, Nov. 2010 – Oct. 2015).

Market share/Distribution	Ferrero	San Benedetto	Nestlé	PepsiCo	Coca-Cola
Ferrero	1.5898	-1.7760	0.7840	-0.0347	-0.2390
San Benedetto	-3.4685	0.9508	-1.0237	0.1233	-0.0107
Nestlé	1.5898	0.9508	0.8755	-0.1741	0.2913
PepsiCo	1.5898	0.9508	-0.1717	-0.0347	-0.3233
Coca-Cola	7.6600	3.2117	-1.5196	-0.0347	1.7784

Moving on to Table 6, we can notice that most direct elasticities are positive, while cross elasticities (which should be negative) often show different signs than expected.

An overall evaluation of the effectiveness of each brand's marketing strategy can be carried out by comparing the basic and average market shares for the whole period.

Table 7 − Average and basic market shares (authors' elaboration, Italy, Nov. 2010 – Oct. 2015).

Brand	Average market share	Basic market share
Ferrero	0.3805	0.0083
San Benedetto	0.3603	0.9888
Nestlé	0.1593	0.0005
PepsiCo	0.0615	0.0024
Coca-Cola	0.0384	8.7291e-08

The basic market share defines the share that each brand would have if all brands had the same coefficients concerning marketing tools' effectiveness and if the intensity of their use were the same for each brand. In brief, it represents the intrinsic attractiveness of each brand.

Basic market shares are very different from the average shares observed in the considered period (Table 7). Ferrero, the brand that invests in advertising the most, manages to obtain a much higher average market share than its basic market share, while a significant erosion of the basic share can be pointed out for San Benedetto.

5. Conclusive remarks

From the combined analysis of the effect on market shares of price, advertising investments and weighted distribution, it can be seen that the availability of the brands within a point of sale also plays a relevant role in influencing own and competitors' market shares.

Price competition does not seem to have considerable direct effects (cross elasticities lower than one). With this in mind, advertising investments, despite not having a major direct effect on market shares, are intended to increase brand awareness in consumers' minds, stimulating its recognition at the time of purchase, while price is the variable that ultimately determines the decision to buy the item. Therefore, the key elements in determining market shares in the examined category are the elasticity to price – especially the direct one – and to the weighted distribution, which is in line with the characteristics of this category of products (i.e., high purchase frequency and weak emotional involvement).

In conclusion, it should be remarked that the analyzed category recorded an excellent performance in the considered period. It is self-evident that our results ought to be expanded by considering a wider range of categories, in order to be able to draw more general conclusions.

References

Cooper, L.G., Nakanishi, M. (1988). *Market Share Analysis: Evaluating Competitive Marketing Effectiveness*. Kluwer Academic Publishing, Boston, USA.

Deleersnyder, B., Dekimpe, M.G., Steenkamp, J.E.M., Leeflang, P.S. (2009). The Role of National Culture in Advertising's Sensitivity to Business Cycles: An Investigation across Continents. *Journal of Marketing Research*, **46**(5), pp. 623-636.

Freo, M., Tassinari, G., Tono, A.G. (2020). Advertising Effectiveness during the Great Recession: the Case Study of Three FMCG Italian Food Categories. *WSEAS Transactions on Business and Economics*, **17**, pp. 368-373.

Istat (2021). Official Website. www.istat.it.

Kashmiri, S., Mahajan, V. (2014). Beating the recession blues: Exploring the link between family ownership, strategic marketing behavior and firm performance during recessions. *International Journal of Research in Marketing*, **31**(1), pp. 78-93.

Lamey, L., Deleersnyder, B., Steenkamp, J.E.M., Dekimpe, M.G. (2012). The effect of business-cycle fluctuations on private-label share: what has marketing conduct got to do with it?. *Journal of Marketing*, **76**(1), pp. 1-19.

Streeck, W. (2016). *How Will Capitalism End? Essays on a Failing System*. Verso, New York, USA.

Tooze, A. (2018). *Crashed: How a Decade of Financial Crises Changed the World*. Penguin, London, UK.

Van Heerde, H.J., Gijsenberg, M.J., Dekimpe, M.G., Steenkamp, J.E.M. (2013). Price and Advertising Effectiveness over the Business Cycle. *Journal of Marketing Research*, **50**(2), pp. 177-193.

The role of the extra-man play actions in elite water polo matches: which elements lead to a good shot?

Alessandro Lubisco

1. Introduction

Many studies on team sports seek to identify which elements set the winning team apart from the losing one and which elements of play lead to victory (Lupo *et al.*, 2014). This is one of the reasons why match analysis has developed a great deal over recent years in many disciplines, including water polo.

In water polo, two teams, each of six outfield players and a goalkeeper, compete for four quarters of 8 minutes' real play in a playing area of 30x20m. All players are involved in both attack and defence. Generally speaking, the attacking team places one player at the centre (position 6, the centre forward) and arranges the others in a semicircle (Fig. 1a). Defence has a number of strategies which range from pressing to one of various types of defence zones. Each team has 30 seconds to complete a play action. If the attacking team still has possession of the ball after a shot, they then have at least 20 available seconds. This is a similar system to that adopted in basketball: 24 seconds for an action and at least 14 seconds following a shot. The team that scores the most goals wins. The match may also finish with a tie and, in direct elimination matches, a penalty shootout is used to determine the winner.

A frequent situation thought to be very significant to the final result of a match (Takagi *et al.*, 2005) is one of numerical superiority usually called extra-man, man-up or a 6-on-5 situation (XM). This occurs when a player is temporarily excluded following a major foul (FINA, 2020) and is sent out of play for 20 seconds.

Coaches dedicate a lot of time to training their team to attack and defend in an XM situation. Briefly, players in attack line up along two lines: two players at 5/6 metres from the goal, each in line with the goal posts, and the others on the two-metre line, a sort of off-side line. So, there are two players in line with the posts and two on the flanks. This kind of attack is known as a 4-2 and is used by most teams. There is an initial attack formation called a 3-3, where 3 players are positioned on the external line and 3 others on the two-meter line. This formation is less frequent.

In reference to the attack formation 4-2, the positions of the players are numbered from 1 to 6 clockwise as you face the goal, starting from the player on the right on the 2-metre line (Fig. 1b).

The defending team generally places three players on the 2-metre line who 'jump' sideways to mark the two opponents either side of each defender and stop the wings, at 2 metres (1 and 4), from scoring by raising their arms. The two external players also move backwards and forwards to cover the goal area with raised arms, stopping the wings from scoring (2 and 3) and intercepting any passes toward the central players on the 2-metre line (5 and 6) as shown by the arrows in Fig. 1b. This formation is called a 2-3 defence.

As there are 20 seconds available for concluding the action (with some exceptions depending on the area where the exclusion takes place), the attacking team, with a series of passes and player movements must quickly try to disrupt the defence and enable a shot. For those in defence these 20 seconds are never-ending because the physical effort involved in mounting an aggressive defence able to prevent the attack from scoring is huge. This is why, for the attack, a coach will aim to improve the players' ability to move the ball quickly from one position to another without leaving the defence enough time to take up positions. And defence has to work on coordination of movement between players so that the attack finds it difficult to score too easily.

It has to be said that nothing is easy in water polo, particularly at high levels. Even in the so-called 1 against 0 situation, when one attacker is alone in front of the goalkeeper, actually scoring

is by no means a foregone conclusion.

This paper investigates the issue of XM actions in detail. More specifically, the study analyses data from a recent European men's water polo championships, whose aim is to identify whether XM actions have any elements that lead to a good shot, meaning a ball in the goal even if it is saved.

Section 2 describes a preliminary analysis of the 48 matches in this championship. The results of the analysis carried out into the XM actions are presented in Section 3. The final section discusses some concluding remarks.

Fig. 1a: Waterpolo player positions
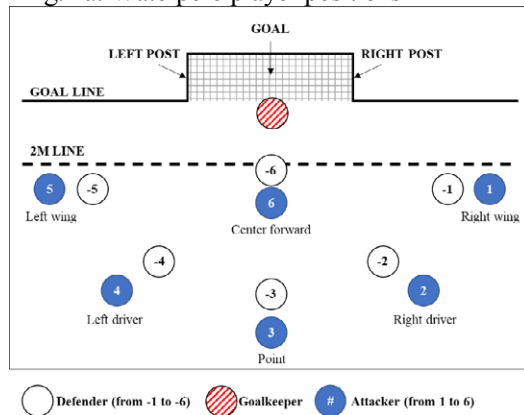
Fig. 1b: Attacker positions in 6-on-5 situation
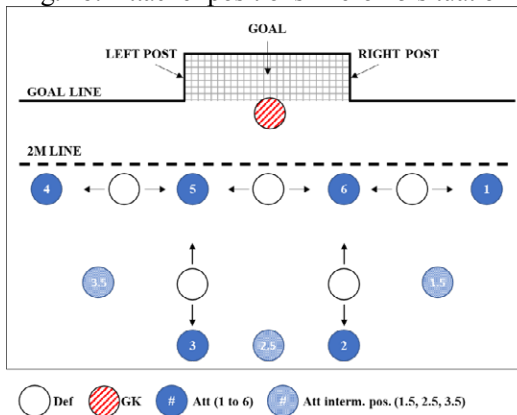
2. Preliminary analysis

The forty-eight matches of the 34th European men's water polo championships held in Budapest in January 2020 were taken into consideration.

To verify the importance of XM actions in water polo matches, a preliminary analysis on all the matches was carried out starting from the information obtained from the play-to-play tables available on the championship site[1].

For each match, the following variables were taken into consideration: outcome, number of possessions, total number of actions (i.e. possessions that end with a shot, an exclusion or a penalty), number of 6-on-6 actions (EA), number of 6-on-5 actions (XM), total number of goals, number of goals scored in 6-on-6 actions (EG) and number of goals scored in 6-on-5 actions (XG).

After regular time, three of the 48 matches ended with a draw and the winner was decided with a penalty shootout: Spain defeated Hungary in the preliminary phase and Serbia in the quarter-finals, but lost the match for 1st place against Hungary. The definition of the winning/losing team in this case corresponds to the result of the match after penalties.

In the tournament, the number of possessions per team averaged 37.8 (SD=3.1) per match. With regard to actions, there were significant differences between winning (mean=30.8; SD=4.1) and losing (mean=26.0; SD=4.1) teams, as is the case for the number of goals per match, in both even (6-on-6) and in man-up (6-on-5) situations (Table 1).

Overall, most goals resulted from even-player actions (6.6 per match against 3.7 in man-up). This is due to the fact that 72% of actions were played in this situation.

Unquestionably, the probability that an even-player action concludes with a goal is lower than the same probability for man-up actions. As you can see in Table 1, 31.2% of even-player actions concludes with a goal, whilst with an extra man the percentage is 46.6%.

Differentiating between winning and losing teams, the percentages become 38.2% and 24.2% respectively in 6-on-6 actions (EG/EA). That is to say, the winning teams score a goal for approximately every three play actions with even players, whilst the losers score one every four.

When considering a numerically superior formation, the winning teams score on average one goal every two play actions (51.9%), whereas losing teams score 41.3% per action (XG/XM). There is, therefore, a distinctly higher performance in situations of numerical superiority, which is emphasised when differentiating between winning and losing teams.

So, when the opponent is given an exclusion it provides an opportunity to play an action with a higher probability of scoring a goal. This is why the objective for most of the game is to give the ball to the centre forward in order to obtain an exclusion.

Table 1: European men's water polo matches. Means, standard deviations, differences between winning and losing teams and p-value of ANOVA F test (significant differences in bold)

	Total (n=96)				Winning teams (n=48)				Losing teams (n=48)				F
	N	%	Mean	SD	n	%	Mean	SD	n	%	Mean	SD	pval
Possessions	3624		37.8	3.1	1838		38.3	3.5	1786		37.2	2.6	0.089
Actions	2726	100.0	28.4	4.7	1479	100.0	**30.8**	4.1	1247	100.0	**26.0**	4.1	**0.000**
6-on-6 (EA)	1964	72.0	20.5	4.5	1091	73.8	**22.7**	4.1	873	70.0	**18.2**	3.8	**0.000**
6-on-5 (XM)	762	28.0	7.9	3.1	388	26.2	8.1	2.7	374	30.0	7.8	3.4	0.642
Goals	988	100.0	10.3	4.6	628	100.0	**13.1**	4.4	360	100.0	**7.5**	2.6	**0.000**
6-on-6 (EG)	633	64.1	6.6	3.8	427	68.0	**8.9**	3.8	206	57.2	**4.3**	1.8	**0.000**
6-on-5 (XG)	355	35.9	3.7	2.1	201	32.0	**4.2**	2.2	154	42.8	**3.2**	1.9	**0.019**
EG/EA			0.312	0.134			**0.382**	0.120			**0.242**	0.109	**0.000**
XG/XM			0.466	0.221			**0.519**	0.223			**0.413**	0.208	**0.018**

3. The extra-man action analysis

In the previous section, analysis underlined the importance of a man-up situation.

The decision was made to proceed with an analysis of all man-up actions in all 48 matches in order to understand if any characteristics of numerically superior play actions can be identified which increase the probability of scoring a goal or at least of making a good throw at the goal.

For this purpose, data from official FINA[2] video footage of the European men's water polo championships were collected and analysed. The dataset is formed of 979 extra-man plays. This number is higher than the total number of play actions in Table 1 (762); the reason being that numerically superior play actions were also considered when the excluded player had already returned to play, but had yet to reach his position in defence.

Focusing attention on the characteristics of an action that depend on the way the team plays, for each of the chosen actions, regardless of the outcome, the following variables were considered: number of passes, action duration (in seconds), sequence of passes (positions in Fig. 1b), time out call (Yes/No).

The following variable were then defined:
- GoodShot: 1 means an action which ended with a goal or a shot saved by the goalkeeper; otherwise 0.
- ZoneCat. This is defined by three categories considering the last zone in a sequence of passes, meaning, for example, the origin of a shot: "Lateral"=Zone 1 or 4, "Posts"=Zone 5 or 6, "External"=Other zone.
- DurationCat. This is defined by three categories considering the duration in seconds of an action: "Less than 11", "From 11 to 15", "More than 15".
- NPassesCat. This is defined by three categories considering the number of passes of an action: "Less than 5", "From 5 to 7", "More than 7".
- Rotation: 1 means an action where at least one player moves to an intermediate position (1.5, 2.5 or 3.5) or an action where players in position 1 or 4 move within the 2-meter line

---

[2] Dailymotion platform

with the ball; otherwise 0.
- LongPasses: 1 means an action with passes between nonadjacent zones (from 1 to 3, from 1 to 4, from 2 to 4 and vice versa); otherwise 0.

The reason behind using the GoodShot variable instead of Goal relates to the fact that a well-played action may not lead to a score because the goalkeeper performs an exceptional save. It can still be defined as a well-played action that satisfies the coach, even though a score would have been preferable.

Pearson's Chi Square test showed a significant association between the occurrence of a good shot and some of the variables considered (in bold), as shown in Table 2.

Table 2: Pearson's Chi Square test results for the selected variables and GoodShot

Variable	p-value
ZoneCat	**0.010**
DurationCat	**0.000**
NPassesCat	**0.000**
Rotation	0.065
LongPasses	**0.028**
TimeOut	0.054

In order to illustrate the effect of significant variables on the probability of performing a good shot, a logistic regression model was estimated. As the number of passes is strongly correlated with the duration of the action (r=0.774), NPassesCat was not included in the model. These results are shown in Table 3.

Table 3: Logistic regression model for GoodShot

	B	S.E.	Wald	df	P-Value	Exp(B)
Duration "From 11 to 15"			15.716	2	0.000	
Duration "Less than 11"	0.352	0.157	5.045	1	0.025	1.423
Duration "More than 15"	0.742	0.191	15.101	1	0.000	2.100
Zone "External"			6.784	2	0.034	
Zone "Lateral"	-0.359	0.171	4.416	1	0.036	0.698
Zone "Posts"	-0.485	0.193	6.277	1	0.012	0.616
LongPasses	-0.015	0.192	0,006	1	0.937	0.985
Constant	0.498	0.171	8.518	1	0.004	1.646

Given substantial heterogeneity in the data, a Nagelkerke pseudo R-squared of 0.21 can be considered acceptable (Hu *et al*., 2006)

The results show that DurationCat and ZoneCat variables both significantly affect the probability of performing a good shot. In particular, an action that lasts less than 11 seconds has 1.423 times the probability of concluding with a good shot than an action that lasts from 11 to 15 seconds. And the probability of "long" actions, as in those which last more than 15 seconds, finishing with a good shot is 2.1 times higher than the probability associated to 'intermediate' ones. The reason may be that short actions are often the result of an extremely fast conclusion. For example, if the centre forward attains an exclusion and finds himself in front of the goal unmarked for a few seconds, this provides a great opportunity to score if the ball is passed to him rapidly. Longer actions, on the other hand, permit the attacking team to 'upset' the defence, thus creating good opportunities for goals.

As far as the ZoneCat variable is concerned, play actions which are more likely to generate a good attempt at scoring are those concluded from the 'external' zone. Attempts from 'lateral' zones

are more difficult because the angle of attack is narrower and the defence covers the goal area more effectively. Conclusions from the goal posts are also difficult because it is hard to get a good ball into that area. The defence covers that area densely and easily succeeds in neutralising the play action.

Despite a significant connection to a good shot, the LongPasses variable was not relevant in the model, even though long passes are thought to contribute to upsetting the defence.

Conclusions

This paper concentrates on extra-man play actions, believed to bear great importance on the outcome of a water polo match. Forty-eight matches comprising the men's 2020 European water polo championships in Budapest were considered.

A significant association was observed between the duration of man-up actions, the origin of the shots and the occurrence of good shots in the 979 actions analysed. Man-up actions that last less than 11 seconds and 'long' actions are more likely to produce good shots. In addition, external shots are more likely to be good shots than posts or lateral shots.

Several characteristics were recorded on each man-up action, but few of them seemed to influence its outcome. This may be explained by the fact that the outcome of a play action is not only linked to the execution of a strategy, but it is influenced by factors which cannot all be measured. The opponents' performance naturally has an effect on the game. When faced by a solid defence, the effectiveness of the attack is likely to suffer negatively. Psychological conditions can also have a positive or negative effect. The coach's role is to motivate his team and bring out the best in them particularly when the opponent is better or the odds at stake are high.

It is not unsurprising that clear cut indications to follow for a numerically superior attack were not found: the result of a match is not only a question of technique. The beauty of team sports also lies in observing the more unexpected result.

References

FINA (2020). *FINA Water Polo rules*, https://www.fina.org/water-polo. Book

Hu, B., Shao, J. and Palta, M. (2006). Pseudo-R^2 in logistic regression. *Statistica Sinica*, **16**, pp. 847-860. Journal Article

Hraste, M., Jelaska, I. and Clark, C. (2020). Impact of Time-Outs on Efficiency of Man-Up in Water Polo: An Analysis of the Differences between the Three Levels of Water Polo Players. *SportMont*, **18** (2020) 3, pp. 67-71. Journal Article

Lubisco A., Mignani S., Trivisano C. (2019). Assessment of game actions performance in water polo: a data analytic approach, in *ASA Conference 2019 Statistics for Health and Well-being*, CLEUP Coop. Libraria Editrice, 2019, pp. 141-144. Paper in Conference Proceedings

Lupo, C., Condello, G., Capranica, L., Tessitore, A. (2014). Women's water polo world championships: technical and tactical aspects of winning and losing teams in close and unbalanced games. *Journal of Strength and Conditioning Research*, **28**, pp. 210-222. Journal Article

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**, pp. 691-693. Journal Article

Takagi, H., Nishijima, T., Enomoto, I., Stewart, A. (2005). Determining factors of game performance in the 2001 world water polo championships. *Journal of human movement studies*, **49**, pp. 333-352. Journal Article

Tucher, G., de Souza Castro, F., Garrido, D., Silva, A. (2014). Relationship between origin of shot and occurrence of goals in competitive men's water polo matches. *Revista Brasileira de Cineantropometria e Desempenho Humano*, June 2014, **16** (2), pp. 136-143. Journal Article

Big data analysis and labour market: an analysis of Italian online job vacancies data

Francesca Giambona, Adham Khalawi, Lucia Buzzigoli, Laura Grassini, Cristina Martelli

1. Introduction

Economists and social scientists are increasingly making use of web data to address socio-economic issues and integrate existing sources of information. The data produced by online platforms and websites could provide a lot of useful and multidimensional information with a variety of potential applications in socio-economic analysis. In this respect, with the internet growth and knowledge, many aspects of job search have transformed thanks to the availability of online tools for job searching, candidate searching and job matching.

In European countries, there is growing interest in designing and implementing evidence-based decision-making tools to analyse Internet labour market data. The analysis of labour market online data could provide useful information, as big data - jointly with official statistics - could help answer the question namely "How to tackle the mismatch between jobs and skills?".

In this regard, the topic of skills gap, how to measure it, and how to bridge it with education and continuous training have been tackled by using the big data collection, as in the Cedefop (European Center for the Development of Vocational Training) initiative (Cedefop, 2018).

This contribution focuses on the issues arising from the use (and the usefulness) of online job vacancies (OJVs) to analyse the most recent Italian data. Data available for the years 2019 and 2020 are analysed to evaluate whether there has been any change in terms of required skills in occupations after the COVID19 pandemic. We use the index proposed by Deming and Noray (2020) that accounts for the change in skills for each occupation (here considered) between 2019 and 2020. Furthermore, some regional information is provided due to the particular importance that the territory has in the Italian labour market.

2. Online job vacancies and data

For some years on, OJVs have received increasing attention as an important source for real-time information on the labour market: thanks to the availability of more and more efficient big data analysis and text mining techniques, an enormous amount of information can be quickly collected and processed to monitor the changes in job demand.

These data provide a detailed and timely description of the jobs: the set of skills and the level of education and experience requested by the companies; the geographic location of the job; the type of contract; the economic sector of the company, etc..

In this sense, even if they cannot be used directly as a support tool for employment policies, they can be considered part of the modern view of Labour Market Information Systems (LMIS, see ETF, 2019), together with more traditional sources, such as statistical and administrative data. Moreover, OJVs also represent an important link between the labour market and the education system, because they provide updated information on the skills required by the market, an essential input to configure effective training offers (OECD, 2020). On the other hand, this type of data also has evident limitations and drawbacks, mainly related to representativeness and, in general, to quality issues (Cedefop, 2019).

Francesca Giambona, University of Florence, Italy, francesca.giambona@unifi.it, 0000-0002-1760-2062
Adham Kahlawi, University of Florence, Italy, adham.kahlawi@unifi.it, 0000-0003-4040-5590
Lucia Buzzigoli, University of Florence, Italy, lucia.buzzigoli@unifi.it, 0000-0003-3297-1023
Laura Grassini, University of Florence, Italy, laura.grassini@unifi.it, 0000-0003-4678-6507
Cristina Martelli, University of Florence, Italy, cristina.martelli@unifi.it

Our study is based on OJVs data produced for Italy by Burning Glass Technologies[1] (BGT), a company that collects millions of online job posting by scanning daily thousands of Internet sources (dedicated portals and company's websites).

The procedure for creating the database is very articulated and complex (ETF, 2019). The data are collected from different sources with various methods (API, scraping, crawling), based on the web portal characteristics, and are pre-processed to eliminate noise, outliers and duplicate entries. Then with the application of text classification algorithms the content of the ads is coded using categories based on reference taxonomies: in short, the taxonomy of variables is standardised according to the official classifications used in the various countries. These data have received increasing attention and have been analysed in numerous research works. Recently, Cammeerat and Squicciarini (2021) have analysed BGT data from a statistical point of view to assess their representativeness. Our data, in particular, refer to the OJVs posted on 239 online job portals in Italy in the period January 2019 - December 2020. The total number of ads is 1,741,621 in 2019 and 1,748,431 in 2020. They contain about 70 variables, most of them referred to official classifications (shown in brackets in the following): opening and closure date of publication, identification and description of occupation and related skills (ESCO classification), job geographic location (LAU and NUTS), economic sector of the company (NACE), educational level (ISCED).

To the aim of this contribution we use the BGT data to explore if between 2019 and 2020 skill changes occur by considering the occupation and regional classification.

3. Methods

Skill change is measured by the index proposed by Deming and Noray (2020) in order to understand if between 2019 and 2020 changes in skills required occurred.

For each year, BGT data collect all skills required for each job vacancy (*JobAds*) and for each occupation. The formulation of the index for the single occupation o is:

$$SCI_o = \sum_{s=1}^{S} \left| \left(\frac{\# JobAds_{os}}{\# JobAds_o} \right)_{2020} - \left(\frac{\# JobAds_{os}}{\# JobAds_o} \right)_{2019} \right|$$

where $\# JobAds_{os}$ is the number of job ads requiring skill s for the occupation o.
This index measures the net skill change in each occupation: the greater the index value the greater the skill change.

Due to the peculiarities of the Italian labour market, it may be useful to report the index value by region instead of occupation, in order to understand if and in which regions there has been the greatest change in required skills. To this aim, the above equation becomes:

$$SCI_r = \sum_{s=1}^{S} \left| \left(\frac{\# JobAds_{rs}}{\# JobAds_r} \right)_{2020} - \left(\frac{\# JobAds_{rs}}{\# JobAds_r} \right)_{2019} \right|$$

where r stands for each Italian region. And, finally, by crossing occupations and regions

$$SCI_{ro} = \sum_{s=1}^{S} \left| \left(\frac{\# JobAds_{ros}}{\# JobAds_{ro}} \right)_{2020} - \left(\frac{\# JobAds_{ros}}{\# JobAds_{ro}} \right)_{2019} \right|$$

---

[1] Source: Burning Glass Technologies. burning-glass.com. 2021.

4. Empirical findings

The index SCI_o is calculated for each occupation available in the BGT data to assess if changes occurred between 2019 and 2020. Highest values (i.e. the highest skill changes) concern mainly occupations related to the ICT as: s*tatistical and mathematical technicians and similar*, *software and application developers and analysts not classified elsewhere*, *web and multimedia developers*, s*oftware developers*, *specialists in databases and computer networks not classified elsewhere*, *web technicians* and *specialists in the design and administration of databases*. We find also some occupations as *public transport controllers and conductors*, *pawnbrokers and loan officers* and *education specialists not classified elsewhere*. Occupations as *geologists and geophysicists* have the lowest SCI values.

Overall, some skills required in 2019 disappear in 2020 such as: *MySQL* or *searching online information*; on the contrary, new skills appear in 2020 such as: *buy raw materials*, *maintain relations with suppliers*, *be updated on social media*, *interpreting the automatic call distribution data* and *create animation*.

For specific occupations, we find some skills that in 2019 are not required. For example *Android* in the occupation *social networking* or also *sell the services* in occupation *statistical and mathematical technicians and similar*. Overall, skills required in 2020 (respect to the previous year) mainly concern the (advanced) *use of computer and statistical tools*, *the ability to adapt to change* and *work in a team, offer support to customers.*

Due to the territorial characteristics of the Italian labour market, it is interesting to investigate if between 2019 and 2020 there was a change in the skills required at the regional level using the index SCI_r. Results highlight the index is higher for Molise, Calabria and Lazio, whilst the index is lower for Friuli Venezia Giulia, Marche and Emilia Romagna.
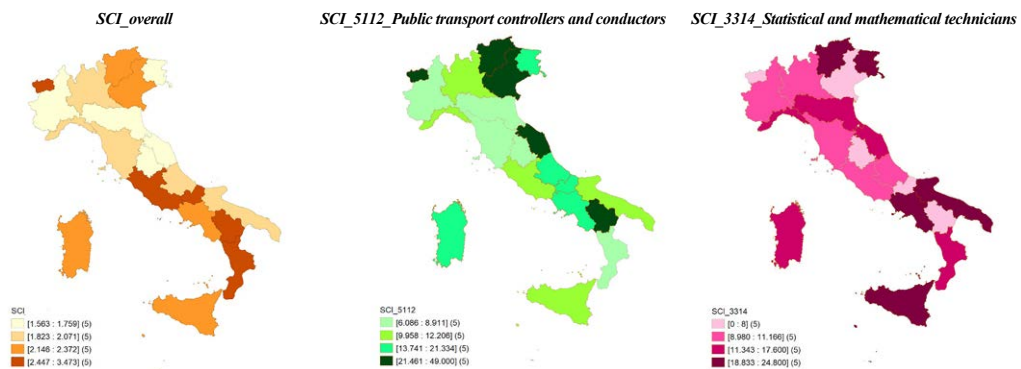
Graph 1: skill change index (SCI) at regional level

If we cross the information about occupation and regions it is possible to analyse, for the occupations with the higher SCI_o, in which region the change was highest and, therefore, whether there are any notable regional differences. Graph 1 displays the SCI_r values quartiles for the overall occupations, and the SCI_{ro} values of some occupations with higher changes.

In this respect, if we consider, for example, those occupations with highest values of SCI_{ro} we can appreciate slight different patterns across regions. In fact, we observe high values of the coefficient of variation (CV) of SCI_{ro} for those occupation with highest skill changes as, for example, CV=0.57 for *mathematicians, actuaries and statisticians*.

5. Some conclusions

The online job vacancies data give us the chance to improve information about labour market with the availability of timely data about the demand of businesses and the skills required for each occupation. In this contribution, by using the BGT data available for the years 2019 and 2020, we apply the skill change index proposed by Deming and Noray (2020) to understand if skills demand changed, for which occupation and if there are Italian regional differences.

Empirical findings suggest that between 2019 and 2020 skill changes occur, especially for some occupations and in some Italian regions. This result proves that the change in the skills required is obviously linked to each occupation (ICT-related occupations are the ones with the greatest dynamism) and to regional business environment.

By crossing occupation with regions, the skills change appears much differentiated between regions proving that for the same occupation the change in skill requirements coming from businesses are not the same, perhaps underlining a different local "perception" with respect to the skills required to carry out the same occupation.

References

Cammeraat, E., Squicciarini M. (2020), Burning Glass Technologies' Data Use in Policy-Relevant Analysis: An Occupation-Level Assessment. OECD Science, *Technology and Industry Working Papers* 2021/05.

Cedefop (2018), Mapping the landscape of online job vacancies. Background report: Italy, https://www.cedefop.europa.eu/en/publications-and-resources/

Cedefop (2019), Online Job Vacancies and Skills Analysis. A Cedefop pan-European Approach, The European Centre for the Development of Vocational Training, Thessaloniki.

Deming, D.J., Noray K. (2020). Earnings Dynamics, Changing Job Skills, and STEM Careers. *Quarterly Journal of Economics*. Forthcoming. ETF (2019), Big Data for labour Market Intelligence", European Training Foundation, pp. 1- 49

ETF (2019), Big Data for labor Market Intelligence, European Training Foundation.

OECD (2020), What skills do employers seek in graduates? Using online job posting data to support policy and practice in higher education, *OECD Education Working Papers,* No. 231.

Sizing & Allocation in Labour Market: business strategies and multivariate analysis

Andrea Marletta

1. Introduction

In Labour Market, the issue of Sizing and Allocation is a largely discussed problem (Mariani, 2002). In this study, this topic has been considered from a statistical point of view. Indeed, the choice to increase or decrease the number of employees after a change in the marketing strategy needs a very accurate analysis. If, for example, a company decides to launch a new product on the market, it could be necessary to recruit new resources. The proposed statistical approach aims to give some hints about how many (Sizing) and where (Allocation) these resources have to be placed. This process is based on the features of the existing market and the territorial geography. Statistically speaking, multivariate analysis techniques have been presented as exploratory tools. In the application, a Principal Component Analysis has been used to investigate the business environment after some qualitative interviews to the board of the company. In a second step, some different scenarios have been proposed to determine the exact number of new resources using a data hybridization technique including internal and external sources. Finally, the allocation of the new hired on the Italian territory has been achieved thanks to the construction of a territorial potential index.

2. Methodological tools

This study is the result of a collaboration of the Bicocca Applied Statistics Center (B-ASC) and a private company requesting a new rule based on a statistical indicator for reorganize their employees after the introduction in the market of a new product. The Bicocca Applied Statistics Center (B-ASC) aims to promote the application of statistical methodologies within private companies and public organisations. The Center's main objective is to represent a point of reference for companies wishing to develop a statistical approach to decision-making processes, using quantitative methods and integrated information processing systems.

In particular, this collaboration aims to offer different scenarios for representatives' activity, representing the most appropriate models to satisfy both the company's needs and its competitiveness within the reference market. The term "scenario" is here intended as a possible reallocation of the workforce following a change in the marketing strategy. To reach this purpose, some internal business data will be compared to Open Data, considering the type of subject of interest, the market dynamics and the prescriptive potential.

This project has been divided into four phases: firstly, a qualitative analysis has been provided through semi-structured individual interviews with company managers involved in the strategic and operational management of the markets; secondly, a structured database has been built through data collection and hybridisation of open data and business internal data; successively, a sizing model has been developed through the synthesis of indicators and weightings; finally, the measure of the actual and potential effort in terms of promotional pressure, indices of territorial potential may be applied to define the placement of the new resources in contiguous and/or nested areas.

From a methodological point of view, multivariate analysis techniques has been proposed as possible tools in order to achieve the company's purposes. Using data from the qualitative analysis based on individual interviews, a Principal Component Analysis has been applied considering the frequency distribution of the terms present in the textual corpus (Jolliffe, 2002). After the construction of the structured dataset, a Principal Component Analysis on some internal and external indicators has been used to synthesize the potential of a specific geographic area. Finally, the results from these two approaches are used to propose some different solutions for sizing and allocation of possible new resources in the company.

3. Principal results

The qualitative analysis based on individual interviews produced some evidences about the vision of the manager board of the involved company. Among the main results, some considerations of the managers have been extracted: "The effort required from the company appears blurred between the global and local vision. It is important to be able to integrate local needs with global strategies. The arrival of new products may be an opportunity for change. In view of a new product launch, the interviewees agree on rethinking the presence on the territory. This may happen in two ways, by acting on the mix of products on offer or on the sales force. A meeting point must be found between the company's revenue and the working efficiency of the employees."

Different scenarios have been proposed as an alternative to the current situation to contemplate a new product's launch addressing a new target. Some managers underlined the "short blanket dilemma": to add a new product, something else should be removed. Otherwise, it is necessary to make an investment. Defining a new structure may help to be more efficient and to manage new products launch in the future. Optimal segmentation and targeting are crucial. Some external barriers should be considered, e.g. regional restrictions.

All the interviews have been analysed to achieve the key concepts and obtain a multi-perspective vision of the company. Firstly, the term frequency has been considered to build a dictionary as a Text Mining technique. From a detailed analysis of the interviews, the main concepts have been extracted. The term frequency allows to obtain a quantitative variable, and for supporting the conceptual analysis, a PCA has been applied on these data. From the PCA, two components explaining the 74% of the variance have been extracted.
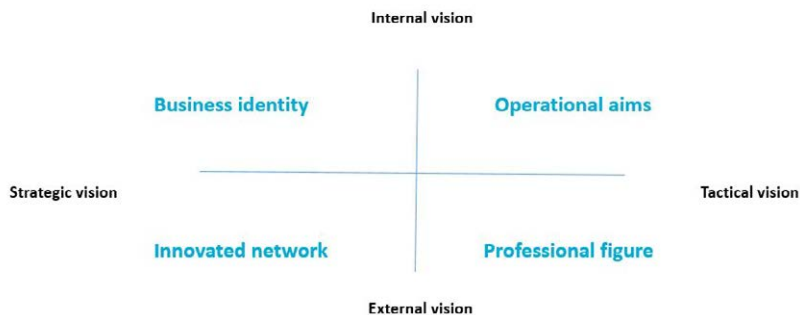
Figure 1: Cartesian plane after PCA qualitative analysis, 2019, Italy

The first component on the horizontal axis represents the continuum between a strategical or a tactical vision. The second component on the vertical axis represents the continuum between an internal or external vision. Using this technique, each term could be associated to one quarter

of the cartesian plane. In figure 1, it is reported the cartesian plane with some key concepts for each quarter. These concepts are the results of the set of words analysed, unfortunately for privacy reasons, it was not possible give more information about respondents and terms used.

After the qualitative analysis, in order to perform the sizing and allocation model, the final dataset containing quantitative indicators has been collected mixing internal and external sources. The internal sources are represented by data about sales of the products in the market and reports showing the daily activity of the employees. The external sources have been collected by using the healthcare data warehouse of National Institute of Statistics (Istat) named Health for All (Istat a, 2020) and the portal demo-istat for obtaining some demographic indexes in the Italian territory (Istat b, 2020).

The hybridization of these sources led to a structured dataset where each row represents an Italian geographical area. The used territorial classification is NUTS 1, NUTS 2 and NUTS 3. The variables for the definition of the potential have been divided into three categories: structural, market and promotional pressure. The first group is composed by some indicators about total, female and female 15-49 years population, birth rate, number of deliveries, physicians and total number of beds in specialized wards. In second group, some KPI about market sales have been reported. Finally, the third is about performance indexes of the employees.

In order to make assumptions about the capacity of a new team to determine the correct sizing, some hypothesis about the number of working days have been assumed. The potential portfolio has been computed only by considering the definition of the total number of physicians in portfolio and computing a number of visits for day. The final sizing model has been obtained through the potential portfolio, the number of physicians and the working days. The original workforce of the team was made of 112 employees. After the launch of a new product in the market, using the sizing model, the proposed new team is composed by 132 elements, with a differential of +20 employees[1].

Once the sizing phase is completed, the allocation phase allows to arrange the new resources in the Italian geographic area. The first allocation is about NUTS 1 and NUTS 2 units. An index of Territorial Potential (ITP) for Area and Region has been computed to detect under-estimated territories (Mariani, 2002). The ITP explains 79% of the variance using the first component of the ACP.

In table 1, for each Italian territory belonging to NUTS 1, it is possible to represent the allocation through the use of the ITP. The area with the biggest increase is South & Islands with a differential of 12 units. Similar results are available at NUTS 2 level.

NUTS 1	ITP	Actual Employees	Proposed Employees	Differential
North-West	23%	28	32	+4
North-East	25%	32	32	0
Centre	25%	28	32	+4
South & Islands	26%	24	36	+12
Italy	100%	112	132	+20

Table 1: Allocation of new hirings in the company for NUTS 1, 2019, Italy

In table 2, it is displayed the ITP for Italian regions. Lombardy is the region with the highest ITP, this means that in North-West area, a possible new hiring could regard this region.

---

[1]In order to respect a non-disclosure agreement between the B-ASC and the interested company, all quantitative results have been blinded and re-scaled.

NUTS 2	ITP
Lombardy	18.1%
Lazio	10.4%
...	...
...	...
Aosta Valley	0.01%
Molise	0.01%
Italy	100.0%

Table 2: ITP for NUTS 2 regions, 2019, Italy

Similar considerations could be hypothesized for NUTS 3 regions.

4. Conclusions

In this work, the problem of sizing and allocation has been considered at business level. In particular, starting from a real case, thanks to the application of some multivariate techniques, an exploratory approach has been proposed to determine the number of new hirings. This approach consists in a multi-steps procedure. Firstly, a preliminary analysis was based on some qualitative interviews to the top managers of the company. These interviews led a text mining analysis, in which through a PCA a dictionary based on frequency terms was obtained. This qualitative analysis allowed to visualize the possible strategies and the different visions proposed by the managers.

Starting from this qualitative analysis, a dataset was built after an hybridization of business and external sources to perform the sizing and allocation model. Each considered variable refers to an Italian geographical area. Similar analysis was performed at NUTS 1, 2 and 3 level. The sizing step was realized by considering the starting number of employees, the working days, the potential portfolio and the number of involved stakeholders. The allocation step was achieved through a PCA based on selected KPI about structural, market and promotional pressure. This PCA led to an Index of Territorial Potential (ITP). At level of NUTS 1, the South & Islands area has been detected as under-estimated, so the majority of new hirings could regard this area. At level of NUTS 2, Lombardy is the region with the highest ITP.

In conclusion, this approach could be considered as a valid alternative to solve the problem of sizing and allocation of new resources in Labour Market when a company chooses to launch a new product.

References

Istat a. (2020). Health for All. https://www.istat.it/it/archivio/14562

Istat b. (2020). Geo-demo. http://demo.istat.it/

Jolliffe, I. T. (2002). *Principal Component Analysis, 2nd edn. Series: Springer Series in Statistics*, XXIX, 487. illus. Springer, NY, 28.

Mariani, P., (2002). *La statistica in azienda, contesti ed applicazioni*. Franco Angeli, Milano.

Post-stratification as a tool for enhancing the predictive power of classification methods

F.D. d'Ovidio, A.M. D'Uggento, R. Mancarella, E. Toma

1. Introduction

As is well known, any decision-making model involving classification algorithms often faces the problem of predictive or diagnostic power (sensitivity or specificity), which tends to decrease rapidly as the asymmetry of the target variable increases (Sonquist et al., 1973; Fielding 1977). For example, segmentation analyses with categorical target variables generally provide very little improvement in purity (or none at all) if the least represented category accounts for less than one-fourth of the cases of the most represented category. The same problem occurs with other theoretically more exhaustive techniques, such as artificial neural networks. In fact, the optimal situation for any classification analysis is the maximum uncertainty, namely the equal distribution of the target variable.

Certainly, some classification techniques are more robust, such as those based on a logit transformation of the target variable (Fabbris & Martini 2002), which is less sensitive to the distribution's shape. However, even this technique is affected by the distributive asymmetry of the target variable, as will be shown below.

Indeed, beginning from the results of a direct survey in which the target variable (binary) was highly asymmetric (12.3% versus 87.7%), the first analysis performed here shows that even logit models with very significant parameter estimates can have an insufficient fit and such low predictive power that they are useless in decision-making processes.

To address this prediction problem, we tested a post-stratification technique originally developed to solve classification problems by making a training sample that is artificially symmetrical in terms of the target variable's distribution.

In this way, a substantial increase in goodness of fit and predictive ability was achieved for both the symmetrized sample and, more importantly, for the original sample, whose probabilities of success are assessed by the parameters estimated by the model.

2. The case study

A sample of participants in a national survey on dietary habits was studied from December 2020 to the end of May 2021 (in continuation of similar surveys carried out since 2018), selecting only those who had regularly completed the proposed questionnaire, corresponding to 2,562 people residing or domiciled in Italy. One of the research topics was the tendency on the part of Italians to eat away from home, i.e, in restaurants or pizzerias, in view of the restrictions necessitated by the COVID-19 pandemic.

The target variable resulted from a question about the frequency with which subjects tended to eat outside home, distinguishing between sporadic customers (who did so, at most, occasionally) and those for whom eating at a restaurant was a usual habit. The percentage of the latter, which had never been very high in previous years, dropped sharply to zero during the pandemic period, but because the survey investigated (even retrospectively) the eating habits of respondents, the result was not quite so poor. However, considering that the pandemic had already affected the social habits of Italians prior to 2021, the response variable shows that

Francesco D. d'Ovidio, University of Bari Aldo Moro, Italy, francescodomenico.dovidio@uniba.it, 0000-0003-1641-039X
Angela Maria D'Uggento, University of Bari Aldo Moro, Italy, angelamaria.duggento@uniba.it, 0000-0001-9768-651X
Rossana Mancarella, ARTI, Agency for Technology and Innovation of Apulia, Italy, r.mancarella@arti.puglia.it, 0000-0001-8179-4970
Ernesto Toma, University of Bari Aldo Moro, Italy, ernesto.toma@uniba.it, 0000-0002-4817-7169

more than 87.7% of respondents (2,248 people) fall into the "non-customers" group and less than 12.3% (314 people) fall into the "restaurant lovers" group.

Despite this outstanding asymmetry, an investigation was conducted into the individual characteristics that were found to be related in some way to the target variable[1] and may explain the motivations for this tendency to eat away from home.

To this end, a common logistic regression model was first developed for exploratory purposes[2]. Such a model is not reported here, because it includes variables with insufficient or zero significance, although it has high statistical significance for the variables gender ($p<0.002$), work position ($p<0.001$) and food-delivery frequency ($p<0.001$); however, the model has minimal fit to the data (Cox-Snell $R^2 = 0.094$; Nagelkerke $R^2 = 0.179$) and minimal predictive power for the category of interest: only 28 cases were correctly identified as "habitual customers" (8.9% of actual cases).

The *correct identification* of "non-customers" (2,224 cases, that is, 98.9% of the subgroup), in contrast, is very relevant, but this result seems trivial.

The overall percentage of correctly identified cases is 87.9%, but it should be noted that, simply assigning the sampling mode "non-usual customer" to predict all cases would have resulted in 100% correct classifications of non-customers and, of course, no correct classifications of usual customers. In short, over 87.7% of cases could be correctly classified simply by assigning the mode value, without the need for complex statistical processing.

Estimating a more articulated logistic model, one that also included the more important interactions among the explanatory variables (but was made parsimonious by using the stepwise forward-deletion criterion, i.e., gradually inserting the most related variables and removing the non-significant ones), did not improve the result.

However, the final model is shown in Table 1, and it is interesting in its own way. The reference categories of the explanatory variables (referred to as *baselines* and shown in brackets after the names in the table) are generally identified with the first category, except for employment position.

This model (although better than the previous one, at least in terms of potential generalization due to the statistical significance found for many variables and items) is also affected by an overestimation of "non-habitual customers," as shown in Table 2. In fact, compared to the almost perfect classification of these (99.1%), few regular customers of the restaurant are also correctly classified, at only 27. Therefore, the overall correct classifications are almost entirely due to the predominance of "non-habitual customers" in the sample, for which the usefulness of the model for predictive and decision-making purposes remains very limited or practically zero.

---

[1] The following individual characteristics were considered: *Gender* (F, M), *Age group* (18 to 80 years), *Highest level of education attained* (from primary or lower secondary school to PhD, also including higher non-university studies), *Employment position* (entrepreneur, full-time employee, part-time employee, self-employed, unemployed, student, retired, other position), *Marital status* ("Married or Cohabiting" to "Single/never married", but also "I prefer not to say"), *Dietary habits* (omnivore, omnivore with reduced meat in diet, vegetarian, vegan), *Average time spent preparing meals at home* ("No time, do not cook at home", and ranging from "Less than 30 minutes" to "4 hours or more"), *Frequency of using food delivery*, *Frequency of buying sustainable food*, *Frequency of buying fresh food*, *Frequency of buying local food*, *Frequency of buying organic food*, *Frequency of buying food "Made in Italy"* (all frequency questions ranging from "Never" to "Always"), *Willingness to pay an extra fee for "sustainable" food* (scale from "definitely not" to "definitely yes"), *Willingness to pay an extra fee for "Made in Italy" food* (same scale as previous question), *Annual income class* ("Not specified", and ranging from "Less than 4,500€" to "Over 130,000€").
[2] The statistical tests used in the analysis are 1) the maximum likelihood ratio test, in terms of improving the fit of the model by adding or removing variables, and 2) the Wald test, which is used to assess the statistical significance of individual parameters.

Table 1. *Estimation of logit model's parameters related to respondents' propensity to consume meals at a restaurant or pizzeria.*

Characteristics of the respondents		B	Std. err.	Exp(B)	p	
Work position	(baseline: *Other position*)				<0.001	***
Entrepreneur		1.076	0.484	2.933	0.026	*
Full-time employee		0.692	0.417	1.997	0.097	°
Part-time employee		-0.039	0.496	0.962	0.937	
Self-employed		0.506	0.451	1.658	0.262	
Unemployed, seeking work		-0.847	0.544	0.429	0.119	
Retired		-0.406	0.744	0.666	0.585	
Student		-0.341	0.439	0.711	0.437	
Marital status	(baseline: *Married or cohabiting*)				0.005	**
Widowed, Separated or Divorced		-0.353	0.429	0.702	0.410	
Single never married		0.407	0.150	1.503	0.006	**
Would not like to provide information		0.700	0.248	2.013	0.005	**
Use of food delivery	(baseline: *Never*)				<0.001	***
Only sometimes		0.430	0.208	1,538	0.038	*
Often		2.322	0.280	10.200	<0.001	**
Very often or always		2.177	0.479	8.818	<0.001	**
Buying of sustainable food	(baseline: *Never*)				<0.001	***
Only sometimes		-1.184	0.280	0.306	<0.001	***
Thick		-1.098	0.279	0.334	<0.001	***
Very often		-1.020	0.276	0.360	<0.001	***
Always		-1.633	0.599	0.195	0.006	**
Gender*Use of food delivery	(baseline: *F*Never*)				0.023	*
M* Only sometimes		-0.390	0.312	0.677	0.211	
M * Thick		-1.241	0.463	0.289	0.007	**
M* Very often or always		-1.608	0.806	0.200	0.046	*
Gender*Purchasing of sustainable food	(baseline: *F*Never*)				<0.001	***
M* Only sometimes		1.120	0.325	3.064	0.001	**
M * Thick		1.117	0.352	3.055	0.002	**
M * Very often		0.585	0.393	1.795	0.137	
M * Always		2.813	0.755	16.656	<0.001	***
Constant		-2.119	0.452	0.120		

Significance of parameters: (°) 10%; (*) 5%; (**) 1%; (***) 1‰

Table 2. *Matrix of correct classification of the model.*

	Expected response		Percentage of correct classification
Observed response	Non-customer	Regular customer	
Not a restaurant customer	2,227	21	99.1
Regular restaurant customer	291	23	7.3
% correct overall classification			87.8

3. Post-stratification for symmetrisation of the target variable

The main reason for the poor performance described above is undoubtedly the extreme asymmetry of the alternatives investigated. Indeed, if about 90% of the observations have one of the two modalities, in practice, any analysis aimed at assessing the probability of the complementary modality will be able to use only a minimal fraction of the necessary information. This phenomenon, which is almost fatal in other statistical techniques based on the search for the best predictability (for example, in the analysis of segmentation, Fabbris, 1997; Fabbris & Martini, 2002), is less relevant in logit analysis, especially when the samples are quite numerous. However, it persists and, sometimes, makes any decision rule impossible or very difficult.

Therefore, here, it was appropriate to experiment with a "Deep Learning" technique that has previously shown excellent results in solving very heavy penalties for symmetry in segmentation analysis and, later, in artificial neural networks elaborated on the basis of

dichotomous response variables (d'Ovidio, Mancarella & Toma, 2016): the formulation of a *symmetric learning sample* constructed by randomly extracting, from the group of statistical units with the majority response, a subgroup of the same size as the one indicating the minority response. The combination of the two subgroups provides a (post-stratified) sample that is almost symmetric in terms of the target variable, although it is undoubtedly smaller in size. In fact, through the above procedure, in addition to the 314 surveyed customers of restaurants and pizzerias, 320 people were randomly selected who ate out only occasionally or never[3]. The corresponding percentages are 49.5% and 50.5%, and the almost perfect symmetry of the distribution of the responses should improve the predictive power of the model.

Table 3, which was elaborated with the same criteria as the previous Table 1, highlights some important differences. First, there is an absence of significant interactions, so *Gender*, whose effect was previously diluted in the interactions, assumes considerable and significant importance in its own right; both the variables *Average time devoted to cooking at home* (but not its specific modalities) and *Willingness to pay an extra fee for food "Made in Italy"* assume statistical relevance; in contrast, *Marital status* and *Frequency of buying "sustainable" food* lose all their relevance and do not appear in the model, while *Use of food delivery services* (which indeed replaced restaurants and pizzerias in terms of the habits of many Italians in the pandemic period) retains statistical significance and much of its relevance. The model fits better, even if sample size is smaller: Cox-Snell $R^2 = 0.157$; Nagelkerke $R^2 = 0.210$.

Finally, the predictive power of the model assumes acceptable values (Table 4), reaching almost two-thirds of correct predictions for the target variable (and surpassing this level in the correct classification of respondents who do not tend to have lunch or dinner outside the home), with 63.4% correct classifications of regular customers of restaurants and pizzerias.

Table 3. *Estimation of logit model's parameters related to respondents' propensity to consume meals at a restaurant or pizzeria, symmetrised sample.*

Characteristics of the respondents	B	Std. err.	Exp(B)	p	
Gender: M (baseline: *F)*	0.752	0.196	2.120	<0.001	***
Work position (baseline: *Other position*)				0.015	*
Entrepreneur	0.593	0.646	1.809	0.359	
Full-time employee	0.317	0.546	1.374	0.561	
Part-time employee	-0.091	0.630	0.913	0.885	
Self-employed	0.112	0.581	1.119	0.847	
Unemployed looking for work	-1.067	0.638	0.344	0.094	°
Retired	-1.022	0.849	0.360	0.229	
Student	-0.193	0.549	0.824	0.725	
Average time devoted to cooking at home (baseline: *No time, no one cooks at home*)				0.050	*
Less than 30 minutes	0.720	1.348	2.054	0.593	
30 min–1 hour	0.968	1.312	2.633	0.461	
1–2 hour	0.319	1.314	1.376	0.808	
2–4 hour	0.760	1.325	2.138	0.566	
4 hours or more	1.061	1.496	2.888	0.479	
Use of food delivery (baseline: *Never*)				<0.001	***
Only sometimes	0.195	0.209	1.215	0.351	
Often	1.785	0.355	5.957	<0.001	***
Very often or always	1.964	0.802	7.125	0.014	*
Willingness to pay extra fee for foods "Made in Italy" (baseline: *Definitely not*)				0.050	*
Probably not	0.200	0.547	1.222	0.714	
Maybe yes, maybe no	-0.844	0.323	0.430	0.009	**
Probably yes	-0.316	0.292	0.729	0.280	
Definitely yes	-0.348	0.317	0.706	0.273	
Constant	-0.892	1.435	0.410		

Significance of parameters: (°) 10%; (*) 5%; (**) 1%; (***) 1‰

[3]The number don't match perfectly between the two groups, because an unavoidable approximation of the computerised procedure of random extraction of the sample of the non-customers respondents.

Table 4. *Matrix of correct classification of the model, symmetrised sample.*

Observed response	Expected response		Percentage of correct classification
	Non-customer	Regular customer	
Not a restaurant customer	217	103	67.8
Regular restaurant customer	115	199	63.4
% correct overall classification			65.6

The striking difference in structure between the model shown in Table 3 and the previous model is obviously due to the different hierarchy of objectives. The model shown in Table 1, while it aimed to identify the characteristics of individuals who tend to eat outside home, necessarily identified, instead, only the variables that characterise individuals who are not accustomed to eating in restaurants. The present model, on the other hand, correctly identified the primary required characteristics, certainly not optimally but well enough for the purposes of the study.

To investigate the reproducibility of the results obtained, it is possible, to calculate the value that the probability of success p of each unit of the total sample assumes using the estimated coefficients, in a logit transformation, for the symmetrised sample (of course, by setting the *baseline category* coefficient to zero):

$$\text{logit}(p) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m.$$

For each subject, the following is then calculated:

$$p = \frac{\exp[\text{logit}(p)]}{1 + \exp[\text{logit}(p)]},$$

rounding the result to the value that identifies the target characteristic "habitual customer" if this probability is close to 1, as well as to the value of the reference characteristic "non-customer" if it is close to zero. In practice, the cut-off line is assumed to be 0.5, in accordance with the given threshold for statistical software.

Thus, once the "expected condition" has been identified (and assigned to a specific record) for each unit of the joint sample, the collected and "expected" data can be easily compared in a contingency table that plays the role of the correct classification matrix.

This transfer (to the totality of the data) of the results obtained with the model derived from the symmetrised sample, as shown in Table 5, provides (as in other experiments previously conducted) results that are fully comparable to those obtained thus far, that is, quite adequate but not optimal. Presumably, beginning from a larger sample and randomly selecting the units to make the modalities of the target variable symmetric would yield a post-stratified sample large enough to guarantee the power and representativeness of the procedure.

Table 5. Matrix of correct classification of the model, applied to the whole sample.

Observed response	Expected response		Percentage of correct classification
	Non-customer	Regular customer	
Not a restaurant customer	1,534	714	68.2
Regular restaurant customer	115	199	63.4
% correct overall classification			67.6

4. Final remarks

The deep learning post-stratification method was first shown to be useful in classification techniques such as segmentation analysis (or artificial neural networks) for symmetrising a categorical response (d'Ovidio, Mancarella & Toma; 2016). In that study, in which no inference was involved, the method provided optimal and robust results. In the first analysis, using the CRT technique, 84% of the minority responses were correctly classified, as compared to 79%

of the alternative, while the original sample analysis provided only 50% correct classification of the interest responses and 99% of the alternative). The same results were obtained by applying the classification rules to the entire dataset (well over one million cases).

In the above research, artificial neural networks, of course, provided better results in the learning and testing samples and were more stable in population reporting (84% to 88% of correct classifications).

The application here shown, thus, demonstrate that post-stratification into symmetric groups provides an effective solution to the problem of the correct representation of relationships by more complex analyses, such as logistic regression. Further applications (including multinomial response variables) could provide a better understanding of the advantages and limitations of this technique.

Information and Acknowledgements.

References

d'Ovidio, F.D., Mancarella, R., Toma, E. (2016). Multivariate data analysis techniques for healthcare organizational efficiency improvement. In: *Proceedings of 5th International Conference "From Challenges to Opportunities: Development of Transition Countries in the Globalization Era"* (Elbasan, AL, December 17). Shpresa Print, Elbasan, AL: 24-39. Book Chapter or Paper in Conference Proceedings.

Fabbris, L. (1997). *Statistica multivariata. Analisi esplorativa dei dati*, McGraw-Hill, Milano. Book.

Fabbris, L., Martini, M.C. (2002). Analisi di segmentazione con una variabile dipendente trasformata in *logit.* In: Carli Sardi L., Delvecchio F. (eds), *Indicatori e metodi per l'analisi dei percorsi universitari e post-universitari*, CLEUP, Padova. Book Chapter or Paper in Conference Proceedings.

Fielding, A. (1977). Binary segmentation: the Automatic Interaction Detector and related techniques for exploring data structure. In: O'Muircheartaigh C.A., Payne C. (eds) *The Analysis of Survey Data. Volume 1; Exploring Data Structures*, Wiley, London: 221-257.Book Chapter or Paper in Conference Proceedings.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced. *International Journal of Emerging Technology and Advanced Engineering*, Vol. 2 (4): 42-47. Journal Article.

Sonquist, J.A., Baker, E.L., Morgan, J.N. (1973). *Searching for Structure*, Institute for Social Research, Ann Arbor, Michigan. Book.

A statistical information system in support of job policies orientation

Adham Kahlawi, Francesca Giambona, Lucia Buzzigoli, Laura Grassini, Cristina Martelli

1. Introduction

One of the main issues in modern labour market governance is about connecting people with jobs (Martin, 2015; Varanasi, 2021); bridging the skills gap is on the agenda of many governs and institutions (Mohla, 2020; Ras, et al., 2017), and many efforts have been done, also at the European level, to address vocational training to the real needs of the different economic sectors. Under the European Blueprint Initiative, for instance, stakeholders work together in sector-specific partnerships, called alliances for sectoral cooperation for skills, which develop and implement strategies to address skills gaps in different sectors[1].

In this perspective, the availability of suitable data centred on skills needs in the labour market is strategic and addresses vocational training investment and lifelong learning politics. However, up to now, data sources are mainly organised around the concept of occupation which is too wide to orient politics and investments. In this work, we intend to use the recommendation systems approach to describe the skills more requested by the different occupations in order to improve the granularity of labour market description.

Born in the era of big data, recommendation systems are a family of information filtering procedures that help users make choices in an extremely rich and variable information context (for a brief, recent review, see Jariha and Jain, 2018). They can also be interpreted as methods of predicting whether a particular user will like a particular item based on its preference structure and characteristics. These methods are widely used in various fields: to suggest the purchase of products to customers in e-commerce; to recommend news articles or blog contents to online content readers; to recommend movies or music to users of streaming services, etc. The two classic entities considered in recommendation systems are users (those who choose) and items (what is chosen): the user-item matrix, also called preference or utility matrix, shows the users by row and the items by column, and each cell contains a number that represents the importance of that item for that user. This number can simply be 0/1 (the user has/hasn't chosen the item) or can be the rating expressed by the user for the item. The matrix is typically sparse because many or sometimes most of the entries are empty: recommender systems consist of filling in the empty cells with what similar users would choose. Additional information about users or items can be added to get better results. This article aims to use recommendation systems to predict the future skills that a person has to acquire to reach a particular profession or to develop himself to improve his chances of getting a job.

The data source is always huge, and the system must be able to produce timely responses by continuously updating the information set that is fed by users' feedback. Therefore, the problem is to combine traditional statistical methods used to develop professional skills with data mining and machine learning techniques, which are able to solve the computational complexity of the system and optimise its performance. Many different approaches can be used to solve this problem (Leskovec et al., 2019). We will refer to model based collaborative filtering methods (Chen et al., 2018), that have received great success in many fields of applications. In this case, no information on users or items is requested, and the user-item matrix is factorised by means of latent factor models to reduce its dimensionality. Different algorithms can be used to map each user and each

---

[1] https://ec.europa.eu/social/main.jsp?langId=en&catId=89&furtherNews=yes&newsId=10035

Adham Kahlawi, University of Florence, Italy, adham.kahlawi@unifi.it, 0000-0003-4040-5590
Francesca Giambona, University of Florence, Italy, francesca.giambona@unifi.it, 0000-0002-1760-2062
Lucia Buzzigoli, University of Florence, Italy, lucia.buzzigoli@unifi.it, 0000-0003-3297-1023
Laura Grassini, University of Florence, Italy, laura.grassini@unifi.it, 0000-0003-4678-6507
Cristina Martelli, University of Florence, Italy, cristina.martelli@unifi.it

item to their corresponding factor vectors (Koren et al., 2009).

In our case, users and items are represented, respectively, by occupations and skills; the user-item matrix is built starting from a database produced by Burning Glass Technologies, which collects online job vacancy ads scanned from Italian online portals and company's websites in 2019 and 2020. The cell (i, j) of the matrix contains the number of ads that require skill j for occupation i. The use of recommender systems in this field of analysis is not new (Al-Otaibi and Ykhlef, 2012; Giabelli et al., 2021; Tavakoli et al., 2020; Valverde-Rebaza et al., 2018), but in this case the recommendation system is based on a dataset referred to Italy, in which occupations and skills follow the ESCO classification (European Skills, Competences, Qualifications and Occupations) (Kahlawi, 2020). In particular, the objective of the analysis is to help the vocational training systems and institutions to answer the question posed by every person looking for a new job or professional opportunities: which are the skills to have to enhance the professional profile? Finally, the matrix factorisation process is performed with the Alternating Least Squares (ALS) method and will be described in the next paragraph.

The results offered by the application of the proposed methodology will show which are the skills more requested in the framework of a specific occupation. Workers, job seekers, vocational training institutions, recruitment companies may take advantage of these results in different ways: Starting from the skills already owned by workers to suggest new skills for them, individuating the closest occupation that matches their skills based on the matrix and then comparing the actual profile with the most requested by the labour market. Alternatively, they may move from the concept of occupation to model updating skills politics.

2. Methodology

The methodology in this article is based on six basic actions, as shown in Figure 1.

- Action 1. The initial dataset contains different columns extracted from the job ads; for example, it has a column representing the occupation requested in the job ads after mapping it to the fourth level of the International Standard Classification of Occupations (ISCO-08). In addition, it has a column that represents the skills requested to be able to do this job. The user-item matrix is built using these two columns, and contains skills in the columns and occupations in the rows. Each matrix cell contains the number of times the skill has been requested for a particular occupation across all jobs ads.

- Action 2. We take the index of matrix cells that contain a value greater than zero, and then we randomly replace 20% of these values with zero. Afterwards, we replace each value greater than zero with the value of one.

- Action 3. For matrix factorisation, we use the ALS algorithm which is implemented in the Python implicit package[2], and built for large-scale collaborative filtering problems. ALS is doing a pretty good job at solving the scalability and sparseness of the compilation data, it is simple and scales well to enormous datasets. ALS has been used to solve different recommendation problems (Lakshmikanth et al., 2021).

- Action 4. First, we identify the occupations whose data has been hidden in preparing the test data. Second, we use the model that we built in the previous action to predict the values that have been hidden. Third, for each occupation, we calculate the Receiver Operating Characteristic Curve (ROC) to get the false positive rates and true positive rates which will be the input to calculate the Compute Area Under the Curve (AUC). Finally, we calculate the mean of AUC values of all occupations. The mean value represents the effectiveness of the model.

---

[2] https://implicit.readthedocs.io/en/latest/als.html

- Action 5. We use the model built in Action 3 to get the three best recommendations for a group of new job seekers.

- Action 6. We calculate the percentage of match between the current job seeker's skills and the skills required in the jobs ads (Match ratio). Then, we take the four job offers with the highest match ratio. Afterwards, we repeat these computations after adding the three skills recommended by the model to evaluate the improvement in job matching for the job seeker who has acquired these three skills.
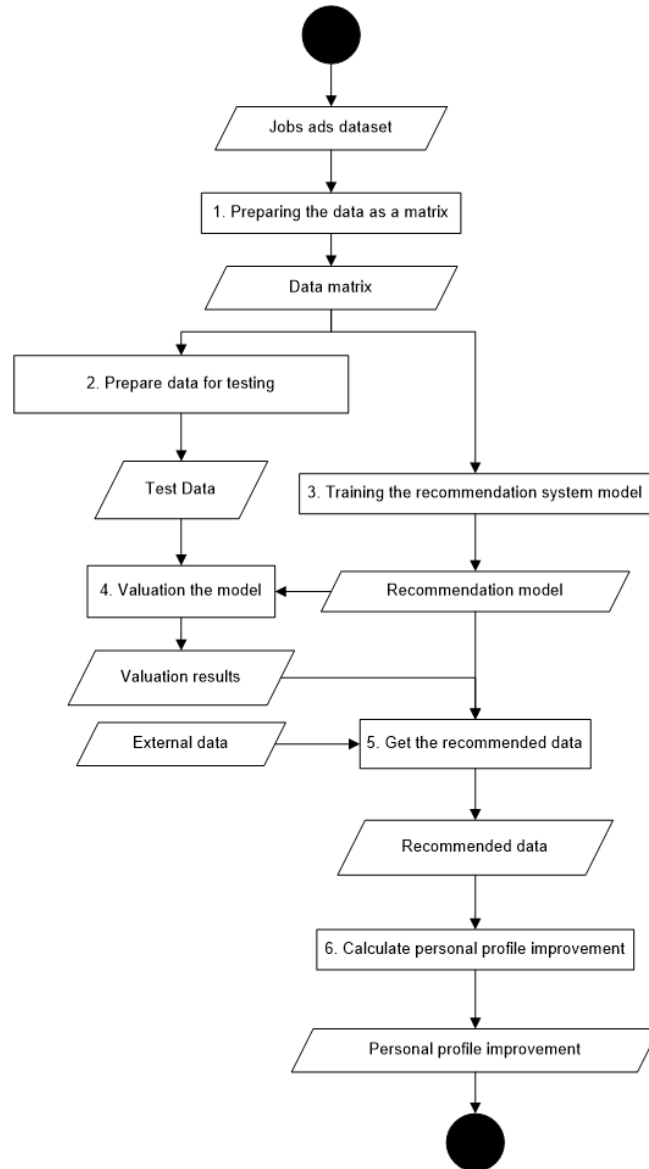
Figure 1 The methodology

3. Results and discussion

The effectiveness indicator of the recommendation system refers to the model's ability to

recommended with an accuracy of up to 95.9 percent, as shown in Table One, which expresses the metadata of this model.

Table 1 Model metadata

Initial Database (job ads)	32,426,926
Matrix shape	(326,1213)
Matrix sparsity	95.5%
Occupations involved in the evaluation process	279
Model effectiveness	95.9%

Table 2 represents the profile of two job seekers and the best three recommendations of our model.

Table 2 Professional profile and the recommend skills for two hypothetical job seekers

Job seeker	Current skills	Recommended skills
First	Teamwork, database, communication, adapting to change, thinking proactively, assisting the client, English, providing information, identifying client wishes, managing time, finance methods, thinking creatively.	work independently, tolerate stress, show enthusiasm
Second	Database, marketing principles, PHP, adjust priorities, CSS, create the front-end design of a website, integrated development software environment, machinery functionality, event planning, communication, financing methods, Scala, prioritise homework, English, pandas	office software, administer ICT systems, SQL

Consequently, Table 3 represents the extent of development that the users will achieve after getting these three skills by showing the top 4 job ads they can apply for. Indeed, it appears clearly from the results how the recommendation system helped users improve their chances of getting jobs directly and based on market demands.

Table 3 Personal profile improvement of two hypothetical job seekers

Job seeker	Moment of progress	Job ads id	Match ratio (%)	Sector
First	Before recommendation	159492369	92	Professional, scientific and technical activities
		247268757	90	Wholesale and retail trade
		161885864	88	Construction
		180547711	88	Wholesale and retail trade
	After recommendation	159492369	92	Professional, scientific and technical activities
		166831331	92	Information and communication services
		180554644	91	Wholesale and retail trade
		180547711	90	Wholesale and retail trade
Second	Before recommendation	543695754	83	Administrative and support service activities
		724284081	80	Manufacturing activity
		175809178	75	Administrative and support service activities
		357363988	75	Accommodation and catering services
	After recommendation	543695754	83	Administrative and support service activities
		363981505	80	Administrative and support service activities
		615486253	80	Professional, scientific and technical activities
		605508081	80	Professional, scientific and technical activities

4. Conclusion

Choosing the skills that a person has to learn to get a job opportunity or develop his job position is an ongoing problem because the labour market is constantly changing, and the skills required to do the job are constantly changing. Thus, the solutions provided have to be able to be continuously updated based on market changes. Indeed, this article proposed a recommendation

system based on a database collected from the labour market and capable of updating itself based on new data that can be obtained in the future from the labour market. Furthermore, the proposed recommendation system improved people's chances of getting new jobs through the skills that it recommended to these people. Finally, this work faced a set of limitations, the most important of which was the size of the matrix built from the initial data, which is why we used the same data for training and testing the model; for this reason, the proposed recommendation system is not considered a complete system and can not find all solutions for all users. Therefore, we will strive in future work to develop this system to become suitable for the largest possible segment of users.

References

Al-Otaibi S.T. and Ykhlef M. (2012), A survey of job recommender systems, International Journal of the Physical Sciences, **7**(29), pp. 5127-5142.

Chen R., Hua Q., Chang Y-S., Wang B., Zhang L., Kong X. (2018), A Survey of Collaborative Filtering-Based Recommender Systems: From Traditional Methods to Hybrid Methods Based on Social Networks, IEEE, **6**, pp.64301-64320.

Giabelli A., Malandri L., Mercorio F., Mezzanzanica M. and Seveso A., (2021), Skills2Job: A recommender system that encodes job offer embeddings on graph databases, Applied Soft Computing Journal **101,** pp.107049.

Jariha P. and Jain S.K. (2018), A state-of-the-art Recommender Systems: an overview on Concepts, Methodology and Challenges, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1769-1774.

Kahlawi A., (2020), An Ontology Driven ESCO LOD Quality Enhancement, International Journal of Advanced Computer Science and Applications (IJACSA), **11**(3).

Koren J., Bell R. and Volinsky C. (2009), Matrix Factorization Techniques For Recommender Systems, IEEE COMPUTER, **42**(8), pp. 30-37.

Lakshmikanth P, P. Radha K, J. V. R. M, (2021), Approaching the cold-start problem using community detection based alternating least square factorisation in recommendation systems, Evol. Intel. **14**, pp. 835–849.

Leskovec J., Rajaraman A., Ullman J. (2019), Mining of Massive Datasets, 3rd Edition, Cambridge University Press.

Martin P M., (2015), Activation and active labour market policies in OECD countries: stylised facts and evidence on their effectiveness, IZA Journal of Labor Policy **4**(4).

Mohla D., (2020), IEEE IAS Standards-A Pathway to Narrow the Technical Skills Gap [Standards News],in IEEE Industry Applications Magazine, **26**(5), pp. 94-95.

Ras E., Wild F., Stahl C., Baudet A., (2017), Bridging the skills gap of workers in Industry 4.0 by human performance augmentation tools: Challenges and roadmap, Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, pp. 428-432.

Tavakoli M., Mol S.T. and Kismih G. (2020), Labour Market Information Driven, Personalized, OER Recommendation System for Lifelong Learners, Proceedings of the 12th International Conference on Computer Supported Education (CSEDU), **2**, pp. 96-104.

Valverde-Rebaza J., Puma R., Bustios P. and Silva N.C. (2018), Job Recommendation based on Job Seeker Skills: An Empirical Study, In: A. Jorge, R. Campos, A. Jatowt, S. Nunes (eds.): Proceedings of the Text2StoryIR'18 Workshop, Grenoble, France.

Varanasi B., (2021), Knowledge Obsolescence and the Future of Work: Relevance of Knowledge and Impact to Jobs, Computational Thinking for Problem Solving and Managerial Mindset Training. IGI Global, pp.64-181.

Linear regression pathmox segmentation tree: the case of visitors' satisfaction to attend a Spanish football match at the stadium

Cristina Davino, Giuseppe Lamberti

1. Introduction

Segmentation trees have been attracting a great deal of attention as model comparison tools, with research mainly motivated by the fact that segmentation trees allow identification of partitions of data characterised by different dependency structures. Few algorithms have been proposed by the statistical community that combine model estimation and segmentation trees, outside the MOdel-based recursive partitioning (MOB) procedure proposed by Zelies *et al.* (2008). In a new approach we generalize the pathmox algorithm developed by Lamberti *et al.* (2016) to the context of linear regression models, using a model comparison test to identify the most significant partitions (i.e., sub-groups) in data. Further developments of the proposed approach will involve extensions to other contexts such as quantile regression.

2. State-of-the-art

Analysis of a dependency model can be furthered by assessing whether a model and/or the impact of regressors on dependent variables differ if heterogeneity is observed. In other words, it may be interesting to assess differences between a global model estimated on the whole set of observations and models based on sub-groups identified on the basis of known categorical variables external to the model. These variables may identify partitions characterised by a dependency structure heterogeneity. The most popular approaches to comparing regression models rely on comparative statistical testing or on recursive methods. The comparison approach consists of comparing coefficients related to a model common to all the data (i.e., a restricted model representing a homogeneous situation) and another model that reflects the interactions between categorical and predictor variables (i.e., an unrestricted model corresponding to a heterogeneous situation). The comparison approach, which allows for analysis of one categorical variable at a time, is reflected in the F-tests developed by Chow (1960) and Lebart *et al.* (1979), based on an assumption of the normality of the residuals of the two models. Comparison is done by calculating restricted deviance (SSR_0) and unrestricted deviance (SSR_1). The latter will be lower if interaction between categorical and predictor variables is significant. Under the null hypothesis, if both types of deviance are equal, then the categorical variables produce no differences in model coefficients. This null hypothesis is tested by computing an F–statistic:

$$F = \frac{(SSR_0 - SSR_1)/(n - 2p)}{SSR_1/p} \tag{1}$$

The recursive approach, based on multiple model comparisons, ranks variables that produce differences in the model coefficients. The outcome is a tree where each node represents a model. Partitions are obtained by comparing the effect of each categorical variable on the model coefficients and choosing the partitions that produce the biggest differences. This approach requires a criterion to quantify differences in the model coefficients. In case of the MOB

Cristina Davino, University of Naples Federico II, Italy, cristina.davino@unina.it, 0000-0003-1154-4209
Giuseppe Lamberti, Autonomous University of Barcelona, Spain, giuseppe.lamberti@uab.cat, 0000-0002-8666-796X

procedure this criterion is based on a fluctuation test that measures coefficient instability (Zelies and Hornick, 2007) as caused by a categorical variable. High instability points to a significant effect of the variable. Tree partitions are defined according to the variables that produce the highest instability.

3. Pathmox in a nutshell

Pathmox (Lamberti *et al.*, 2016), developed to detect heterogeneity in models, is a recursive algorithm based on segmentation trees. While pathmox was introduced in the context of partial least square structural equation modelling, it can be generalized to other contexts when a suitable test for comparing models is available. The algorithm applies binary segmentation principles to produce a tree with different models in each node. It starts by fitting a global model to all the data (i.e., the tree root) and identifies models with the most significant differences in child nodes. The most different models are identified by minimizing the sum of the squares of the residuals of the models estimated in each child node. The available data are recursively partitioned according to categorical variables – not included in the model – that yield the most significant differences in the child nodes. Partitions are identified using a test that determines the degree of difference between two compared sub-models. Finally, pathmox avoids overfitting using stopping rules based on maximum depth, minimum node size and non-significance of the partitioning criterion. As the partitioning criterion we propose the hypothesis test as proposed by Lebart *et al.* (1979) and Chow (1960) to compare two linear regression models.

4. Visitors' satisfaction to attend a Spanish football match: a pathmox application

We applied the pathmox approach in an empirical analysis to measure the visitors' satisfaction to attend a Spanish football match at the stadium. The sample consisted of visitors aged 18 years and older who attended Barcelona Football Club home matches during the 2017/2018 season. Visitors were selected using a no-random selection based on convenience. Three hours before matches started, randomly selected visitors were approached by seven researchers, who had previously reviewed and resolved any doubts regarding the questionnaire. The visitors were told about the purpose of the research and were asked to collaborate. If they agreed, they were asked to supply an email address to receive an online version of the questionnaire to be completed after the match. The questionnaire was available in the Catalan, Spanish, English and French languages to avoid bias due to the understanding of the questions by tourists. We offered the possibility of accessing the questionnaire through a QR code if they did not want to give an email address. Finally to encourage participation, respondents were entered in a lottery to win an authentic Barsa football shirt. A total of 944 visitors were invited to take part; the response rate of 38.45% meant that 362 usable questionnaires were collected. Men represented almost three-quarters (71.27%) of the respondents (women 28.72%), and nearly half (48.34%) were aged ≤30 years (34.52%, 31-45 years, and 16.71% ≥46 years). Involvement was strong, moderate, and slight in 32.50%, 48.76%, and 18.45% of the respondents, respectively. The percentage of tourist was 40,88% (no tourist 59.12%). 69.06% indicated that it was not the first time that they went to the Camp Nou Stadium.

The questionnaire was designed with closed questions answered on a 5-point Likert scale and aimed to measure the *visitors' satisfaction* in terms of perceived benefits of attending a Barcelona Football Club match (adapted from Ahrholdt et al., 2017; Oliver, 2010), *image of the football team* measured as visitors' perception of the attributes, players, management, and

condition of the club (adapted from Beccarini, and Ferrand. 2006), and *stadium service quality* measured as visitors' perception of service performance, based on evaluations of several service dimensions as tickets price, accessibilities, stadium facilities (adapted from Ahrholdt et al., 2017). The following categorical variables, reflecting specific visitors' characteristics, were considered as potential sources of heterogeneity: *gender*, *age* ($\leq$30, 31-45, $\geq$46 years), *involvement* (strong, moderate, slight), *tourist* (yes or not), and *first time at the stadium* (yes or not).

Pathmox analysis results are reported in Figure 1 and Table 1. We set maximum depth to two levels, bounded the final number of segments to a maximum of four and set the minimum admissible node size to 10% of the total sample. The significance threshold for the partitioning algorithm was p=0.05. The pathmox algorithm identified *involvement* as the variable with the greatest power, distinguishing between not involved– slight – (Node 2) and involved – strong and moderate – (Node 3). Not involved visitors were differentiated according to the variable *tourist* defining two terminal nodes: Node 4 (no tourist) and Node 5 (tourist). Involved visitors were further differentiated according to *age*: visitors aged $\leq$30 years form one group (Node 6) while visitors aged >30 years (Node 7) form another. On the basis of involvement combined with age and tourist, we could characterise partitions and assign labels to sub-groups. Thus, Node 4 can be defined as the group of not involved-local visitors, Node 5 as not involved-tourist visitors, Node 6 as younger-involved visitors, and Node 7 as older-involved visitors. Finally, the global model coefficients were compared with the coefficients for the four models estimated for the sub-samples identified by the terminal nodes (Table 1), showing that, in terms of satisfaction, not involved-local visitors primarily valued the image of the football team, not involved-tourist visitors valued more the quality of the stadium, younger-involved valued both image and quality in a similar way, and the older-involved valued again primarily the image of the football team.
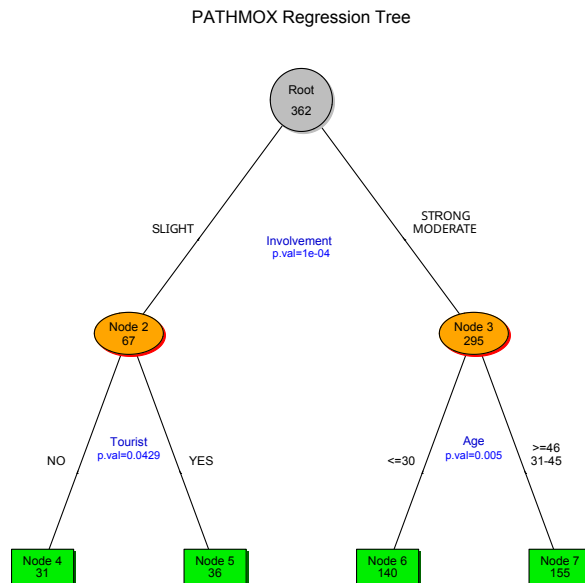
Figure 1: Pathmox tree

	Image of football team	Stadium service quality
Global model	0.493*	0.253*
Node 4: not involved-local	0.892*	0.468*
Node 5: not involved-tourist	0.319*	0.750*
Node 6: younger-involved	0.287*	0.243*
Node 7: older-involved	0.548*	0.166*

* indicates significance according to the t-test (p-value <0.05)

Table 1: Coefficient comparison for global and terminal nodes.

5. Discussion and conclusion

Our results suggest that pathmox can be used to compare regression models, opening up a future research line in other contexts such as quantile regression. From a decision-making perspective, the paper contributes evidence exemplifying how an apparently representative global model can in fact mask different relationships between variables due to heterogeneous data, underlining the importance of accounting for heterogeneity when defining new polices. While the algorithm allows partitions to be identified where differences between model coefficients are greatest, it has the limitation that no overall significance criterion is considered once each partition is identified. This important aspect needs to be considered in a future version of the algorithm. Note that pathmox aims to identify the most significantly different sub-groups, unlike a classic decision tree where the objective is to obtain the best prediction based on splitting observations into sub-groups. Therefore, the only similar method is the MOB proposed by Zelies *et al.* (2008), which, however, uses a different criterion to identify the best partitions. A comparison of both approaches will be a natural next step in our research.

References

Ahrholdt, D.C., Gudergan, S.P. and Ringle, C.M. (2017). Enhancing service loyalty: the roles of delight, satisfaction, and service quality. *Journal of Travel Research*, **56**, pp. 436–450.

Beccarini, C., and Ferrand, A. (2006). Factors affecting soccer club season ticket holders' satisfaction: the Influence of club image and fans' motives. *European Sport Management Quarterly*, **6**(1), pp. 1–22.

Chow, G.C. (1986). Test of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, pp. 591–605.

Lamberti, G., Aluja, T, Sanchez, G. (2016). The Pathmox approach for PLS path modeling. *Applied Stochastic Models in Business and Industry*, **32**, pp. 453–468.

Lebart, L., Morineau, A, Feenelon, J.P.(1979). *Traitement des donnees statistiques*. Dunod, Paris.

Oliver, R.L. (2010) *Satisfaction: a behavioral perspective on the consumer*. Armonk, NY: M.E. Sharpe.

Zeileis, A., Hornik, K., (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, **61**, pp. 488–508.

Zeileis, A., Hothorn, T., Hornik, K., (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, **17**, pp. 492–514.

Exploring competitiveness and wellbeing in Italy by spatial principal component analysis

Carlo Cusatelli, Massimiliano Giacalone, Eugenia Nissi

1. Introduction

The statistical observation of a complex social phenomenon such as of *well-being* includes a series of logical operations that lead to empirical indicators suitable for its study. The so-called statistical operationalization, which in this case will include the following basic steps:
- definition of the complex concept of well-being, which cannot be measured directly;
- specification of the concept, or its decomposition into directly observable and measurable dimensions and sub-dimensions;
- choice of indicators, measuring instruments of each dimension and sub-dimension.

It is not easy to define the concept of well-being: it concerns a social phenomenon with very different semantic and interpretative orientations that have their roots in various disciplines, such as Economics, Sociology, Psychology, Urban Planning and others. A concept that embraces all aspects of living can therefore only be part of a theoretical model built ad hoc. To begin to evaluate the problems that led to the construction of the model used in this work, we want to briefly recall the excursus that over time has led to the different formulations of an evolving concept: *quality of life*.

Some authors[1] identify the birth of the problem in question in the period of the industrial revolution. At that time, the profound changes in the living conditions of the European population led some researchers[2] to study, together with the income earned by the new types of workers, the price of a basket of goods consumed by them and the composition of their expenses, identifying significant differences between the structure of the budgets of the less well-off classes (in which almost all of the income was consumed for food) and that of the wealthier classes (who saw the relative share of food expenditure diminish). Such research, of a purely descriptive nature, spread during the nineteenth century in various European countries[3] and, essentially linked to the study of the subsistence budgets of working families, maintained a purely economic significance of studies on the level of survival of these families.

Subsequently, together with the evolution of the economic conditions of the working families, the concept under examination took on a different meaning, becoming more and more identified with that of satisfying needs that are no longer exclusively nutritional. One of the terms relating to this evolving phenomenon was that of *standard of living*, by which some scholars[4] defined the set of goods and services used in families whose type of life (*mode of life*) was determined by different parameters and social characteristics. According to a different

---

[1] See, e.g., W.S. and E.S. Woytinsky, *World population and production. Trends and outlook,* The Twentieth Century Fund, New York, 1953; J. Fourastié, *Machinisme et bien-être, niveau de vie et genre de vie en France de 1700 à nos jours,* Les éditions de minuit, Paris, 1962.

[2] See, e.g., F. Le Play, *Les ouvriers européens; études sur les travaux, la vie domestique et la condition morale des populations ouvriéres de l'Europe, précédées d'un exposé de la méthode d'observation*, Paris, 1855.

[3] See, e.g., E. Engel, *La consommation comme mesure du bien-être des individus, des familles, des nations*, in Bulletin de l'Institut International de Statistique, Tome II, Roma, 1887.

[4] See, e.g., A. Bowley, *The nature and the purposes of the measurement of social phenomena*, P.S. King & Son Ltd., London, 1923.

Carlo Cusatelli, University of Bari Aldo Moro, Italy, carlo.cusatelli@uniba.it, 0000-0003-3770-3060
Massimiliano Giacalone, University of Naples Federico II, Italy, maxgiacit@yahoo.it, 0000-0002-4284-520X
Eugenia Nissi, Gabriele d'Annunzio University, Italy, nissi@unich.it, 0000-0003-3440-601X

approach, other authors[5] addressed themselves towards measuring the individual psychic satisfactions that income could provide through the consumption of goods and servcies, formulating the utilitarian concept of *niveau de confort*. Still others[6], considering a broader meaning of well-being, came to the concept of a *level of living* which, by loosening the centrality of economic needs, also penetrated into the social areas of *welfare*.

Totally emancipated from definitions of an economic matrix that are not suited to an approach inspired by components of a purely social nature, the concept of *quality of life* dates back to the early Seventies, a period in which a particular and active political-social climate was developing. as opposed to the prevailing economist conception of progress: in Western societies characterized by high rates of economic growth (the *most*, the quantity), it was beginning to doubt that such growth could always be equivalent to social progress (the *best*, the quality). Roads choked by traffic, air pollution, difficult accessibility to services theoretically aimed at all citizens, the spread of new forms of poverty, difficulties in interpersonal relationships are just some of the phenomena that are easy to find in the most industrialized and, particularly, in urbanized contexts, where wealth and population are concentrated but also inequality and social hardship.

Another fundamental guideline of the debate on the quality of life was the requirement that the concept also contemplate the subjective aspects of human existence: the term *quality* implies in fact a personal judgment that is generally not measurable except through subjective indicators. Through the latter it is in fact possible to grasp the internalization of social problems by individuals (attitudes, judgments, perceptions, concerns, etc.). Furthermore, subjective indicators make it possible to complete and specify the information collected by means of objective indicators regarding the aspects (material and otherwise) of the quality of life, towards which individuals perceive a different satisfaction.

Given the interdisciplinary value of the phenomenon, and lacking a unanimous definition of the concept of well-being, the theoretical model formulated in this work stems from the consideration that the healthy evolution of the territory must correspond to a trend towards economic growth that also brings with it another type prosperity: efficiency and effectiveness of services, interpersonal relationships, culture, good housing conditions, and so on. But if on the one hand urbanization offers such undoubted localization advantages, on the other it produces growing disadvantages; therefore the individual who expects high living standards from it may, on the contrary, find himself paying the price attributable to the malaise produced by the degradation: atmospheric pollution, crime, etc.

Assuming that the urbanized one is still the space in which these expectations can be realized more easily, it becomes inevitable to ask the question of its livability when planning and governing the modernization processes of today's society. Therefore, in order to study the phenomenon in question in relation to a more circumscribed collective than that of the entire population of a nation (to which most of the existing studies on the subject refer for reasons of international comparability), it is advisable to examine a territorial area of great socio-demographic importance: the provincial one. The statistical units of this research are therefore represented by the 107 Italian provinces with respect to the objective indicators that will be described later.

The insights expressed so far highlight the evanescence of the concept of well-being which must therefore be stopped in its empirically measurable dimensions. Well-being is an important element as it discriminates against human aggregates by enriching the image of a territory, strengthening its attractiveness, highlighting, in short, its state of health, that is the ability to fulfill its different roles: it is a place of private residence, social place where meetings and

---

[5] See, e.g., Bureau International du Travail (BIT), *Les méthodes d'enquête sur les budgets familiaux*, Etudes et Documents, Série N, n.9, Genève, 1926.

[6] See, e.g., D.E. Christian, *International social indicators: the OECD experience*, Social Indicators Research, 1974, n °1, p. 169-186.

interactions are easier, public place where collective demand services can be used to the extent that agglomeration externalities exist. These are precisely the guidelines on which the model adopted here to evaluate the problem of quality of life is based: an integral concept of human development capable of grasping, in addition to the economic one, also other aspirations.

Nowadays a significant amount of data and information is available on topics of local interest which, especially at an objective level, can lead to a fairly complete examination of well-being according to the directions just outlined. Social concerns often examined in order to identify and compare levels of well-being between nations constitute a useful reference in the selection of thematic areas and indicators that can be adopted for this study. The topics that most interest the sphere of well-being can be summarized in the following areas of investigation: environment, housing conditions, roads, work, public services, crime, cultural level.

Within the aforementioned areas, suitable means of indication must therefore be sought. Their identification constitutes a very important phase of this study: after having defined and specified the object of the research and identified the statistical units of reference (the Italian provinces), it is necessary to select the indicators that are best able to measure the various aspects of the well-being.

Once the indicators were chosen on the basis of the foregoing, the data collection often provided only the raw material of the final product represented by the measuring instruments of well-being. The next step therefore concerned the construction of the indicators: this phase consists of all those operations in which some initial data are weighted or variously combined with each other in order to make them statistically comparable and theoretically representative of the phenomenon to be studied.

The simplest family of indicators is that of the so-called primary measures: they concern the amount of individual characters possessed by each statistical unit. At a higher level of complexity are the simple (or elementary) weighted indicators, constructed by dividing the primary measure by a reference variable (which is often another primary measure) called the basic measure; this operation, eliminating the source of variation determined by the basic measure, has the purpose of legitimizing the comparability of the data relating to the various statistical units.

Often, then, the need arises to combine the different simple indicators into compound (or synthetic) indicators, especially when the relationship between phenomenon and elementary indicator is not simply one-to-one, but rather problematic and complex: the same phenomenon can in fact be measured by means of several simple indicators, all different but sometimes partial with respect to the final dimension to be represented. They can be integrated into one or more models that give as an answer the level of each constitutive aspect of the phenomenon considered (partial compound indicators), or its overall level (a global compound indicator). These aggregations also make it possible to better visualize the state conditions, especially when one wishes to make comparisons between different realities, which are also necessary in this research so that the representation of the livability of our country can be interpreted.

The work is organized as follows: in section two some methodological aspects related to principal component analysis for spatial data will be presented; finally, an application to the the data of BES at local level NUTS 3 will be presented.

2. Spatial Principal Component Analysis

Principal Component Analysis (PCA) is one of the most popular multivariate statistical technique used for reducing data with many dimensions, and often wellbeing indicators are obtained using PCA: it is implicitly based on a reflective measurement model that it non suitable for all types of indicators. Mazziotta and Pareto (2013) in their paper discuss the use and misuse of PCA for measuring well-being. The classical PCA is not suitable for data collected on the territory because it does not take into account the spatial autocorrelation present in the data.

Spatial PCA techniques, specifically designed for spatial effects are available. The Geographically Weighted Principal Component (GWPC) is a method that adapt PCA for spatial effects. Given n observation x_i it depends from its location i in the space with coordinates (u, v) such that supposing $x_i \sim \mathcal{N}(\mu(u,v), \Sigma(u,v))$, the Geographically Weighted Principal Components are obtained through the decomposition of geographically weighted variance-covariance matrix (Harris et al 2011):

$$\Sigma(u, v) = \boldsymbol{X}^t \boldsymbol{W}(u, v) \boldsymbol{X}$$

where $\boldsymbol{W}(u, v)$ is a diagonal matrix of weights. Different kernels functions can be employed to generate the diagonal matrix of weights. Hence the local principal component at location (u_i, v_i).

$$\Sigma(u_i, v_i) = \boldsymbol{L}(u_i, v_i) \boldsymbol{V}(u_i, v_i) \boldsymbol{L}(u_i, v_i)^T$$

where $\boldsymbol{L}(u_i, v_i)$ is the matrix of geographically weighted eigenvectors and $\boldsymbol{V}(u_i, v_i)$ is the diagonal matrix of the geographically weighted eigenvalues.

Considering a set of p variables, the GWPC provides p components, p eigenvalues, p set of component loadings and p sets of component scores for each location in the study area.

An alternative way to assess the spatial variability of data in PCA is to consider the Locally Weighted Principal Component (LWPC) applied to the situation when data are not described well by an universal set of principal component (Tipping and Bishop 1999,). This technique use a moving window weighting approach in the data space. For each individual LWPCA around x, neighboring data point are first weighted according to some distance decay kernel function. Each observation is then multiplied by its respective weight and a standard PCA algorithm is (locally) applied to this weighted data.

Spatial effects can also be taken into account when PCA is combined with a measure of spatial autocorrelation. Jombart et al. (2008) have introduced a modification of PCA (called sPCA) to investigate the pattern of spatial variability of multivariate spatial pattern. The presence of spatial autocorrelation is measured using Moran's I (Moran, 1950). sPCA provides PC scores that summarize both the aspatial variability and the spatial autocorrelation structure in geographical space.

3. Empirical Results

The application consider the data of BES at local level NUTS 3, a system of equitable and sustainable well-being indicators at small-regions level that are consistent with the national Bes measures. To meet the statistical information needs of local communities, Istat designed Bes at local level in cooperation with local authorities, investigating the specific information needs of Italian Municipalities, Provinces and Metropolitan Cities and tuning a shared theoretical framework. Bes measures at local level maintain a high level of quality and consistency with the Bes indicators system and constantly follow the evolution of the Bes framework. The two frameworks share a core of common and harmonized indicators. In addition, Bes at local level includes specific well-being indicators, concerning some issues that are related to responsibilities and functions of local authorities (Istat, 2020).

The set of indicators, illustrating the 12 domains relevant for the measurement of well-being, is updated and illustrated annually in the Bes report. In 2020, the set of indicators has been expanded to 152 (it was 130 in previous editions), with a deep revision that takes into account the transformations that have characterised Italian society in the last decade, including those linked to the spread of the COVID-19 pandemic.

The first step in spatial analysis is to asses if a source of spatial correlation is present in the

data or not. Among all the possible alternatives, Moran's *I* is the test used for assessing the presence of spatial autocorrelation. The results of Moran's *I* test reveal a statistically positive spatial autocorrelation for each pillar (Tab. 1).

Table 1. Moran's *I* test of spatial autocorrelation.

Indicator	Statistics	p-value
Life expectancy at birth (male)	0.3481	6.064e^{-08}***
Life expectancy at birth (female)	0.3891	1.542e^{-09}***
Infant mortality rate	0.1827	0.002119***
Mortality rate for road accidents (15-34 years old)	-0.01504	0.5337
Age-standardised cancer mortality rate (19-64 years old)	0.2299	0.0001871***
Age-standardised mortality rate for dementia and related illnesses (people aged 65 and over)	0.2715	1.568e^{-05}***
Participation primary school	0.0508	0.1858
Participation in upper secondary education	0.2683	1.944e^{-05}***
Participation in tertiary education (19-25years old)	0.1767	0.0028***
Early leavers from education and training	0.5551	2.2e^{-16}***
Young people who do not work and do not study	0.5509	2.2e^{-16}***
Level of literacy	0.4651	8.681e^{-13}***
Level of numeracy	0.6191	2.2e^{-16}***
Employment rate of people 20-64 years old	0.4766	3.406e^{-13}***
Non-participation rate (15-74 years old)	0.4836	1.39e^{-13}***
Incidence rate of fatal occupational injuries or injuries leading to permanent disability	-0.0755	0.8381
Employment rate of women with and without children	0.1601	0.005711**
Per capita adjusted disposable income	0.3995	7.233e^{-10}***
Distribution of IRPEF incomes	0.4753	4.056e^{-13}***
Quality of dwellings	0.0622	0.1392
Number of people in workless households	0.5518	2.2e^{-16}***
Households with suffering bank debts	0.3152	6.986e^{-07}***
Volunteers in no-profit organizations (per 100 residents aged 14+)	0.2663	1.202e^{-05}***
No-profit organizations	0.3142	7.304e^{-07}***
Social cooperatives	0.3234	3.047e^{-07}***
of paid workers in local units of social cooperatives	0.2441	7.753e^{-05}***
Homicide rate	0.0589	0.1454
Burglaries	0.3746	6.744e^{-09}***
Pickpocketing	0.0496	0.1843
Robberies	-0.0542	0.7635
Presence of historic rural landscapes	-1.072	0.5659
Libraries	0.3598	1.271^{-08}***
Museums	0.3183	8.222^{-08}***
Visitors of libraries	0.2037	0.0003068***
Visitors of museums and similar institutions	-0.0427	0.7128
Drinkable water supplied every day per capita	0.4138	1.651e^{-10}***
Exceeding of the daily limit for the protection of human health for PM10 (Maximum number)	0.2217	0.000314
Urban parks and gardens	0.4735	4.718e^{-13}***
Protected Natural Areas	0.0242	0.307
Urban green areas	0.0386	0.1728
Noise pollution	-0.0283	0.628

The results shows that spatial dimension is relevant for all sustainable well-being dimensions and within each dimension, for most indicators. This evidence gives other further elements in favour of the application of Spatial PCA (sPCA) than a classical PCA in investigating determinants of provinces well-being for construction spatial composite indicator.

There is a strong spatial differentiation between positive and negative spatial correlation for each principal component obtained, recorded in the North and the South part of the country, respectively.

The spatial nature of sPCA ensure that the percentage of total variance explained can be decompose into pure variability and spatial autocorrelation.

The spatial patterns in the proportion of the explained variance vary significantly across the studied region, allowing in such a way to highlight territorial urban differences. In general, for the majority of the urban well-being domains the highest PTVs (Proportion of the Total Variance) are located in the Province capital cities in the south of Italy.

For the domain Health, the thematic maps of the local principal components reveal the

presence of a global spatial pattern.

By mapping the spatial variation of the first local component of the Education domain, we find out that it mainly considers the participation to primary school and this elementary indicator dominates in most urban areas, with some exceptions for a number of Province capital cities of Calabria and Sicily, where the leading variable is represented by the early leavers from education and training, aimed to capture the problem of school dropout. The dimension Education is characterized by local and global pattern.

For work and life balance pillar we can observe a lesser geographic variation in the influence of each variable on the first component: for the majority of cities the dominant variable is the employment rate of women.

References

Allardt, E. (1981). Experiences from the comparative Scandinavian study, with a bibliography of the project. *European Journal of Political Research*, **9**, pp. 101-111.

Andrews, F., Szalai A. (1980). *Quality of life: comparative studies*, Sage, London, (UK).

Bowley, A. (1923). *The nature and the purposes of the measurement of social phenomena*. P.S. King & Son Ltd., London, (UK).

Bureau International du Travail – BIT (1926). *Les méthodes d'enquête sur les budgets familiaux*, Etudes et Documents, Série N, **9**, Genève, (CH).

Christian, D.E. (1974). International social indicators: the OECD experience. *Social Indicators Research*, 1974, **1**, pp. 169-186.

Engel, E. (1887). La consommation comme mesure du bien-être des individus, des familles, des nations, in *Bulletin de l'Institut International de Statistique*, Tome II, Héritiers Botta, Roma, (IT).

Fourastié, J. (1962). *Machinisme et bien-être, niveau de vie et genre de vie en France de 1700 à nos jours*. Les éditions de minuit, Paris, (FR).

Jombart, T., Devillard, S., Dufour, A. B., Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, **101**(1), pp. 92-103.

Harris, P., Brunsdon, C., Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, **25**(10), pp. 1717-1736.

ISTAT(2020), Rapporto BES.

Le Play, F. (1855). *Les ouvriers européens; études sur les travaux, la vie domestique et la condition morale des populations ouvriéres de l'Europe, précédées d'un exposé de la méthode d'observation*, Imprimerie Impériale, Paris, (FR).

Mazziotta, M., Pareto, A. (2013). Methods for constructing composite indices: One for all or all for one. *Rivista Italiana di Economia Demografia e Statistica*, **67**(2), pp. 67-80.

Moran, P.A.P. (1950), "Notes on Continuous Stochastic Phenomena," *Biometrika*, **37**, pp. 17-33.

Tipping, M. E., Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **61**(3), pp. 611-622.

Woytinsky, W.S., Woytinsky, E.S. (1953). *World population and production. Trends and outlook*. The Twentieth Century Fund, New York City, (NY).

Total Process Error framework: an application to economic statistical registers

Roberta Varriale, Fabiana Rocci, Orietta Luzi

1. Introduction

As many other National Statistical Institutes (NSIs in the following), in recent years Istat has given new impetus to the renewal of its overall strategy for the production of Official Statistics. In this strategy, the production of the required outputs in all the statistical production areas is obtained based on the combined use of both primary and secondary sources of information. Primary data are those obtained by direct surveys, while secondary data correspond to information that are made available to NSIs by external bodies, and that are used by NSIs for statistical purposes (Memobust, AA.VV. 2014). Actually, one of the fundamental principles of the new Istat production strategy is the massive and integrated use of micro data from administrative sources (hereafter AD), which are used in particular for the construction of statistical registers. Besides other methodological aspects, this deep change in the statistical production paradigm requires to adapt standards and tools for the evaluation and documentation of data quality for the final users of the registers outputs and, more generally, of the outputs of multisource processes.

In this context, the Total Process Error (TPE) framework has been recently proposed in literature for assessing the quality of multisource processes, such as the production process of statistical registers. TPE framework can be used both to support the multisource process design and to monitor an overall production process, and can provide key elements for the assessment of the quality of both the processes and their statistical outputs.

In this paper, we describe how the TPE framework can be used referring, as a case study, to the Istat Register for Public Administrations. The production process of this register is still under construction, and is characterized by a modular structure depending on the different sub-population covered by the register itself. By using the TPE, we focus on the different steps and critical "decision points" of the production process for the different modules of the register. In section 2, we describe the main elements of the TPE, in section 3 we describe its application to the Register for Public Administrations.

2. The Total Process Error framework

Total Process Error (TPE) framework has been recently proposed in literature for assessing the quality of multisource processes (Rocci *et al*., 2022). The TPE framework represents an evolution of the Zhang's two-phase life-cycle approach (Zhang, 2012).

The TPE includes two phases of assessment, that can be described as: Phase 1. Assessment of single data sources w.r.t. original source purposes; Phase 2. Combination/re-use/integration of data sources w.r.t. target statistical purposes, that can be further splitted in: Phase 2a. Assessment of single data sources w.r.t. target statistical purposes and Phase 2b. Assessment of the combined data sources w.r.t. target statistical purposes. For each phase, some potential errors that may arise together with specific indicators to assess them are identified.

The TPE also includes an operative tool to connect the steps of a multisource production process to the phases of the quality evaluation framework: actually, this tool consists of a cross-classification scheme describing the link between the process steps of an entire production process and the above mentioned phases of the TPE framework. The cross-classification scheme may be used both to support the design of the statistical production process and to monitor the whole process once it has been put into production. Furthermore, the scheme allows to use the TPE in a

very flexible way to represent different production processes. Table 1 shows the cross-classification scheme for a multisource production process using AD composed by *N* steps.

Table 1. Cross-classification scheme: production process steps vs TPE phases

Process step	Phase		
	1. Assessment of single AD w.r.t. administrative purposes	**2. Combination/re-use/integration of AD w.r.t. target statistical purposes**	
		2a. Assessment of single AD w.r.t. target statistical purposes	**2b. Assessment of the combined AD w.r.t target statistical purposes**
1			
2			
…			
N			

3. The register for public administrations, territorial bodies

The economic Register for Public Administrations (hereafter Frame PA) is the result of an Istat project started in 2019. Frame PA is a *satellite* register of the *base* Register of Public Administrations (S13 hereafter). The latter defines the Italian public administrations as a subset of the Italian business register units. The difference between *base* and *satellite* register is in the role they play in the statistical production system, given the target (sub)populations and variables they are referred to. Following Wallgren and Wallgren (2014), we can define the base registers as the ones that represent the statistical reference populations for all the statistical processes (individuals/hoiseholds, economic units, etc.) and the satellite registers as those releasing additional variables usually representing specific phenomena. The information contained in the final statistical Register Frame PA will be, for each statistical unit, both structural information coming from the Register S13, and some economic variables respecting accountancy definitions.

Frame PA includes different subpopulations. Nowadays, Istat is working on the subpopulation of Local Authorities, including municipalities, unions of municipalities, provinces, mountain communities, metropolitan cities, regions and autonomous provinces.

The first step to build Frame PA for Local Authorities (hereafter Frame PALA) is to select the statistical units from the Register S13, together with some structural information (address, number of employees, etc). Subsequently, information from AD sources is extracted, integrated and treated to produce the final output, that are some economic variables according to the statistical target accountancy definitions. The main AD sources concerning the economic variables of Local Authorities are the Public Administration Database (BDAP), and the Information System on the Operations of Public bodies (SIOPE). BDAP records the accounting variables of balance sheets according to the Financial Statement Management Schemes; SIOPE is a system of digital collection of profits and payments made by treasurers and cashiers of all Public Administrations. Both BDAP and SIOPE can be can be queried at different times of a reference year to acquire periodic updates.

Following the subject matter experts' indications, taking into account the target population and variables of the Frame PALA, the BDAP has been defined as the primary source of information, as it is provides information consistent with the statistical target accountancy definitions. This choice implies that, after drawing and integrating information from BDAP and SIOPE, missing information in BDAP need to be estimated (imputed), by using SIOPE as auxiliary variables.

Different features of BDAP source characterize the Local Authorities: information on

municipalities, unions of municipalities, provinces, mountain communities and metropolitan cities is affected by total missing values, while information on regions and autonomous provinces (22 bodies in total) usually do not suffer of this problem.

Three variables are considered in the process, both on the revenues and the expenses sides. Let Y_1^{BDAP}, Y_2^{BDAP} and Y_3^{BDAP}, with $(Y_2^{BDAP} + Y_3^{BDAP}) = Y_4^{BDAP}$, be the variables observed in BDAP and Y_4^{SIOPE} the variable observed in SIOPE corresponding to Y_4^{BDAP}. The revenues and expenses are specified in Frame PALA across 148 and 22 "items", respectively, that are grouped in Titles. We will refer to the 148 and 22 items as the Frame PALA "theoretical scheme".

In case of total missing values from BDAP, such as for municipalities, unions of municipalities, provinces, mountain communities and metropolitan cities, missing information in BDAP have to be fully imputed, by using SIOPE information as auxiliary variables.

Table 2 shows the coverage of BDAP at different times during 2020 and 2021 for units belonging to the base Register S13 population. The reference year for data of both Register S13 and BDAP is 2019.

Table 2 – Coverage od BDAP source with respect to the target population (Register 2013), for Local authorities type – Number of respondents. Year 2019.

Local authorities type	Total population, Register S13, 2019	Respondents, July 2020	Respondents, October 2020	Respondents, May 2021
Provinces (excluded autonomous provinces) and metropolitan cities	100	80	93	98
Municipalities	7914	6455	7521	7806
Mountain communities	151	62	71	83
Unions of municipalities	562	282	324	363
Total	8727	6879	8009	8350

The presence/absence of total missing values in BDAP, makes the design of the Frame PALA production process for the two groups of local authorities completely different. Tables 3 and 4 show how the cross-classification scheme may be used to support the design of these two production processes.

Without going into the details of the two production process steps, it is clear that the process relating to the population of municipalities, unions of municipalities, provinces, mountain communities, metropolitan cities is more complex, and comprehend both an integration and an imputation step that are not present in the production process of Frame PA for the populations of regions and autonomous provinces. This means that this process is characterised by additional critical "decision points" and potential errors that may arise. The indicators linked to these steps (and phases) will be useful to support the design of these two different production processes (Rocci *et al.*, 2022).

Table 3. Frame PA, *municipalities, unions of municipalities, provinces, mountain communities, metropolitan cities*: production process steps vs TPE phases.

Process step	Phase		
	1. Assessment of single AD w.r.t. administrative purposes	**2. Combination/re-use/integration of AD w.r.t. target statistical purposes**	
		2a. Assessment of single AD w.r.t. target statistical purposes	**2b. Assessment of the combined AD w.r.t target statistical purposes**
1	Quality assessment of each candidate AD source (*BDAP, SIOPE*)		
2		Quality assessment of each AD source (*BDAP, SIOPE*) in terms of Frame PA purposes	
3			Integration of AD sources (*BDAP, SIOPE)*, by following a "theoretical scheme"
4			Imputation of the total missing values of the variable Y_4^{BDAP}
5			Imputation of the (totally) missing values of the variables Y_1^{BDAP}, Y_2^{BDAP} and Y_3^{BDAP}
6			Computation of the output Frame PA variables as aggregation of Y_1^{BDAP} values for different items

Table 4. Frame PA, regions and autonomous provinces: production process steps vs TPE phases.

Process step	Phase		
	1. Assessment of single AD w.r.t. administrative purposes	**2. Combination/re-use/integration of AD w.r.t. target statistical purposes**	
		2a. Assessment of single AD w.r.t. target statistical purposes	**2b. Assessment of the combined AD w.r.t target statistical purposes**
1	Quality assessment of each candidate AD source (*BDAP*)		
2		Quality assessment of BDAP source in terms of Frame PA purposes	
3			Transformation of the BDAP variables in the Frame PA output variables (computation of the output Frame PA variables as aggregation of Y_1^{BDAP} values for different items)

In the future, Frame PA will comprehend additional statistical populations, characterized by a different structure of information sources. Therefore, the production process of the output

economic variables will have different steps and critical "decision points". TPE was a useful tool in the design phase of the Local Authorities component, it will be used in the design phase of the other components and will also be used for their monitoring once it is put into production.

References

AA.VV. (2014). *Memobust Handbook on Methodology of Modern Business Statistics*. Available at: https://ec.europa.eu/eurostat/cros/system/files/Memobust%20glossary%20def.pdf.

Rocci F., Varriale R., Luzi O. (2022). Total process error: An approach for assessing and monitoring the quality of multisource processes. Forthcoming in *Journal of Official Statistics*, June 2022.

Wallgren, A. and B. Wallgren. (2014). *Register based statistics: Administrative data for statistical purposes*. John Wiley & Sons, Ltd.

Zhang, L.C. 2012. Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, **66** (1), pp. 41-63.

HEALTH AND WELL-BEING

Development of an innovative methodology to define patient-designed quality of life: a new version of a wellknown concept in healthcare

Barbara Bartolini, Serena Bertoldi, Laura Benedan, Carlotta Galeone, Paolo Mariani, Francesca Sofia, Mariangela Zenga

1 Introduction

Quality of life (QoL) is a concept embracing several aspects and functionalities of people's lives. Some of the areas affected are health, relationships, socializing, leisure (The International Society for Quality of Life Research). The achievement of a good QoL is recognized as an essential aim of health assistance, regardless of the pathology and the administered therapy (M. Asadi-Lari et al., 2004). QoL is therefore a pivotal parameter used by clinicians to evaluate how treatments and therapies influence patients' functionality and emotional state, aiming to ameliorate interventions and their outcomes.

QoL is determined by indices assessed by administering questionnaires that can be either generic or disease-specific (D. L. Patrick & Deyo, R. A., 1989; R. Rabin & de Charro, F., 2001; J. E. Ware, Jr. et al., 2016). Currently, the majority of the QoL questionnaires are designed with the main contribution of clinicians and, therefore, include items that are centered on the disease rather than on its multifaceted impact on people's life. These tools are useful for clinicians in determining the best clinical approach, but may fail to truly grasp the patients' perspective, needs, aspirations, perceptions and emotional state, resulting in a major drawback that sets medical care on clinical parameters alone. A proper tool defining the patient's perception of the pathology is missing.

To bridge this existing gap, the definition of a bottom-up patient-designed QoL index could provide a new, patient-centric, unbiased tool to evaluate the patients' perception of their own well-being. Here we describe the development of an innovative methodology to define patient-designed QoL, based predominantly on patients' contribution.

2 Working group and methodology

To define a patient-centric QoL tool, we used a consensus technique aiming to favor the expression of the major players involved in dealing with the pathology.

The Delphi method is currently widely used in academic research, industry, social sciences and healthcare to reach consensus (R. Boulkedid et al., 2011; I. R. Diamond et al., 2014; M. K. Murphy et al., 1998; Robinson N Trevelyan EG, 2015). The main goal is to collect different opinions to be evaluated by the panel, with the aim of reaching pluralistic evaluations of an issue. In the Delphi panel, the participants are either technical or non-technical experts (i.e., patient representatives) reporting their own point of view (G. Mazziotta Marbach, C. & Rizzi, A., 1991).

In our model, patients and healthcare professionals constitute the working group to build the settings and assertions of the questionnaire.

Barbara Bartolini, HPS, Health Publishing and Services, Italy, bbartolini@aboutpharma.com, 0000-0001-9646-9310
Serena Bertoldi, Science Compass, Italy, serena.bertoldi@sciencecompass.it
Laura Benedan, University of Milano-Bicocca, Italy, laura.benedan@unimib.it, 0000-0003-0427-2487
Carlotta Galeone, University of Milano-Bicocca, Italy, carlotta.galeone@statinfo.org, 0000-0003-1934-5167
Paolo Mariani, University of Milano-Bicocca, Italy, paolo.mariani@unimib.it, 0000-0002-8848-8893
Francesca Sofia, Science Compass, Italy, francesca@sciencecompass.it
Mariangela Zenga, University of Milano-Bicocca, Italy, mariangela.zenga@unimib.it, 0000-0002-8112-5627

The items of the QoL questionnaire were defined during focus groups involving a panel of patients, two clinicians, one statistician and one facilitator.

Patients are the active players that identify the settings and the main items. According to the European Patients' Academy on Therapeutic Innovation (EUPATI) (K. Warner et al., 2018) definition, patients involved may be:

- Individual Patients, i.e., "persons with experience of living with a disease. Their main role is to contribute with their subjective experience";
- Patient Organization Representatives, i.e., "persons who are mandated to represent and express the collective views of a patient organization on a specific issue or disease area".

Clinicians offer the technical knowledge of the pathology, supporting the precise description of the items. Facilitators act as a guide for discussion among the parties, with the important role of harmonizing the contribution of all the participants to avoid a decisional bias due to the influential opinion of the clinicians on patients' decision.

The Pseudo-Delphi we propose is an iterative process with subsequent steps aiming to identify a shared solution. It is a flexible method, useful when the identification of the statistical model to be applied is uncertain and when the gist of the discussed problem is not completely known.

The Pseudo-Delphi steps are as follows:

- Problem definition and identification of the expert panel
- First round to collect the basic aspects of the research
- Definition of the first questionnaire with open questions
- Administration of the questionnaire to each expert, who respond anonymously
- Second round to discuss questionnaire's answers and definition of the Likert scale and the closed questions
- Definition of the second questionnaire, open questions and closed questions with the respective agreement scale
- Administration of the second questionnaire to each expert
- Third round, with discussion on the questionnaire's answers, identification of the scale of importance with the evaluation of the possible consequences of each decision and the feasibility of each defined option
- Moderated feedback, result aggregation and sharing with the experts. In this round, the new questionnaires resulting from the previous opinions are also shared
- Repetition of the questionnaire, if necessary, to reach the agreement.

Working anonymously allows to avoid the prevalence of a charismatic individual over the others, which can freely express their opinions without any social pressure.

Moreover, the feedback control allows the experts to be provided with all the information needed to reach the final agreement. With the Pseudo-Delphi method all the participants can analyze and re-consider a variety of aspects included in the questionnaire. This method entails a great effort for the methodologists in estimating and summarizing facts, quantitative data and subjective variables. This approach is of value in the analysis of real-world data, especially in QoL evaluation (S. Pietersma et al., 2014).

The workflow starts from the patients' evaluation of a list of settings identified by a literature search. Patients independently provide social, economic, and organizational information related to their pathology and their relationship with the healthcare system. This allows the identification of the settings of the questionnaire.

The setting evaluation is carried out by every single patient anonymously, to avoid any kind of psychological subjection from the healthcare professional's opinion. At the end of this process, the group meets in a roundtable session to openly discuss about the settings emerged

and to rank them according to the perceived order of importance. This step is crucial to skim the settings and find those to be included in the questionnaire, to make it usable in daily practice.

Within each setting, a series of assertions are generated individually and anonymously by each patient. Following similar steps, a set of assertions is identified and included in the questionnaire. Every assertion is then associated to a four-point Likert scale. On the basis of the score, a synthetic patient-centric QoL index is then defined (Figure 1).
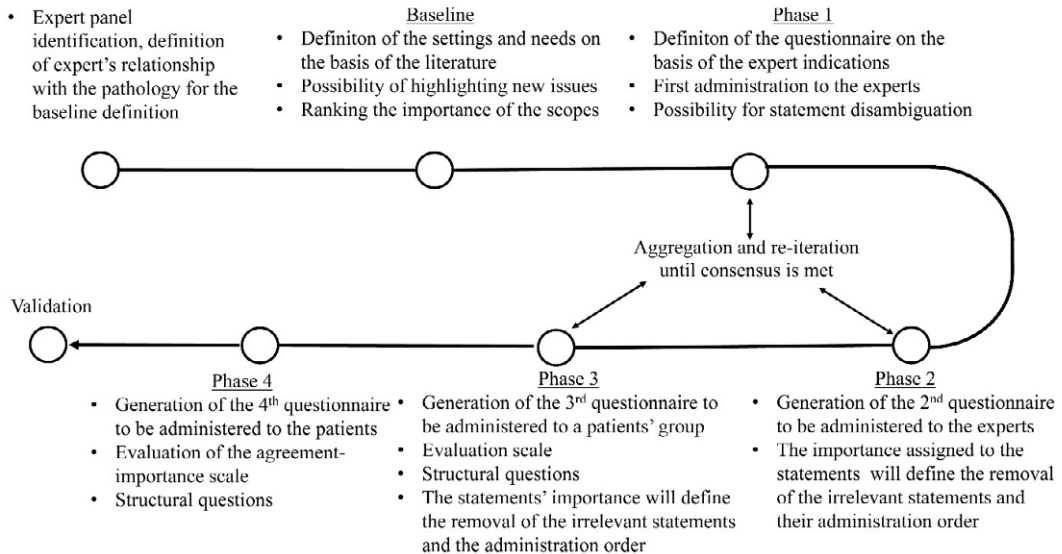
Figure 1. Flow chart showing the generation of the QoL questionnaire.

The final version of the questionnaire contains the scales for agreement and importance measures. They aim to link the agreement to one item with the importance for the patient in her/his life and they are built on the basis of the Customer Satisfaction Techniques. After the identification of the different settings (e.g., physical, emotional, social, functional and economic) the level of agreement and the level of importance of the statements within each setting are rated on a 4-points Likert scale (response categories: not at all; a little; quite a bit; very much) by the participants. The methodology allows the production of a composite index for "uneasiness", which will be then compared to the internal control –provided by the evaluation of each own QoL on a one to ten scale. The composite index is defined as follows.

Let x_{ijs} (i=1,…, n; j=1,…, k_s; s=1,...S) be the agreement of the i-*th* respondent on the j-*th* statement for the *s-th* setting. The categories on the agreement part for a statement are treated as numeric variable where "not at all" =0.001; "a little" =0.33; "quite a bit" =0.67; "very much" =1. In this case we transform the variable at 4 categories in 3 categories where the distance between each successive item category is equivalent and equal to 0.33. The agreement on *not at all* is treated as the lack respect to the statement. Moreover let w_{ijs} (i=1,…, n; j=1,…, k_s; s=1,...S) be the importance given by the *i-th* respondent to the *j-th* statement for the *s-th* setting. In this case the categories for the importance for a statement are "not at all" =0.25; "a little" =0.5; "quite a bit" =0.75; "very much" =1.

An indicator on the *j-th* statement for the *s-th* setting given by the *i-th* respondent is given by

$$u_{ijs} = x_{ijs} \cdot w_{ijs} \qquad \text{(I)}$$

The u_{ijs} takes values in [0.00025; 1]. For each value of u_{ijs}, it is possible to find the correct combination of x_{ijs} and w_{ijs}.

The questionnaire includes a section with structural questions exploring the current state of the disease, personal evaluation about the psychological state and the type of assistance received, geographical and demographic characteristics. This information completes the patient profile and can be used for further analysis and stratification.

For the *i-th* respondent, it is possible to create an uneasiness score for the *s-th* setting as

$$U_{is} = \sum_{j=1}^{k_s} x_{ijs} \cdot w_{ijs} \qquad \text{(II)}$$

In (II) the statements running in the opposite direction for the *s-th* setting are reversed for the score. The U_{is} could take values in [$k_s \cdot 0.00025$; k_s].

For the *i-th* respondent, the total composite index is given by:

$$TU_i = \sum_{s=1}^{S} U_{is} \qquad \text{(III)}$$

that takes values in [$0.00025 \sum_{s=1}^{S} k_s$; $\sum_{s=1}^{S} k_s$]. The linear transformation of (III) in

$$TU_i^{10} = (10 - 1) \cdot \frac{TU_i - \sum_{s=1}^{S} k_s \cdot 0.00025}{\sum_{s=1}^{S} k_s (1 - 0.00025)} + 1 \qquad \text{(IV)}$$

allows that $TU_i^{10} \in [1; 10]$. The TU_i^{10} represents the synthetic patient-centric QoL index. It is possible to compare TU_i^{10} respect to the *i-th* respondent to the score of the quality of life of the *i-th* respondent *QoLi*.

3 Conclusions

With this pilot study we suggest a methodology to set up a questionnaire for the identification of a synthetic index that allows the evaluation of the overall QoL of patients, regardless of the clinical data. The index enhances the patients' awareness of their subjective experience with the disease and enables them to better present their situation to the clinicians. This methodology can be considered in light of the idea of improving patient engagement as highlighted by the EUPATI PARADIGM project (P. Spindler & Lima, B. S., 2018). This methodology needs to be further validated through administration to patients suffering from different pathologies, and compared to the methodologies already available from international sources. An index directly generated by the patients can provide a descriptive model helpful not only to patients, but also to clinicians and third parties, that can be further integrated with clinical details to obtain an overall view of the course of treatment for each patient.

References

Asadi-Lari, M., Tamburini, M., & Gray, D. (2004). Patients' needs, satisfaction, and health related quality of life: towards a comprehensive model. *Health Qual Life Outcomes*, **2**, pp.32.

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. *PLoS One*, **6**(6), pp.e20476.

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. *J Clin Epidemiol*, **67**(4), pp.401-409.

Marbach, G. M., C., & Rizzi, A. (1991). *Le previsioni. Fondamenti logici e basi statistiche*. ETASLIBRI.

Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F., Askham, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development. *Health Technol Assess*, **2**(3), pp.i-iv, 1-88.

Patrick, D. L., & Deyo, R. A. (1989). Generic and disease-specific measures in assessing health status and quality of life. *Med Care*, **27**(3 Suppl), pp.S217-232.

Pietersma, S., de Vries, M., & van den Akker-van Marle, M. E. (2014). Domains of quality of life: results of a three-stage Delphi consensus procedure among patients, family of patients, clinicians, scientists and the general public. *Qual Life Res*, **23**(5), pp.1543-1556.

Rabin, R., & de Charro, F. (2001). EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*, **33**(5), pp.337-343.

Spindler, P., & Lima, B. S. (2018). Editorial: The European Patients Academy on Therapeutic Innovation (EUPATI) Guidelines on Patient Involvement in Research and Development. *Front Med (Lausanne)*, **5**, pp.310.

The International Society for Quality of Life Research. *What is QOL?* Retrieved 27 May from https://www.isoqol.org/what-is-qol/

Trevelyan EG, R. N. (2015). Delphi methodology in health research: how to do it? *Eurpean Journal of Integrative Medicine*, **7**, pp.423-428.

Ware, J. E., Jr., Gandek, B., Guyer, R., & Deng, N. (2016). Standardizing disease-specific quality of life measures across multiple chronic conditions: development and initial evaluation of the QOL Disease Impact Scale (QDIS(R)). *Health Qual Life Outcomes*, **14**, pp.84.

Warner, K., See, W., Haerry, D., Klingmann, I., Hunter, A., & May, M. (2018). EUPATI Guidance for Patient Involvement in Medicines Research and Development (R&D); Guidance for Pharmaceutical Industry-Led Medicines R&D. *Front Med (Lausanne)*, **5**, pp.270.

Measuring the impact of healthcare indicators on academic medical centers' scientific production

Corrado Cuccurullo, Luca D'Aniello, Massimo Aria, Maria Spano

1. Introduction

The Academic Health Centers (AHCs), also known as Academic Health Science Centers (AHSCs) or Academic Medical Centers (AMCs) are hospitals where the activities of scientific research, teaching, and patients care are fully integrated. These complex institutions pursue a triple mission: research, teaching, and care, having an enormous impact on society and the nation's health.

Recently, policymakers and practitioners give more and more great importance to the AMCs' scientific activity for both welfare and Country competitiveness. However, there is no commonly agreed definition of AMCs because their structure and composition are different from the context in which an AMC is located. Indeed, some scholars comment *"when you have seen one Academic Health Centre, you've seen one Academic Health Centre"* (Sanfilippo, 2009). AMC structural and operational characteristics could affect their scientific production and impact. These factors are the scope of services, the location, the size, the market and so on.

Our study aims to investigate and determine which are the possible factors impacting the research productivity and impact of AMCs. We develop a model to assess the academic value of AMCs by considering these factors and how they are related to healthcare performance, measured in terms of scientific productivity, impact, and growth. We focus our research on Italian public-owned AMCs - that is 20 public AMCs as "Aziende Ospedaliere Universitarie", 9 public AMCs as "Ex Policlinici Universitari a gestione diretta", 23 public-owned "Istituti di Ricovero e Cura a Carattere Scientifico" (IRCCS) (Ministry of Health - *www.dati.salute.gov.it*). We retrieve structural information mainly from AMC websites and research data from bibliographic indexing databases (e.g. Web of Science, PubMed) in the period 2010-2019.

Our analysis is articulated in two steps. First, we identify different groups of AMCs by applying a Hierarchical Cluster Analysis (HCA). These groups share common structural and operational characteristics. Second also test the presence statistically significant differences in terms of research productivity and impact among the resulting groups through the Analysis of Variance (ANOVA). Any group is a peculiar AMC configuration.

Corrado Cuccurullo, University of Campania Luigi Vanvitelli, Italy, corrado.cuccurullo@unicampania.it, 0000-0002-7401-8575
Luca D'Aniello, University of Campania Luigi Vanvitelli, Italy, lucadaniello94@gmail.com, 0000-0003-1019-9212
Massimo Aria, University of Naples Federico II, Italy, massimo.aria@unina.it, 0000-0002-8517-9411
Maria Spano, University of Naples Federico II, Italy, maria.spano@unina.it, 0000-0002-3103-2342

2. Data and methodology

As we said above, AMC could be very different in terms of different structural and operational characteristics. Aiming at covering these several aspects we collect data from different sources, from the official websites of AMCs, from official documents published by the Italian Ministry of Health, and on Google Maps. Table 1 provides an overview of the variables considered in our study. The table shows the variable synthetic label, how it was encoded, and the source of data. It is worth noting that some variables (e.g. *Type of AMC, Geographical localization*) do not change value during the years, but some others (e.g. *Structure Dimension*) have changed the value in the reference period 2010-2019.

Table 1 Main structural and operational characteristics of Italian public-owned AMCs

Descriptors	Labels	Sources
Type of AMC (AMC)	AOU - Azienda Ospedaliera Universitaria (1); AOU_SSN - Ex Policlinici Universitari a gestione diretta (2); IRCCS - Istituti di Ricovero e Cura a Carattere Scientifico (3)	*http://www.dati.salute.gov.it/dati/dettaglioDataset.jsp?menu=dati&idPag=68*
Geographical localization (GEO_LOC)	Metropolitan areas (1); Other (0)	*https://temi.camera.it/leg18/temi/tl18_province-1.html*
Buildings typologies (LAYOUT)	Pavillion (1); Monoblock (0)	*https://www.google.com/maps*
Emergency Department (ED)	Presence (1); Absense (0)	*https://www.salute.gov.it/portale/documentazione/p6_2_8_1_1.jsp?lingua=italiano&id=17*
Service mix (S_mix)	Generic (1); Specialized Hospital (0)	*http://www.dati.salute.gov.it/dati/dettaglioDataset.jsp?menu=dati&idPag=96*
Structure Dimension	As a proxy of AMC dimension we consider 4 quantitative variables measured in the reference period (2010-2019): minimun (MIN_PL) and maximum n° of beds (MAX_PL); minimum (MIN_REP) and maximum (MAX_REP) n° of hospital wards.	*http://www.dati.salute.gov.it/dati/dettaglioDataset.jsp?menu=dati&idPag=96*
Type of care organization (ORG)	Division by pathology/organ or intensity of care/mixed (i.e. related to the patient (severe, chronic)) (1); Division by medical specialties (e.g. surgery, urology, orthopedics) (0)	*Official websites of AMCs*
Hospital turnaround plans (PAR)	Yes (1); No (0)	*https://www.cergas.unibocconi.eu/sites/default/files/files/Capitolo-17.pdf*
Regional Health System turnaround plans (SSR)	Yes (1); No (0)	*https://www.salute.gov.it/portale/pianiRientro/dettaglioContenutiPianiRientro.jsp?lingua=italiano&id=5022&area=pianiRientro&menu=vuoto ; (Ferrè et al. 2012)*

We carried out a HCA to identify homogenous groups of AMCs by minimizing their distance within groups (clusters) and, at the same time, maximizing distance among groups. HCA is a multivariate technique that allows the visualization of the association structure among statistical observations at different levels of granularity.

We choose an agglomerative algorithm where each observation is initially considered as a single-element cluster. At each step of the agglomerative procedure, the two clusters that are the most similar are combined into a new bigger cluster, using a specific linkage criterion. This procedure is iterated until all observations are in a single cluster. The different solutions are sequentially nested and displayed in a tree structure, known as a dendrogram. Here, we used the Ward linkage algorithm (Ward, 1963) with the Gower's distance (Gower, 1971), the most popular distance for mixed-type variables.

Regarding the research activity of AMCs, we retrieved on Web of Science (WoS) indexing database – launched by the Institute for Scientific Information (ISI) and now maintained by Clarivate Analytics – all the publications from January 2010 to December 2019. To identify the publications related to each AMC, we searched by full name affiliation (e.g. "IRCCS FND MILANO" for the Fondazione IRCCS Istituto Nazionale Tumori Milano, "IRCCS Ca Granda Ospedale Maggiore Policlinico" for the Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico). We limit our search by document type and selected only Articles, Proceedings Papers, Review Articles, and Book Chapters in the English language. The records were exported into PlainText format. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) was used for the selection process of the publications (Liberati *et al.*, 2009). We used three bibliometric indicators to capture the different aspects of their research activity in terms of productivity (*n. of publications/total affiliated authors*), impact *(total citations/n. of publications)*, and the annual percentage growth rate for article publication (*percentage growth rate 2010-2019*).

Analysis of Variance (ANOVA) (Jaccard *et al.*, 1984) and Tukey's Post-hoc test were used to inspect differences among the clusters resulting from the HCA.

3. Findings and conclusion

HCA performed on the main characteristics of AMCs returned the dendrogram in Figure 1. We choose the solution into five clusters, highlighted in different color in the graphical representation. Interestingly, there is a natural separation among healthcare institutions with respect to the variable *Type of AMC*. For instance, IRCCS are almost all included in the Cluster 5 (orange) and in the Cluster 2 (green). They differ only with respect to the *SSR* and *PAR* variables, because the Cluster 5 includes all IRCCS subjected to both Regional Health system turnaround plans [*SSR=1*] and Hospital turnaround plans [*PAR=1*].

The Cluster 1 (blue) includes the 75% of AOU and a small portion of IRCCS (13%). All these AMCs are mainly characterized by a more articulated architectural structure [*LAYOUT=1*] and by the presence of an Emergency Departement [*ED=1*]. All of them are not subjected to both Regional Health system turnaround plans [*SSR=0*] and Hospital turnaround plans [*PAR=0*].

The remaining 25% of the AOUs fall within the Cluster 3 (red). They differ from the AOUs in the Cluster 1 because of their dimension. Indeed, these AMCs are all organized in a monoblok [*LAYOUT=0*] and therefore, they have on average a lower number of beds and wards. Finally, the Cluster 4 (lighblue) includes about the 80% of AOU_SSN, all of them localized in metropolitan areas [*GEO_LOC=1*], with an Emergency Departement [*ED=1*] and mainly organized in pavilions [*LAYOUT=1*].
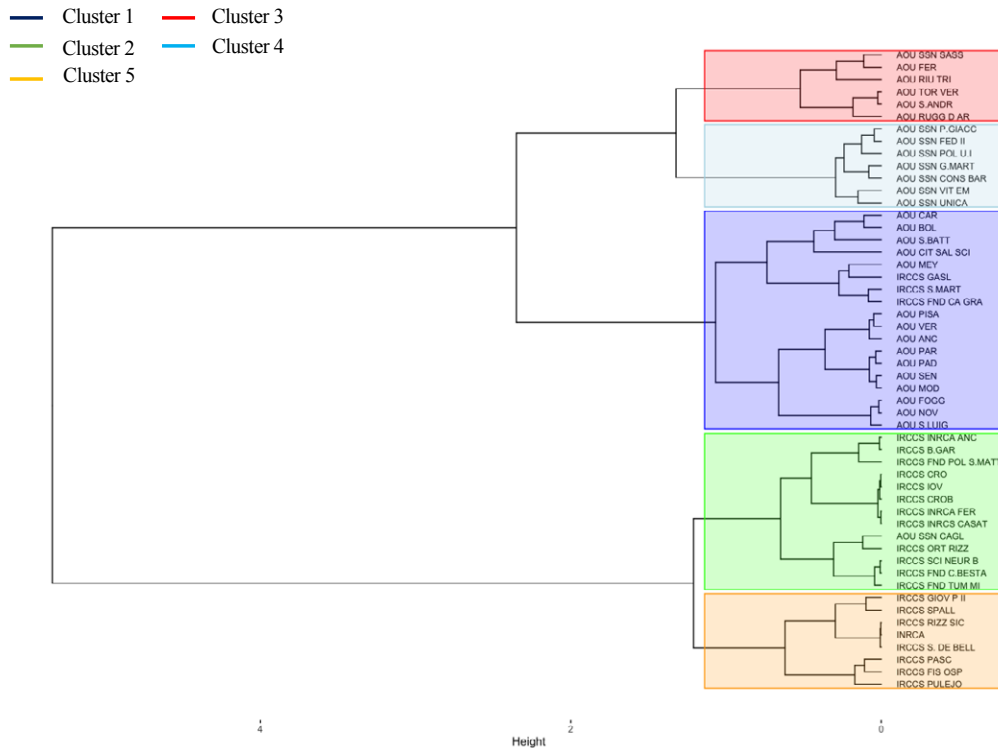
Figure 1 Dendogram resulting from HCA of Italian public-owned AMCs

Table 2 shows the results of a one-way ANOVA. We found a statistically-significant difference in average in our clusters by N. of publications per affiliated authors (F stat = 2.994, P-value = 0.0281*) and by Total citations per N. of publications (F-stat = 4.523, P-value = 0.003**) but not by Growth rate (F-stat = 0.307, P-value = 0.872).

Table 2 Mean and standard deviation by clusters and ANOVA analysis among clusters.

	N. of publications per affiliated authors		Total citations per N. of publications		Growth rate (%) 2010-2019	
	AVERAGE	SD	AVERAGE	SD	AVERAGE	SD
Cluster 1 ∗ (n=18)	2.067	1.043	22.597	3.239	0.68 (68)	1.547
Cluster 2 ∗ (n=13)	2.322	1.108	25.047	8.491	1.85 (185)	7.835
Cluster 3 ∗ (n=6)	1.518	0.431	16.120	1.882	0.11 (11)	0.411
Cluster 4 ∗ (n=7)	0.987	0.365	17.497	3.294	0.63 (63)	2.559
Cluster 5 ∗ (n=8)	2.211	0.808	24.403	5.599	0 (0)	1.259
ANOVA TEST						
F stat	2.994		4.523			0.307
P-value	0.0281*		0.003**			0.872

*P-value of the F statistic * Significant 0.01 < p-value≤ 0.05. ** Significant p-value ≤ 0.01.*

We noted that AMCs in Cluster 2, Cluster 5 and Cluster 1 including all the IRCCSs and the 75% of AOUs are more productive than the others with an average value of N. of publications per affiliated authors greater than 2. This result is reflected also on the impact of their research with an average value of total citations per N. of publications greater than 22. From these

preliminary results we could observe that the AMCs, where the research activity is regulated by strict guidelines (IRCCS) push these institutions to produce more and more with respect to AOUs and AOU_SSN where more time is probably devoted to teaching and patient care.

References

Aria, M., Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, **11**(4), pp. 959-975.

Ferrè, F., Cuccurullo, C., & Lega, F. (2012). The challenge and the future of health care turnaround plans: evidence from the Italian experience. Health Policy, 106(1), 3-9.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, **27**, pp. 623–637.

Jaccard, J., Becker, M. A., Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin*, **96**(3), 589.

Liberati A., Altman D. G., Tetzlaff J., Mulrow C., Gøtzsche P. C., Ioannidis J. P. A., Clarke M., Devereaux P. J., Kleijnen J., Moher D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, **62**(10), pp. e1-e34.

Sanfilippo F. (2009) Academic health centers. *Information, Knowledge, Systems* Management, **8**, pp. 383-414.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, **58**(301), pp. 236–244.

EGIPSS model for the evaluation of performance in healthcare

Pietro Renzi, Alberto Franci

1. Introduction

Debate about the performance of healthcare systems has been amplified by the current Covid19 pandemic. The impact of this crisis has served to highlight the fragility of many such systems and the key need for policymakers and health service managers across the world to evaluate their performance. Arguably the strategic development of any healthcare system should aim to reduce health inequalities, and therefore, as a minimum, it is necessary to monitor its performance in addressing inequalities in both health and its social determinants.

The situation in Italy is a case in point, where there are demands for better quality of care, higher productivity, better responsiveness, more efficiency and better sustainability. All of these are expressions of the same question, viz. how to improve the performance of health services and health workers? However, measuring healthcare performance presents difficulties because of its multidimensional nature, which can easily lead to conceptual and methodological confusion. As a consequence, there is a scarcity of models which fully analyse performance at healthcare system level. Unsurprisingly, virtually all current performance frameworks include quality of care as a key element, with effectiveness, productivity and efficiency also being recurrent themes. Examples include the World Health Organisation's (WHO) World Health Report 2000, the Organisation for Economic Co-operation and Development's (OECD) framework (2004), and the Nuti's framework (2008). In contrast, social outcomes of healthcare and equity are missing or little developed in most frameworks, with Australian and Canadian national frameworks being notable exceptions. Given this situation, Sicotte et al. (1999) developed the comprehensive Evaluation Globale et Intégrée de la Performance des Systèmes de Santé (EGIPPS) framework for the assessment of the performance of Health Care Organizations (HCOs). Therefore, the main aims of this paper are:

- To describe the key features of the EGIPPS framework;
- To present the authors' version of the EGIPSS framework; and
- To illustrate how it can be used, with reference to an "Area Vasta" of Italy's Marche region, to the Republic of San Marino and to other territories.

2. Key features of the EGIPSS model

In the healthcare sector, one framework stands out: the EGIPSS framework developed by Sicotte et al. (cit.), which is a comprehensive approach to the assessment of performance of HCOs. It includes **goal achievement, service production and adaptation to the environment as core dimensions of performance**, and usefully adds a focus on **values and culture**. EGIPSS is geared towards North American settings and has been mainly used in OECD countries. For example, it acts as the basis of WHO-Europe's framework for assessing hospitals, to assess accreditation schemes, to analyse how actors and stakeholders of an HCO define performance and to explore how HCOs learn.

In this paper, the authors present a practical, simplified version of the EGIPSS framework. Keeping the key strengths of this framework, some elements were redefined based on

Pietro Renzi, University of San Marino, San Marino, pietro@unirsm.sm, 0000-0001-6200-7265
Alberto Franci, University of Urbino Carlo Bo, Italy, alberto.franci@uniurb.it, 0000-0002-8157-9792

concepts of integrated healthcare systems and public service. Inspiration was found in Parsons' social system action theory to develop an integrative framework of performance, with the performance of a HCO considered to be multi-dimensional. More specifically, it is the result of the interaction between four organisational functions (see Figure 1). Consequently, the success of an organisation depends not only on how each of these functions is organised, but also on how they are aligned with each other. Performance is therefore understood as something more comprehensive than merely efficiently producing desired outputs. Furthermore, it incorporates the managerial approach of the New Public Management (NPM). The framework also describes **six equilibriums** or **alignments** between these four functions, which can be best understood as tensions that may arise between the functions as a result of a change in one of them (Figure 1).

The **tactical alignment** links the Goal Achievement and Service Production function. This deals first with the appropriateness of the service provision in relation to the goals: *"To what extent do the service production processes contribute to attaining the goals? Are they effectively producing the output needed to reach the goals?"*.

The **allocative alignment** links the Interaction with the environment and the Service Production function. It first deals with resource acquisition. Questions that can be used to assess this include: *"Are the obtained resources adequate to organise the service production function? Is the service production function optimal in relation to available resources?"*.

The **strategic alignment** examines the link between the Goals that the HCO is pursuing and its Environment. Here, questions include whether the organisational goals correspond with the needs of the population and other key actors.

The **legitimating alignment** is about the congruence of the Goal Attainment function with the Culture and Values Maintaining function, and questions how the strategic choice of goals influences and shapes the organisational values.

The **operational alignment** covers the congruence of the Culture and Values Maintaining function with the Service Production modalities, and the impact of the Service Production system on the organisational culture and values.

Finally, the **contextual alignment** between Culture and Values Maintaining function and Adaptation to the environment deals with how the social, political and cultural dimensions of the environment influence the organisational culture and its core operational values.

Figure 1 below sets out the model:

Fig. 1: EGIPSS model
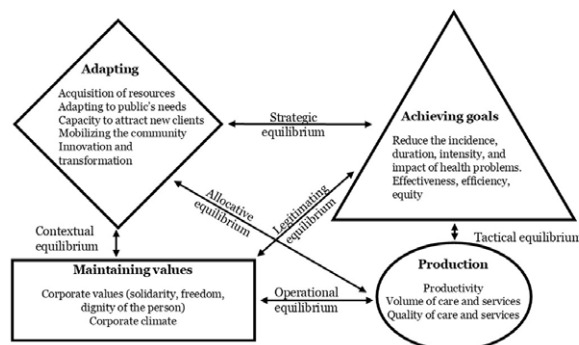
Source: Sicotte et al. (1998)

3. Materials and methods

The research made use of various statistical sources (Istituto Nazionale di Statistica (ISTAT), Centro Studi Investimenti Sociali, Osservasalute (CENSIS), Istituto Superiore di Sanità and an array of survey methods. The indicators used in the performance evaluation

model were determined through an in-depth study of the existing literature (Sicotte, cit.) and in collaboration with experts from the two locations involved in the study.

Data relating to the Republic of San Marino was provided by its Health Authority, its Istituto per la Sicurezza Sociale (ISS RSM) and by the Office of Statistics of the Republic. The data for the "AreaVasta" of the Marche Region were local and regional statistical sources, plus some internal information sources.

Once the list of indicators was identified, a "balise" (according to French terminology) or benchmarking (according to English terminology) of excellence was determined for each indicator. This represented a norm/guide/method against which results could be compared, thereby enabling opinions and judgements to be formed. The EGIPSS model and methodology incorporates performance indices that enable comparisons to be made based on excellence, which can then be weighted relative to each other within a set of dimensional and subdimensional categories. For example, the Adaptation function covers the dimension of 'Availability of resources' which has two subdimensions of 'Healthcare expenditure and financing' and 'Health workforce' (see tab.2 below). The weights used were based on the original weights provided by the model which were in turn validated by a panel of experts representing the various stakeholders of the healthcare systems studied, chosen according to their skills. This validation process incorporated the DELPHI method (Fabbris et al 2007).

One analytical issue involves establishing the relationship between an indicator and performance. The approach adopted was to determine a balise of excellence for each indicator, that involved values considered to be 'high performing'. The sense of variation in the relationship between an indicator and its associated performance can be positive, negative or parabolic. An overall performance achievement index for a subdimension and dimension is calculated by applying the assigned weights to the calculated percentage of achievement of the balise for each indicator and then aggregating the results. This can be done at each level, on the basis that if the weights are expressed as a percentage their sum within a sub-dimension, dimension and function must be equal to 100. The process can be repeated for the four functions provided by the model.

Once the percentage of achievement of the balise has been calculated, it is possible to assign a qualitative scale of performance. This serves to add precision, with the values used in this study shown in table 1:

Tab.1: Levels of performance

Level of performance	Values
Very worrying	$X < 65\%$
Worrying	$65\% \leq X \leq 75\%$
Good	$75\% \leq X < 90\%$
Excellent	$X \geq 90\%$

4. Results

This section presents a synthesis and summary of the results derived from the application of the EGIPSS model in the "Area Vasta" of the Marche region and the Republic of San Marino (the latter involves a more reduced version, due to a lack of certain data in local information systems).

Comparisons were made using the indicators relating to the Adaptation function, the Service Production function, Goal Attainment, and the Culture and Values Maintaining function, which are set out in tabs. 2, 3, 4 and 5, respectively. In addition to the above tables the authors sought a helpful diagrammatical presentation that enables the reader to judge the relative performance of different healthcare organisations in terms of 'Strategic equilibrium'. Figure 2 below illustrates this for the relationship between Infant Mortality and Healthcare Expenditure and Financing for the three healthcare organisations studied.

Tab. 2: Adaptation function

Dimension	Subdimension	Weight	Indicator	Balise Québec	ISS RSM	% Balise	Area Vasta 1 ASUR Marche	% Balise
Availability of resources	Healthcare expenditure and financing	50,0%	Total health expenditures (€/capita)	€ 2.556,67	€ 2.503,23	97,91%	€ 1.081	42,3%
		25,0%	Health expenditures for mental health (€/capita)	€ 134,67	€ 74,10	55,03%	€ 350	100,0%
		25,0%	Health expenditures for elderly care (over 65) (€/capita)	€ 1.734,00	€ 809,49	46,68%	€ 331,88	19,1%
	Total of subdimension	33,0%	**Total of subdimension**			74,4%		49,3%
	Health workforce	27,5%	Practising physicians (/1000 population)	1,05	0,56	53,69%	0,84	80,0%
Population needs	Adequacy of diagnostic tests	0,0%	Number of examinations in tomodensitometry(CT) (/1000 population)	125,70	128,34	100,00%	195,46	100,0%
		0,0%	Number of examinations in Magnetic Resonance Imaging (RMI) (/1000 population)	32,50	118,03	100,00%	118,95	100,0%
Vitality	Administrative health	100,0%	Administrative expenditure to the total health expenditure	0,05	0,04	100,00%	3,82%	100,00%

Tab. 3: Service Production function

Dimension	Subdimension	Weight	Indicator	Balise Québec	ISS RSM	% Balise	Area Vasta 1 ASUR Marche	% Balise
Accessibility	Users' perception of the accessibility and speed of services	50,0%	Patient satisfaction level with accessibility	89,0%	81,6%	91,6%	86,75%	97,47%
		50,0%	Patient satisfaction level with regard to the speed of the requested services	77,7%	72,0%	92,7%	57,30%	73,75%

Dimension	Subdimension	Weight	Indicator	Balise Québec	ISS RSM	% Balise	Area Vasta 1 ASUR Marche	% Balise
Technical quality	Relevance of care	50,0%	Caesarean sections (/100 live births)	21,9%	26,6%	82,3%	28,50%	76,84%
Humanisation of care	Perception of patients on the level of humanisation	33,3%	Patient satisfaction level regarding the environmental characteristics where care is provided	85,8%	77,6%	90,4%	91%	100%
		33,3%	Patient satisfaction level with the empathy shown towards them	87,7%	87,1%	99,3%	93,66%	100%
		33,3%	Patient satisfaction level with the ethics and professionalism shown by the staff towards them	87,7%	83,8%	95,5%	92,25%	100%

Tab. 4: Goal Attainment function

Dimension	Subdimension	Weight	Indicator	Balise Québec	ISS RSM	% Balise	Area Vasta 1 ASUR Marche	% Balise
Effectiveness	Infant mortality	33,3%	Infants underweight (percentage)	5,7%	4,1%	100,0%	2,05%	100%
		33,3%	Infant mortality (percentage)	3,3%	0,0%	100,0%	0,29%	100%
		33,3%	Mortality in newborns (percentage)	3,5%	0,7%	100,0%	0,13%	100%
	General health	50,0%	Life Years at age 65 (years)	19,9	20	100%	17,5	88%
Global Satisfaction	Population satisfaction	33,3%	Number of complaints, (/10.000 population)	29,1	10,68	100%	3,55	100%

Tab. 5: Culture and Values Maintaining function

Dimension	Subdimension	Weight	Indicator	Balise Québec	ISS RSM	% Balise	Area Vasta 1 ASUR Marche	% Balise
Quality of working life	Workplace	16,70%	Budget dedicated to training (percentage)	1,040%	0,216%	20,8%	0,60%	57,7%

Fig. 2: Strategic equilibrium
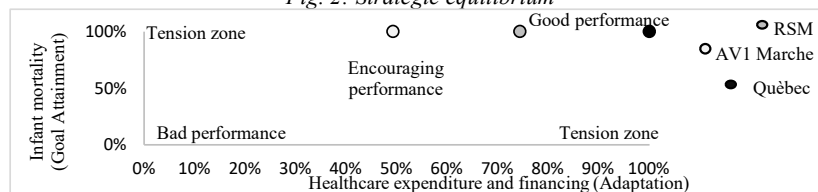
5. Weaknesses and strengths of the results obtained

The model provides an overview of the performance, especially for the "Area Vasta" of the Marche, through a prism of 107 indicators. These indicators can present warning signs regarding accessibility, technical quality, efficiency and fairness of a system.

Among the inevitable weaknesses that can affect any frameworks of this nature it should be pointed out that the evaluation of performance is equivalent to comparing the result of an indicator, of a dimension or of a sub-dimension, to a given standard. Where such standards exist they have been used. Otherwise, the performance evaluation was set against external objectives (such as those determined by Canada, the WHO and OECD) or comparisons with the similar results from other countries. The authors' choice on Canada was justified by the fact that this country, by establishing its own empirical standards of excellence, utilised its own comparisons with the EU15 countries (Vrijens et al., 2016). This approach, the only practicable one available, made it possible to position the areas studied in relation to those states that have similar healthcare systems. However, it should be noted that the interpretation of relative performance between micro-areas and States appears very delicate due to the methodological and contextual differences that can compromise the validity of the comparisons. Further constraints were the absence of relevant indicators or the lack of data in the information systems of the two areas.

6. Conclusions

Performance evaluation is a process that enables the holistic analysis of healthcare systems utilising measurable indicators. Its role, therefore, is to improve the quality of the decisions being taken by all the staff in the healthcare arena and those services that can impact on people's health. The principles and orientation of the EGIPPS model are useful for assessing healthcare systems at any level, whether it be a country, a province or even a local community.

7. References

WHO (2000), *The World Health Report 2000 - Health systems: improving performance.* Geneva: World Health Organisation.

Hurst J., Jee-Hughes M. (2001), Performance measurement and performance management in OECD health systems. In *Labour Market and Social Policy Occasional Papers*. Paris: OECD.

Sicotte C., Champagne F., Costandriopoulos A. P., (1999), La performance organisationnelle des organismes publics de santé. *Ruptures, revue transdisciplinaire en santé*, 6(1), pp 34 - 46

Fabbris L., Martini M.C. (2007), Graduates' Job Quality Dimensions According to a Delphi-Shang Experiment. In: *Effectiveness of University Education in Italy,* eds. Fabbris L., Physica-Verlag HD

Nuti S. (2008), *La valutazione della performance in sanità*, Il Mulino Editore.

Vrijens F., Renard F., Camberlin C. et al. (2016), *Performance of the Belgian health system – Report 2015*, Belgian Health Care Knowledge Centre (KCE), KCE Reports 259C.

Unsupervised spatial data mining for the development of future scenarios: a Covid-19 application

Yuri Calleo, Simone Di Zio

1. Introduction

In the framework of Future Studies, the development of future scenarios can contribute within the social context by providing inputs at the decision-making level in order to take action in the present. However, this implies an effort and a long-time frame in the first two phases of the scenario development (typically *Framing* and *Scanning*) which require a long desk research, such as reading of documents, research of the scientific literature or the consultation of experts for identifying the key factors (Bishop et al., 2007). In particular, the goal of the scanning phase is to define a number of basic driving forces which constitute the base for the construction of alternative futures scenarios. Some scholars (Kayser and Shala, 2020) estimated an average time of two weeks, in which typically the research team compose a panel aimed at understanding the object of study. Recently, with the exponential growth of social networks, users are constantly in connection with each other, disseminating textual, multimedia, and geographical content on a daily basis. It therefore follows that given the enormous increase in data sources within them and given the communication with which users share ideas, thoughts, and information, all this could be exploited in the context of scenario building.

From the premises made so far, we have developed a new approach that uses unsupervised classification models aimed at speeding up the first two phases of scenario development and optimizing the entire process. To capture the topics and the relevant key factors we used Machine Learning methods, including text-mining (Kayser and Blind, 2017) and Spatial Data Mining techniques. The goal of this work is to provide an answer to the following questions: "Is it possible to obtain information on the object of study by extracting key factors from Twitter?", "Does this approach speed up the Scanning phase?". And, above all, "What contribution can spatial data mining offer to the process of development of future scenarios?". To apply the method, we extracted a dataset from Twitter containing textual and geo-spatial content relating to Covid-19.

2. Materials and Methods

The approach used here applies unsupervised classification models belonging to Machine Learning and aims to extract the major topics within a dataset of tweets, in order to use them as key factors in the scenarios' development process. During the month of November 2020, a dataset of 60.000 tweets was extracted through the use of the Streaming API System using 95 keywords and hashtags related to the discussions on Covid-19 (Uhl and Schiebel, 2017). After extracted the matrix, we proceeded to import it into Python to clean and manipulate it, and then we applied the techniques useful for our analysis (after this phase, the remaining tweets resulted in 29.949). The first step carried out saw the conversion into numbers, better defined as "number vectors" (Atenstaedt, 2012) of the data matrix, through the "lemmatisation" and "tokenization". In the processing of a specific language, the vectors of numbers are determined by textual data, in order to reflect various linguistic properties of the text, where a coding of the characteristics is necessary (Goldberg, 2017). First of all, we tried to have a qualitative general view of our dataset by applying the text-mining technique using the bag-of-words model that extracts and flexibly represents the data of a given text describing

Yuri Calleo, Gabriele d'Annunzio University, Italy, yuricalleo@yahoo.com, 0000-0002-0190-6061
Simone Di Zio, Gabriele d'Annunzio University, Italy, s.dizio@unich.it, 0000-0002-9139-1451

the occurrence of words within a document or corpus of documents. The same extracts in a document only the words known and therefore present in a vocabulary assigned to it, while any other information is discarded a priori. We then applied a Sentiment Analysis to understand the degree of polarity of the terms found within the dataset, using two distinct algorithms, called *Vader* and *Afinn,* in order to have a comparison between the two results obtained. We decided to use these algorithms since they are two of the most used in Sentiment Analysis for social networks (cfr. Narasamma et al. 2021; Mayor & Bietti, 2021; Tan & Guann, 2021), and compared to the others, they are able to specifically decipher abbreviations and emojis in the corpus of documents.

They are tools based on lexical rules in relation to what is published mainly on social networks (Hutto & Gilbert, 2014) using a vocabulary of words generally labelled a priori (manually) and subsequently acquired by the model based on their semantic orientation (for example they can be labelled as positive, negative, or neutral). Both algorithms assign a final score based on a sum of the valence scores of the terms in the text and normalized usually between the negative (-1) and the positive (+1) extremes (Huang et al. 2019).

After having a general view of the dataset, in order to understand the most cited terms and their polarity, we used topic modelling to extract possible topics and keywords from the tweets. In this case, we used the Latent Dirichlet Allocation (LDA) (Tong and Zhang, 2016) with the following term frequency function of term t:

$$\text{tf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where $f_{t,d}$ represents the raw count of the term t in document d.

It is based on a distributive hypothesis of statistical measurement, through the extraction of a series of topics from a corpus of documents. This process is carried out through the mapping of every single document with a good part of the words present (Wang & Grimson, 2007), and the model assigns to each topic a word arrangement determining the key factors.

The LDA assumes that the topics follow a Dirichlet distribution (Minka, 2000), in fact the similarity of documents and topics is controlled by hyperparameters known as α and β; if α is low it will assign fewer topics to each document, while when α is high, we will have the opposite. A low β value will use fewer words in the topic modelling process, while a high value will use more words, thus making the topics more similar to each other. The LDA, in fact, does not know a priori the number of topics or terms to be extracted. The model produces a vector that contains the coverage of each topic for the document to be modelled: $\boldsymbol{c} = (c_1, c_2, \dots)$ where c_1 is the coverage of the first argument and so on.

To answer the research questions, we propose an analysis of georeferenced data that will optimize all process by adding important spatial information. Here we use the expression "georeferenced data" in a broad meaning, including any kind of information useful to link a tweet to a geographic object, where the object can be a unit of a vector shapefile layer (like for example a country). Numerous studies have been conducted on Twitter using text-mining or open-mining techniques (Pang & Lee 2008; Taboada et al., 2011; Liu 2012; Poria et al., 2014). Few studies, on the other hand, have focused on the construction of future scenarios starting from the extraction of georeferenced data from social networks. The spatial aspect, in our case, becomes of fundamental importance, as having a geographical view of the subject would benefit the development process.

In the scientific literature, some studies (including Haining, 2010) have highlighted the importance of georeferenced data and therefore the presence of such information in the data (if any) is worth to explore. Actually, through web mining it is not easy to extract spatial information, given that the geographic coordinates (latitude and longitude) are rarely available. The social networks themselves, while previously freely providing quantities of data relating to the positions of users, recently they try to protect themselves by not having such data extracted in a substantial way. In our case, having used a streaming API system extraction, it was possible to model them.

First, we replaced the missing values with the wording "data_2 NA", after this first step, we have obtained 20.372 tweets with a geographic information included. Subsequently, we linked each tweet to the corresponding country from which it was written by means of information on the location (e.g. village or city), so that, for example, a tweet from Paris is assigned to France. From that, we obtained the frequency distribution of the number of tweets for each country, for each topic and for each key factor permitting to calculate the discussion rate in a single topic and in a single country. These relative frequencies are then reported in a GIS software (Q-GIS Development Team, Open Source Geospatial Foundation Project. http://qgis.osgeo.org.) to create cartograms for the topics, in order to have a representation of the spatial distributions of the same.

3. Results and discussion

The results obtained fully answer the research questions. In fact, it was possible, through the use of text-mining and spatial data mining techniques to extract the influencing factors from our dataset for future scenario development. From sentiment analysis it was possible to measure the polarity of the terms within our matrix, identifying more positive words than negative ones in both algorithms. In support of this analysis, the results are shown in Table 1.

Table 1: Sentiment Analysis

Algorithm	Positive	Negative	Neutral
Vader	18261	11688	51
Afinn	21119	8830	51

The key factors were extracted through topic modelling which highlighted 5 topics with 6 related keywords (Table 2). The first topic focuses on the *health* aspects, important to understand how the existing pandemic has brought health problems, causing discomfort and death. But beyond the physical aspect, the psychological aspect has also been affected, in fact we can find the presence of the term "anxiety" within the keywords that compose our topic. However, the vaccination uncertainty that persisted in November 2020 should not be underestimated, this aspect is of fundamental importance precisely because it fuelled – and feeds – discussions and conspiracy theories (see topic 3). The second topic describes the *political* aspects, and it is worth noting that the dataset, having English as language and having been extracted during the American elections, was affected by a strong influence of the same. This perspective can be observed from the terms: "government" and "trump". The third topic is reserved for denial and *conspiracy*, as we can see from the words "forced", "reality", "protest" and "planning".

Table 2: Topic modelling

Topic	key1	key2	key3	key4	key5	key6
Health	covid	quarantine	death	pandemic	vaccine	anxiety
Politics	politics	government	trump	gates	underperformance	progressivism
Conspiracy	forced	talking	reality	protest	blaming	planning
Economy	economy	bottomed	lockdown	employee	million	recession
Society	social	distancing	app	people	black	track

The fourth topic refers us to the *economic* field, in fact governments (see keyword "recession") were suffering for the pandemic. Citizens too (see keyword "employee") – forced to close shops, companies, etc. to prevent the spread of infections, they found themselves in a difficult grip to overcome with consequences at work and personal level. Finally, the fifth topic regards the *social* context, and shows how the pandemic issue has had implications in the social structure. Social distancing adopted by governments prevented the normal development of social activities. Not only that, the spread by governments of applications aimed at tracking movements has also had a debate on social networks, specifically on possible complications and on the possible violation of citizens'

privacy. The results are shown in Table 2. After analysing the keywords for each topic, we constructed a cartogram for each of them (Figures 1-5).

Figure 1: Health

Figure 2: Politics

Figure 3: Conspiracy

Figure 4: Economy

Figure 5: Society
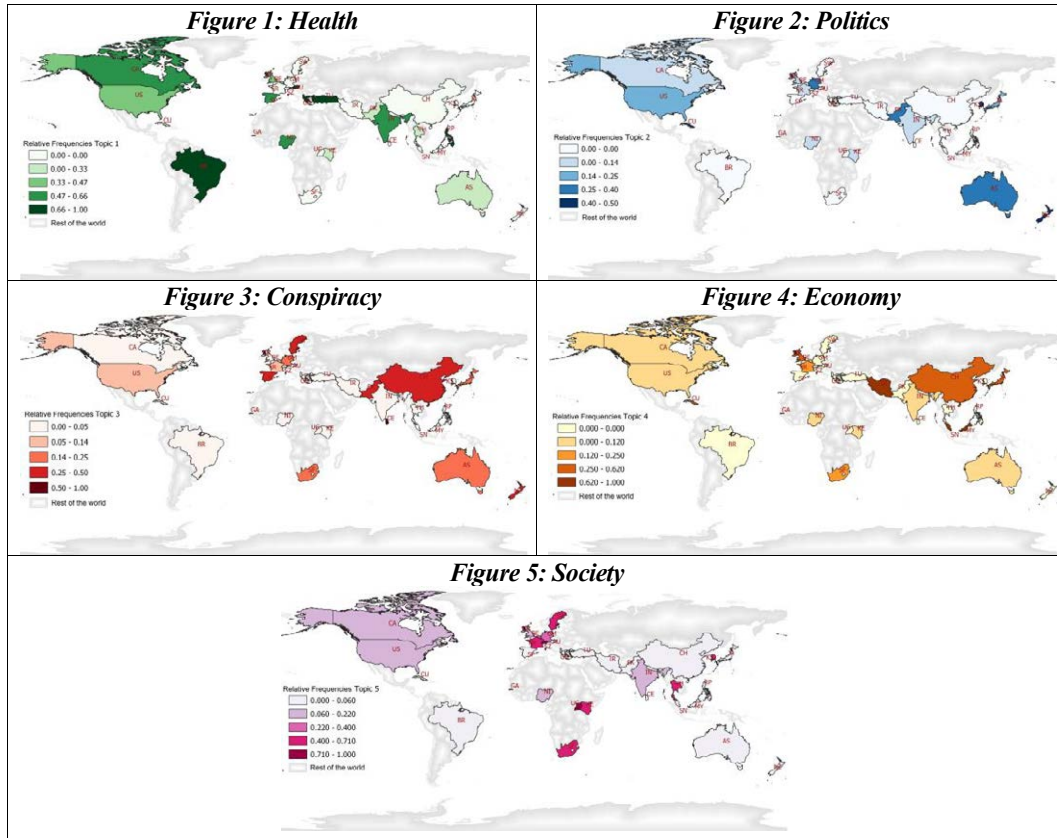
Specifically, topic 1 (fig. 1), concerning health, it was carried out in Austria, Brazil, Canada, Greece, Philippines and Turkey. These rates of discussion may be higher than in other countries because during the period studied these countries were experiencing more infections and deaths from Covid-19. As for topic 2 (fig. 2) which analyzed political discussions on Twitter, it was more discussed in American countries, Australia, Germany, South Korea and New Zeland, probably due to the political involvement of the american elections carried out in November 2020. Topic 3 (fig. 3) – which analyzed the conspiracy aspect – sees a multitude of countries involved, take for example Spain, Sri Lanka, New Zeland, China and Pakistan. China, first saw the virus appear in its territory, and subsequently it had to interface with conspiracy theories about the nature of the virus trying to disprove them rapidly. Topic 4 (fig. 4), depicting the discussion rates of the economic topic, was most discussed in Iran, China, Japan, Malaysia and United Kingdom, probably because they were particularly affected by the economic damage that has occurred resulting in a strong response from central governments. The last topic (fig. 5), depicting the discussion of topics of a social nature, finds its foundation in Singapore, Switzerland, Uganda and Sweden. A specific note must be addressed in the analysis of African territories, in fact it is possible to find a strong rate of discussion in some countries such as Nigeria, Uganda, Gambia and Kenya compared to other continents, probably due to the social problems added by the pandemic issue to those already existing.

Since the world scale does not allow to highlight all the details, especially for the smaller countries, in Figure 6 we report the five cartograms with a focus on the European region.

The analysis of georeferenced data has fully answered our research questions, given that the results can be used in the context of futures studies in order to implement the initial process of

constructing futures scenarios. This approach provides an effective tool for the development of future scenarios greatly reducing the timing of the *Framing* and *Scanning* phases. Furthermore, it provides a contribution to the future research from a statistical-spatial field and, in particular, in the field of spatial scenarios.

Starting from these results, the scenario planning process will continue with the forecasting phase (Bishop et al., 2007; Hines and Bishop, 2015), which consists of the generation of a sufficient number of alternative futures.

Figure 6: Focus on Europe area
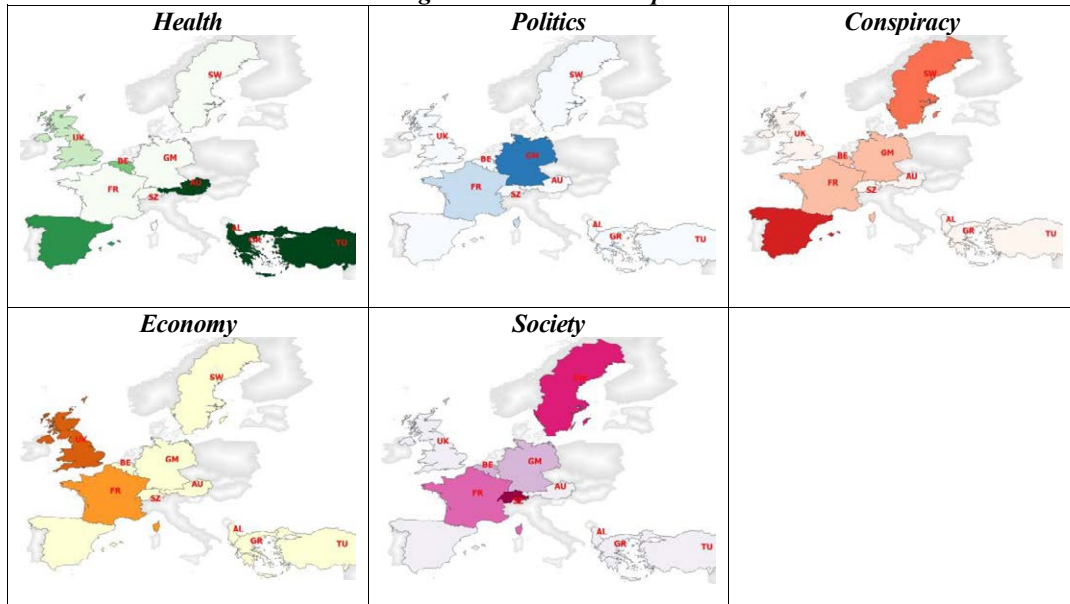
4. Concluding remarks

The approach developed above confirms the possibility of introducing the use of text-mining and spatial data mining within the first two phases of the scenario development (Framing and Scanning). It was therefore possible to extract the influencing factors in a short time frame without any literature review of the object studied and without the consultation of experts. Our study, in addition to providing elements for speeding up the process, enrich the analysis through the spatial component that offers important insights, when it is possible to observe the dynamics on geographical distributions. Understanding in which situations and in which parts of the globe a certain key factor is spoken of, means that much more information is provided. The analysis of Twitter data is only a starting point, in fact, in future studies additional social networks could also be taken into consideration (e.g., Reddit, Facebook, Instagram etc.). Furthermore, it will be possible to analyse much larger datasets in order to have a more complete vision of a given subject.

We recommend that subsequent studies focus on the spatial analysis, too often underestimated in futures studies, but capable of providing important information and, if combined with text-mining techniques, it could lead to an important turning point in the process of scenario and/or spatial scenario development.

It is worth noting that the method proposed in this paper produces spatial data that can be analyzed with the typical tools of spatial statistics. For example, a spatial autocorrelation analysis could reveal similarities between adjacent countries, even if in this case study it was not possible given the very low contiguity of the nations included in the dataset.

References

Atenstaedt, R. (2012). Word cloud analysis of the BJGP. *British Journal of General Practice*, **62**(596), pp. 148-148.

Bishop P., Hines A., Collins T. (2007). The current state of scenario development: An overview of techniques, *Foresight*, **9**(1), pp. 5–25.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, **10**(1), pp. 1-309.

Haining, R. P. (2010). The nature of georeferenced data. Handbook of applied spatial analysis. *Springer*, Berlin, Heidelberg, pp. 197-217.

Hines A., Bishop P., (2015). *Thinking about the Future: Guidelines for Strategic Foresight*, 2nd Edition, Hinesight Edition, Huston (TX).

Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, **167**, pp. 26-37.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, **8**(1).

Kayser, V., & Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. *Technological Forecasting and Social Change*, **116**, pp. 208-215.

Kayser, V., & Shala, E. (2020). Scenario development using web mining for outlining technology futures. *Technological Forecasting and Social Change*, **156**, 120086.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, **5**(1), pp. 1-167.

Mayor, E., & Bietti, L. M. (2021). Twitter, time and emotions. *Royal Society open science*, **8**(5), 201900.

Minka, T. (2000). *Estimating a Dirichlet Distribution*. MIT Technical Report, Cambridge, (US).

Narasamma, V. L., Sreedevi, M., & Kumar, G. V. (2021). Tweet Data Analysis on COVID-19 Outbreak. *Smart Technologies in Data Science and Communication*, Springer, pp. 183-193.

Pang, B., & Lee, L. (2008). Using very simple statistics for review search: An exploration. In Coling 2008. *Companion volume*: Posters, pp. 75-78.

Poria, S., Cambria, E., Winterstein, G., & Huang, G. B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, **69**, pp. 45-63.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, **37**(2), pp. 267-307.

Tan, M. J., & Guan, C. (2021). Are people happier in locations of high property value? Spatial temporal analytics of activity frequency, public sentiment and housing price using twitter data. *Applied Geography*, **132**, 102474.

Tong, Z. and Zhang, H., (2016). May. A text mining research based on LDA topic modelling. In *International Conference on Computer Science, Engineering and Information Technology*, pp. 201-210.

Uhl, A., Kolleck, N. and Schiebel, E., (2017). Twitter data analysis as contribution to strategic foresight-The case of the EU Research Project "Foresight and Modelling for European Health Policy and Regulations" (FRESHER). *European Journal of Futures Research*, **5**(1), pp.1-16.

Wang, X., & Grimson, E. (2007). Spatial Latent Dirichlet Allocation. *NIPS*, **20**, pp. 1577-1584.

Supporting decision-makers in healthcare domain. A comparative study of two interpretative proposals for Random Forests

Massimo Aria, Corrado Cuccurullo, Agostino Gnasso

1. Introduction

Today, the availability of data is growing exponentially in all sectors, especially in the healthcare sector. Machine Learning (ML) techniques allow to analyze big data to exctrat knowledge and support healthcare activities (Miotto et al., 2018), such as models for the diagnosis of complex diseases (Dhillon and Singh, 2019), (Aria et al., 2020). Despite the use of ML is spreading in many applications, it is characterized by some limitations and disadvantages.

ML main drawback corresponds to its lack of interpretability which does not allow users to represent causal relationships and interactions between predictors and response. This leads to the inability to learn how particular decisions are made. From this problem derives the definition of the Black Box model, a highly accurate model with a large complexity that cannot be represented by a relational structure. In other words, it is not possible to visualize how it internally works.

Furthermore, the opaque nature of these models hinders application in various sectors, especially in critical ones such as healthcare. To undertake a decision-making process, having faith in a machine learning model is essential, to feel reassured when analyzing and using it.

Ribeiro et al. (2016) identify a different but at the same time-related definitions of trust: trust in a prediction and trust in a model. Trusting a prediction implies that the user will take a certain action based on it; it is important to determine this confidence given that the model will be used to make decisions think for example of the use of a decision-making process in the clinical field, the consequence of acting with absolute confidence on the predictions obtained without being able to understand how they are obtained. Having faith in a model is equivalent to evaluating the model as a whole and testing its ability to generalize with appropriate evaluation metrics. A problem that recurs in using data from real contexts is that they are often significantly different and the chosen metric may not be adequate, therefore an inspection procedure of individual predictions and their interpretations may be the optimal choice.

In this work, we pay attention to one of the most used, accurate, and performing models in Machine Learning, the Random Forest model (RF) (Breiman, 2001).

Random Forest is an evolution of Bagging which aims to reduce the variance of a statistical model, simulates the variability of data through the random extraction of bootstrap samples from a single training set, and aggregates predictions on a new record (see Breiman, 1996). Being an evolution of Bagging, Random Forest aims to obtain even more different and unrelated trees. It is known as an efficient ensemble learning model, as it ensures high predictive accuracy, flexibility, and immediacy; it is recognized as an intuitive and understandable approach to the construction process, but is also considered a Black Box model due to the large number of deep decision trees produced within it (Haddouchi and Berrado, 2019).

Massimo Aria, University of Naples Federico II, Italy, massimo.aria@unina.it, 0000-0002-8517-9411
Corrado Cuccurullo, University of Campania Luigi Vanvitelli, Italy, corrado.cuccurullo@unicampania.it, 0000-0002-7401-8575
Agostino Gnasso, University of Naples Federico II, Italy, agostino.gnasso@unina.it, 0000-0002-9220-9754

The results deriving from the use of the Random Forest are valuable. Various studies have confirmed RF effectiveness in many sectors, such as biomedical for genetic selection (Díaz-Uriarte and De Andres, 2006). Breiman et al. (2001) states that Random Forest has A + performance but, having a prediction process that is difficult to understand, evaluates an F on interpretability. This leads to Occam's dilemma (Domingos, 1998) (Domingos, 1999).

The poor interpretability has prevented the adoption of the model in some sectors where there is little or no tolerance for errors, such as healthcare and clinical context (Ahmad et al., 2018). Having set the common goal of interpretability, in recent years the scientific community has fueled considerable interest in Interpretable Machine Learning, which today is an extremely open and active research field with numerous approaches that continually emerge every year (Adadi and Berrada, 2018) (Du et al., 2019) (Guidotti et al., 2018).

This research focuses on the comparison between two approaches proposed in the literature that attempt to overcome the interpretative problem. These approaches, Node Harvest by Meinshausen (2010) and inTrees by Deng (2019), are based on a post-processing interpretation method. They are also defined as Rule Extraction (Haddouchi and Berrado, 2019) approaches as they are focused on the extraction of rule sets. Both proposals use an understandable model based on the rules extracted from a Random Forest. The general idea is to identify a representative weak model to provide the interpretation. This one is selected from the sequence of weak models generated by the ensemble procedure. In particular, Node Harvest selects the set of rules through weights that are assigned based on quadratic programming with linear inequality constraints. Performing this task manages to coincide with two objectives, such as interpretability and accuracy in prediction.

Similarly, inTrees obtain interpretable information through the extraction and processing of rules deriving from a tree ensemble sequence. The extracted rules are used for the realization of a learner, which serves to make predictions on new data.

inTrees works through a series of algorithms that, at first, extract the rules and classify them; subsequently, they carry out a pruning phase on each rule, eliminating the rules that produce background noise or that are irrelevant. Subsequently, these algorithms select a compact set of rules considered relevant and not redundant. Frequent interactions are extracted and finally, everything is summarized in a learner that will be used to make predictions on new data.

2. Comparison Study

We compare Node Harvest and inTrees on four health datasets.
Comparison analysis is performed in an empirical context, where their performance is evaluated using performance metrics. These are obtained from the output and are compared to a reference standard (Aria et al., 2021).

The metrics that evaluate the performance of predictive models, when used for classification, are based on the confusion matrix, which contains the expected and observed class labels, as well as the predicted target category and the source category, as can be seen from Table 1 which represents the structure of a 2x2 confusion matrix.

Regarding comparison, the goal is to compare these approaches through the use of different health datasets. The analysis is conducted on four binary classification health datasets. These datasets are available in the UCI Machine Learning repository. They have different characteristics (see Table 2).

Table 1: Confusion Matrix

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	*TP (True Positive)*	*FP (False Positive)*
Predicted Negative Class	*FN (False Negative)*	*TN (True Negative)*

Table 2: Main characteristics of the selected health datasets.

Datasets	Obs.	Qual. Feat.	Quant. Feat.	0/1 Response Rate	Unbalanced Response
Diabetic Retinopathy Debrecen	1151	3	16	118/120	False
EEG Eye State	14980	1	15	2375/1822	False
Cardiovascular Disease	10500	7	5	883/707	False
Pima Indians Diabetes	768	8	1	130/45	False

The analysis follows the following structure: we proceed with carrying out the random forest for each of the four datasets to obtain the performance of the standard model, in terms of the confusion matrix and prediction of the target variable; the extraction of the set of rules is carried out to investigate the paths taken by each observation, of which the most important and frequent rules of the set itself will also be shown.

Finally, the comparison of the various sets of rules obtained from the two investigated methodologies is performed. The final performance evaluation is conducted through nine parameters obtained from the confusion matrices: Accuracy, Precision, Sensitivity, Specificity, G-Mean, F1 Score, Youden's Index, Balanced Accuracy, Kappa (see Sokolova et al., García et al., Akosa).

Examples are provided of the outputs obtained from the Node Harvest and inTrees approaches. These examples derive from the analysis conducted on Pima Indians Diabetes data: Node Harvest allows you to view the set of rules through an explanatory plot, provided in figure 1, while inTrees allows easy reading through summary tables that show the most frequent rule sets, such as in the table 3.

Table 3: inTrees (STEL) on Pima Indians Diabetes: set of decision rules that are easily applicable to new data. The impRRF value measures the relative percentage decrease in the Gini index for each rule derived from the random forest. The impRRF consider the length of each rule as a proxy of its complexity.

len	freq	err	condition	pred	impRRF
3	0.279	0.307	X[,2]>129.5 & X[,3]<=102 & X[,6]>27.2	1	1
2	0.326	0.366	X[,2]>114.5 & X[,8]>28.5	1	0.301
3	0.054	0.138	X[,1]>6.5 & X[,7]>0.6 & X[,7]<=1.41	1	0.162
4	0.134	0.278	X[,2]>96 & X[,5]<=34 & X[,6]>29.8 & X[,8]>30.5	1	0.144
2	0.84	0.282	X[,2]<=165.5 & X[,3]>39	0	0.139
1	0.553	0.219	X[,8]<=30.5	0	0.092
4	0.024	0.154	X[,1]<=4.5 & X[,2]<=168.5 & X[,5]>250 & X[,6]>29.85	0	0.088
3	0.184	0.232	X[,2]>127.5 & X[,6]>31.4 & X[,8]>24.5	1	0.073
2	0.786	0.261	X[,2]<=162 & X[,6]<=40.75	0	0.071
4	0.119	0.375	X[,3]<=77 & X[,5]<=118 & X[,6]>27.55 & X[,8]>30	1	0.060

Table 4 shows the nine performance metrics calculated on the four health datasets. The highest score, for each metric, is marked in bold. First of all, the interpretative solutions

Figure 1: Rule set plot obtained from Node Harvest on Pima Indians Diabetes.

proposed by Node Harvest (NH) and inTrees (STEL) represent an understandable approximation that provides an accurate summary of Random forest structure. All datasets show accurate measures very close to the reference value, provided by RF.

Focusing on the comparison, inTrees obtained higher scores in all the analyzed datasets. In particular, for EEG Eye State and Diabetic Retinopathy Debreceen, it shows much higher classification performances. It worth to noting, Node Harvest reports higher scores of sensitivity for all datasets. Maybe, it depends on the fact that this classifier can better recognize positive observations.

3. Conclusion

InTrees represents an excellent strategy for obtaining interpretative learners from Random Forest models.
The results deriving from this methodology are just as good, considering that the simplified rules based on the STEL classifier can be implemented in any programming language.

This work is a starting point for understanding the potential of Interpretable Machine Learning, which requires the development of innovative approaches that can meet the interpretative needs of each application context, such as the healthcare framework. A more complete comparative analysis should focus on analyzing data characterized by unbalanced responses and the presence of missing data (D'Ambrosio et al., 2012), and multiclass responses.

Table 4: Summary tables on the performance metrics performed on the four health datasets.

(a) *Diabetic Retinopathy Debrecen*

	RF	NH	STEL
Accuracy	0.64	0.64	**0.70**
Balanced Accuracy	0.64	0.65	**0.71**
Kappa	0.29	0.29	**0.41**
Specifity	0.65	0.48	**0.68**
Sensitivity	0.64	**0.82**	0.74
Precision	0.63	0.59	**0.65**
G-mean	0.64	0.63	**0.71**
F1	0.64	0.68	**0.69**
Youden's Index	0.29	0.30	**0.42**

(b) *EEG Eye State*

	RF	NH	STEL
Accuracy	0.92	0.68	**0.69**
Balanced Accuracy	0.92	0.67	**0.69**
Kappa	0.84	0.34	**0.38**
Specifity	0.89	0.48	**0.65**
Sensitivity	0.94	**0.85**	0.73
Precision	0.91	0.66	**0.72**
G-mean	0.92	0.64	**0.69**
F1	0.93	**0.75**	0.73
Youden's Index	0.83	0.33	**0.38**

(c) *Cardiovascular Disease*

	RF	NH	STEL
Accuracy	0.73	0.69	**0.71**
Balanced Accuracy	0.73	0.69	**0.71**
Kappa	0.47	0.38	**0.43**
Specifity	0.69	0.55	**0.66**
Sensitivity	0.78	**0.84**	0.77
Precision	0.71	0.64	**0.70**
G-mean	0.73	0.68	**0.71**
F1	0.74	0.72	**0.73**
Youden's Index	0.47	0.38	**0.43**

(d) *Pima Indians Diabetes*

	RF	NH	STEL
Accuracy	0.71	**0.74**	0.72
Balanced Accuracy	0.67	0.67	**0.69**
Kappa	0.35	**0.39**	0.38
Specifity	0.55	0.42	**0.57**
Sensitivity	0.80	**0.93**	0.80
Precision	0.78	0.74	**0.78**
G-mean	0.66	0.62	**0.68**
F1	0.79	**0.82**	0.79
Youden's Index	0.35	0.35	**0.38**

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, **6**.

Ahmad, M. A., Eckert, C., and Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 559–560.

Akosa, J. (2017). Predictive accuracy: A misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum*, pp. 2–5.

Aria, M., Cuccurullo, C., and Gnasso, A. (2021). A comparison among interpretative proposals for random forests. *Machine Learning with Applications*.

Aria, M., D'Ambrosio, A., Iorio, C., Siciliano, R., and Cozza, V. (2020). Dynamic recursive tree-based partitioning for malignant melanoma identification in skin lesion dermoscopic images. *Statistical papers*, **61**(4).

Breiman, L. (1996). Bagging predictors. *Machine learning*, **24**(2):pp. 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1):pp. 5–32.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, **16**(3):pp. 199–231.

D'Ambrosio, A., Aria, M., and Siciliano, R. (2012). Accurate tree-based missing data imputation and data fusion within the statistical learning paradigm. *Journal of classification*, **29**(2):pp. 227–258.

Deng, H. (2019). Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, **7**(4):pp. 277–287.

Dhillon, A. and Singh, A. (2019). Machine learning in healthcare data analysis: a survey. *Journal of Biology and Today's World*, **8**(6):pp. 1–10.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, **7**(1):pp. 3.

Domingos, P. (1998). Occam's two razors: the sharp and the blunt. In *KDD*,

pp. 37–43.

Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data mining and knowledge discovery*, **3**(4):pp. 409–425.

Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, **63**(1):pp. 68–77.

García, V., Mollineda, R. A., and Sánchez, J. S. (2009). Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian conference on pattern recognition and image analysis*, pp. 441–448. Springer.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, **51**(5):pp. 1–42.

Haddouchi, M. and Berrado, A. (2019). A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pp. 1–6. IEEE.

Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, pp. 2049–2072.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, **19**(6).

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pp. 1015–1021. Springer.

Media and fake news: An analysis of citizens' attitudes toward misinformation in European countries

Mauro Ferrante, Anna Maria Parroco

1. Introduction

The rapid changes determined by the rise of Internet and by the recent development of social media in everyday life have led to profound consequences on the quantity and quality of information made available and on the mechanisms of their dissemination. Today, information is increasingly shared through decentralized mechanisms in which social media play a role as a distribution channel, thanks to tools and platforms that enable peer-to-peer sharing mechanisms (Baldacci & Pelagalli, 2017). The rapid spread of on-line misinformation is one of the most-discussed issue today and has been identified as one of the top-trends in modern societies by the World Economic Forum (2013), partly because of the link between these processes and political communication. Among the reasons behind the relevance of this phenomenon, in addition to the already mentioned process of decentralization of the information, it is possible to identify also: the loss of control by the media on the dissemination process, now increasingly determined by algorithms that decide what, when and to whom to show in an unpredictable way; the growing power of Internet giants, such as Google, Facebook, and Twitter; to mention but a few, in deciding who to allow to publish news, what news to show, to whom to show it and how to earn from this process. This because among the scope of on-line disinformation, it is possible to identify the intention of generating interaction on social media, to gain profits from advertising or to discredit someone image (Figueira & Oliveira, 2017). It is therefore important to better understand citizens' attitude and trust toward media, and eventually to identify the potential determinants of different attitudes.

Starting from these premises, the present work aims at analysing the attitude of European citizens toward fake news and disinformation. After briefly discussing the growing literature on fake news and disinformation, by virtue of the availability of micro-data from the Flash Eurobarometer survey on "Fake news and disinformation online" (European Commission, 2018), a segmentation of users is proposed according to their attitude towards different types of media. Secondly, clusters are characterized both in terms of socio-demographic characteristics and in relation to users' behaviour and opinions regarding misinformation. In consideration of the social and political relevance of misinformation, potential strategies to face with fake news and online misinformation are discussed.

2. Background

Fake news and misinformation are not new phenomena. However, starting from U.S. Presidential election in November 2016 a rapid increase in the use of the term "fake news" has been observed (Rose, 2020). Also, terms such as "post-fact" and "alternative facts" emerged in new media communications. These terms are referred to deliberate distortion of the news with the aim of having an influence in public opinion and to exasperate the internal divisions in the society (Martens et al., 2018). This determined a rise in preoccupation for fake news and for their capability in generating confusion among the public. When the term fake news is used, the reference is generally to deliberate fraudulent media products. It is indeed a more severe judgement compared to "biased news". Also, fake news is something different from on-line satire. However, except for striking situations, in many cases it is not easy to identify the border between satire and discredit

intention. Allcott and Gentzkow (2017) define fake news as "news articles that are intentionally and verifiably false and could mislead readers" (Allcott and Gentzkow, 2017, p. 213) with facts entirely false. Dentith (2016) points out that fake news is an "allegation that some story is misleading – it contains significant omissions – or even false – it is a lie – designed to deceive its intended audience" (Dentith, 2016, p. 66), with facts that may be entirely false, contain partial truths, or omissions that would undermine the real fact. Fake news and misinformation have attracted the interest of researchers and institutions in identifying the mechanisms of dissemination of fake news and, eventually, potential strategies for their identification (Shao, 2018). The use of social media to get information has even more amplified the fake news issue: the news, due to its bounce rate, is likely to be contaminated, that is to undergo considerable changes until it becomes itself a fake news. It is undisputed that nowadays mainstream media have been progressively displaced by social media as a source of information. Consequently, individuals must be able to select reliable or unreliable information.

Some studies focused on the principal factors causing fake news and they found that both micro and contextual variables act (Kim & Kim, 2020). People's attitudes and citizens perception towards fake news have been recently investigated by several authors (Reuter et al., 2019; Borges-Tiago et al., 2020; Quan-Haase et al., 2018; Dinev et al., 2009; Fletcher et al., 2018). They agree that age, education, tech-profile, and cultural and ideological differences among users are relevant variables in shaping the attitudes towards fake news and disinformation. Reuter et al. (2018), referring to a survey conducted in Germany, find that people who are younger or more educated show more ability to identify fake news, and liberal or left-wing persons are more critical; Borges-Tiago et al. (2020) show that citizens attitude towards fake news is different among European countries and report that younger and tech savvy users recognize fake news most likely than others. Quan-Haase et al. (2018), highlight the importance of information literacy characteristics and information technology skills and Dinev et al. (2009) focus on cultural dimension. Finally, Fletcher et al. (2018), in presenting the results on Italian and French attitude towards fake news, stress the relevance of policy makers, private and public companies in acting to regulate information sources. Nonetheless, few studies have assessed whether populations can be segmented according to their attitude toward media. The present work aims at filling this gap and to assess whether these segments exhibit specific characteristics, both in terms of socio-demographic profile and according to media use.

3. Data and Methods

This study uses micro-data from the European Commission Flash Eurobarometer 464 on "fake news and disinformation online" (European Commission, 2018). The survey carried out in 28 Member States in 2018 on a sample of about 26 thousand respondents interviewed via telephone, aims at exploring EU citizens awareness and attitude toward fake news and disinformation online. Detailed information on the survey, as well as the questionnaire and micro-data are made available by the European Commission through the official portal for European data: https://data.europa.eu.

With the aim of identifying the main determinants of consumer attitudes towards misinformation and fake news, in a first step clusters users have been identified in relation to their attitude toward media, for the six different media types considered (i.e. Printed newspapers and news magazines; Online newspapers and news magazines; Online social networks and messaging apps; Television; Radio; Video hosting websites and podcasts). Secondly, the degree of association of socio-demographic characteristics and of media usage with the proposed cluster is explored in order to characterize different profiles of users across European countries.

In consideration of the categorical nature of data concerning the level of trust in media k-mode clustering (Huang, 1997) has been implemented. According to this approach, let $A_1, A_2, \ldots, A_6$ the set of attributes, describing the categorical space Ω, representing the users' opinion on the different media types considered, where the domain $DOM(A_j)$ of each categorical attribute A_j is given by the three answers' categories, namely: *trust, don't trust, don't use that media type*. A categorical object $X \in \Omega$ is represented by the set of attribute-value pair for each of the set of attributes considered,

and it can be represented as a vector [$x_1, x_2, ...,x_6$]. Let X = {X$_1$, X$_2$, …, X$_n$} be the set of *n* categorical objects observed in the *n* sample units. We write X$_i$ = X$_z$ if $x_{i,j} = x_{z,j}$, for $1 \leq j \leq 6$, i.e. if two generic sample units, *i* and *z*, have the same value for any of the 6 considered attributes.

The *k*-mode algorithm is an extension of the *k*-means clustering procedure to categorical variables (Chaturvedi et al., 2001) and it aims to partition the objects into *k* groups such that the distance from objects to the assigned cluster modes is minimized. A mode of X is a vector Q = [q_1, q_2, ..., q_6] that minimises a dissimilarity measure *d*, which is computed by counting the number of mismatches in all variables (simple-matching distance). The *k*-mode algorithm works iteratively by selecting initial *k*-modes of each cluster, allocating each unit to the cluster with the nearest mode according to *d*, then retesting the dissimilarity of units against the current mode and eventually reallocating the units to the cluster with the nearest mode iteratively, until no unit has changed cluster after a full cycle test of the whole dataset (Huang, 1997). In the present work the package *KlaR* implemented in *R* software has been used for the analysis.

Having been identified the clusters which characterise our sample according to their attitude toward different types of media, a regression modeling approach was undertaken to quantify the degree of association of socio-demographic characteristics and of users' behaviour and opinions regarding misinformation with cluster membership. The covariates included in the model were: 1) *Gender,* 2) *Age,* 3) *Occupation,* 4) *Social network use (How often do you use online social networks?),* 5) *Reading and sharing attitude on social network (Do you read or share things when using social network?)* 6) *Presence of fake news and misinformation in the media (Do you come across news which misrepresent reality or are even false?)* 7) *Confidence in the ability to detect fake news (Are you confident that you are able to identify news or information that misrepresent reality or is even false?)* 8) *Perception on the danger of misinformation and fake news (Is the existence of news or information that misrepresent reality a problem in your country or for democracy in general?).* By considering the categorical nature of the response variable, a multinomial logistic regression model was implemented.

4. Results

The dataset under analysis is constituted by 26,576 respondents residing in one of the EU28 countries. As also reported by the EU Commission, at an aggregate level, most respondents tend to trust news and information they receive through radio (70%), television (66%) and printed media (63%). However, less than half (47%) trust online newspapers and magazines, and lower proportions trust video hosting websites and podcasts (27%) and online social networks and messaging apps (26%). Also, these results are consistent across all the Member States (European Commission, 2018). Before implementing the clustering algorithm, those cases containing missing data for at least one of the covariates examined for the analyses were removed. This reduced the dataset to 22,384 cases. Then, to implement *k*-mode algorithm, according to the elbow method, a number of clusters equal to 5 was fixed. Modes of each item corresponding to the attitude for the different types of media considered are reported in table 1.

The results of the *k*-mode clustering procedure highlight different segments of users according to their attitude toward different type of media. Those here called *Impatients* are constituted by users which tend to trust on online social network, radio, and television, whereas they tend to do not trust on printed or online newspapers or magazines. On the contrary, it is possible to define as *Traditionalists* those who trust mainly on traditional sources of information, such as printed or online newspapers and radio. Also, they tend not to trust on social network, television, and they do not use video-hosting websites or podcasts. A particular group of users constitute those who can be defined *Sceptics*, which tend not to trust to any type of media. A fourth group are here named as the *News buff*. They trust on media coming from printed or online newspapers, radio, and television. They are very similar to the *Traditionalists*, except for a trust on television compared to *Traditionalists* who do not. Finally, the last group can be labelled as *Credulous*. They believe in almost any type of media; the only exception being represented by video hosting websites which

generally are not used by this type of users.

Table 1. Cluster modes according to the attitude toward different types of media.

Cluster ID	Printed newspapers and news magazines	Online newspapers and news magazines	Online social networks and messaging apps	Television	Radio	Video hosting websites and podcasts
Impatients (*n* = 3024)	Don't trust	Don't trust	Trust	Trust	Trust	Don't trust
Traditionalists (*n* = 3401)	Trust	Trust	Don't trust	Don't trust	Trust	Don't know
Sceptics (*n* = 3594)	Don't trust	Don't trust	Don't trust	Don't trust	Don't trust	Don't trust
News buff (*n* = 6428)	Trust	Trust	Don't trust	Trust	Trust	Don't trust
Credulous (*n* = 5937)	Trust	Trust	Trust	Trust	Trust	Don't know

Table 2 summarizes beta-coefficients, odds ratios (OR) and related *p*-values of the multinomial logistic regression model. From an analysis of the results in Table 2, all the considered factors appear significant, although differences emerge in their effects in relation to the various clusters considered. Conditionally to the other variables, and considering the cluster of *Traditionalists* as baseline, *Gender* is significantly associated to the cluster of *Sceptics*, with a risk of being "Male" of about 1.30 higher compared to the baseline, whereas being "Female" is associated with the cluster of *Credulous* (OR=1.11). In terms of *Age*, being of an age comprised between "25 and 39 years old" and "older than 55 years old", decreases the 'risk' of belonging to the *Impatients*, compared to the other categories. Also being older than 55 years old decreases the risk of belonging to the *News buff*, which tend to be younger compared to the *Traditionalists*. Different occupation profiles characterize the various clusters. Being "manual worker" or "not worker" increases the 'risk' of belonging to the *Impatients* (OR=1.46 and 1.23, respectively); similarly, "not workers" are more likely to belong to the *Sceptics* (OR=1.23), compared to the other occupation categories. Whereas being "self employed" is negatively associated with the cluster of *Credulous* (OR=0.75). Regarding *social network use*, frequent users are associated with the *Credulous* and *News buff* clusters, whereas being a non-frequent user is associated with the *Sceptics*. A more active behaviour in terms of *reading or sharing things on social media* characterizes the *Impatients,* the *Sceptics* and the *News buff*, compared to *Traditionalists*. On the other hand, *Credulous* tend not to read or share things on social media, thus indicating a more passive behaviour. *Sceptics,* as expected, tend to come across news which they think misrepresent reality or are even false. The other clusters perceive less this risk, compared to the *Traditionalists*. Nonetheless, the *Sceptics* are less confident on their capability in identifying fake news; the same holds for the *Impatients*, whereas *News buff* and *Credulous* are more confident from this perspective. Finally, a perception of fake news and disinformation as a problem in the country or for democracy in general characterizes mainly the *News buff*. On the contrary, those who do not perceive this problem are more likely to belong to the *Credulous*.

In summary, the analysis of the segmentation results reported in Table 1, jointly with the results of the logistic regression, suggest that *Impatients* and *Credulous* seems to be those at more risk for fake news and misinformation on-line. Also, they do not perceive misinformation as a problem for democracy in general. In the case of the *Impatients* an active behaviour in terms of on-line sharing emerged, thus potentially determining an active role in the spreading of on-line misinformation, also in consideration that both groups are constituted by regular social network users.

Table 2. Multinomial regression coefficients, odds ratios (*Exp(β)*), and p-values. The baseline is the cluster of "Traditionalists" for the response variable.

Cluster	Variable	Categories	*β*	*p-value*	*Exp(β)*
Impatients	Gender (Male = Ref.)	Female	0.000	0.997	1.000
	Age (15-24= Ref.)	25 - 39	-0.280	0.022	0.756
		40 - 54	-0.204	0.082	0.815
		>=55	-0.304	0.006	0.738
	Occupation (Employees = Ref.)	Self-employed	-0.028	0.754	0.973
		Manual workers	0.376	0.000	1.456
		Not working	0.208	0.002	1.231
	Social network use (several time a month or less= Ref.)	Almost everyday	0.044	0.580	1.045
		At least once a week	-0.026	0.789	0.974
	Read, or share on social network (No = Ref.)	Yes	0.216	0.005	1.241
	News misrepresent reality? (No = Ref.)	Yes	-0.152	0.031	0.859
	Able to identify news that misrepresent reality (No=Ref.)	Yes	-0.204	0.000	0.816
	Misinformation is a problem for democracy? (No=Ref.)	Yes	-0.060	0.551	0.942
	Intercept		0.180	0.283	
Sceptics	Gender (Male = Ref.)	Female	-0.266	0.000	0.767
	Age (15-24= Ref.)	25 - 39	-0.010	0.932	0.990
		40 - 54	0.010	0.930	1.010
		>=55	-0.202	0.065	0.817
	Occupation (Employees = Ref.)	Self-employed	0.221	0.006	1.248
		Manual workers	0.129	0.222	1.138
		Not working	0.210	0.001	1.234
	Social network use (several time a month or less= Ref.)	Almost everyday	-0.142	0.062	0.868
		At least once a week	-0.219	0.020	0.803
	Read, or share on social network (No = Ref.)	Yes	0.280	0.000	1.323
	News misrepresent reality? (No = Ref.)	Yes	0.449	0.000	1.568
	Able to identify news that misrepresent reality (No=Ref.)	Yes	-0.266	0.000	0.766
	Misinformation is a problem for democracy? (No=Ref.)	Yes	-0.105	0.281	0.901
	Intercept		-0.042	0.799	
News buff	Gender (Male = Ref.)	Female	-0.068	0.122	0.935
	Age (15-24= Ref.)	25 - 39	-0.216	0.039	0.806
		40 - 54	-0.173	0.088	0.841
		>=55	-0.390	0.000	0.677
	Occupation (Employees = Ref.)	Self-employed	-0.245	0.001	0.783
		Manual workers	-0.166	0.077	0.847
		Not working	-0.312	0.000	0.732
	Social network use (several time a month or less= Ref.)	Almost everyday	0.236	0.001	1.266
		At least once a week	0.159	0.057	1.172
	Read, or share on social network (No = Ref.)	Yes	0.271	0.000	1.311
	News misrepresent reality? (No = Ref.)	Yes	-0.254	0.000	0.775
	Able to identify news that misrepresent reality (No=Ref.)	Yes	0.103	0.037	1.108
	Misinformation is a problem for democracy? (No=Ref.)	Yes	0.005	0.955	1.005
	Intercept		0.928	0.000	
Credulous	Gender (Male = Ref.)	Female	0.104	0.020	1.109
	Age (15-24= Ref.)	25 - 39	-0.041	0.719	0.960
		40 - 54	0.090	0.409	1.094
		>=55	0.156	0.128	1.169
	Occupation (Employees = Ref.)	Self-employed	-0.291	0.000	0.748
		Manual workers	0.052	0.588	1.054
		Not working	0.022	0.707	1.022
	Social network use (several time a month or less= Ref.)	Almost everyday	0.306	0.000	1.358
		At least once a week	-0.192	0.029	0.825
	Read, or share on social network (No = Ref.)	Yes	-0.141	0.035	0.868
	News misrepresent reality? (No = Ref.)	Yes	-0.772	0.000	0.462
	Able to identify news that misrepresent reality (No=Ref.)	Yes	0.281	0.000	1.325
	Misinformation is a problem for democracy? (No=Ref.)	Yes	-0.227	0.008	0.797
	Intercept		0.987	0.000	

5. Conclusion

The results of the present work show different attitudes of European citizens towards the media, and this is related not only to socio-demographic characteristics, but also to their behavior and

opinions regarding misinformation. In considering the relevance of misinformation and fake news in contemporary times, it is important to identify potential strategies for tackling misinformation. Indeed, the role of countering misinformation is the responsibility of a variety of actors. Policymakers could promote a climate of calm discussion around decision that have to be made. The media could make greater efforts to promote unbiased reporting and ensure high standards of quality. It is incumbent on public institutions to provide support and monitoring misinformation, just as social media should pay more attention to the content disseminated through their platforms, playing a role that increasingly resembles that of a publisher. But a key role is represented by education and training, to act on the side of the final recipients of information and make the effects of misinformation less dangerous.

Reflecting on the limitations of this study and future research, it was not possible to include other potentially relevant information, such as the ones regarding tech-profile and cultural background of users since no information are provided from the Eurobarometer survey. It is likely that these aspects markedly affect users' attitude toward media. Finally, the proposed clusters have not been validated in other contexts, a deeper analysis through other data sources and in relation to different geographical areas is required to investigate the validity of the proposed users' segments.

References

Allcott, H., & Gentzow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, **31**(2), pp. 211-36.

Baldacci, E. & Pelagalli, F. (2017). *Communication of statistics in post-truth society: the good, the bad and the ugly*. Luxemburg: Publications Office of the European Union. KS-TC-17-005, 2017.

Borges-Tiago, T., Tiago, F., Silva, O., Guaita Martinez, J. M., & Botella-Carrubi, D. (2020). Online users' attitudes toward fake news: Implications for brand management. *Psychology & Marketing*, **37**(9), pp. 1171-1184.

Dinev, T., Goo, J., Hu, Q., & Nam, K. (2009). User behaviour towards protective information technologies: the role of national cultural differences. *Information Systems Journal*, 19(4), 391-412.

Dentith, M. R. (2016). The problem of fake news. *Public Reason*, **8**(1-2), pp. 65-79.

European Commission (2018) *Fake news and disinformation online. Flash Eurobarometer 464*. Document available at: https://op.europa.eu/it/publication-detail/-/publication/2d79b85a-4cea-11e8-be1d-01aa75ed71a1/language-en [last accessed 07/2021]

Figueira, Á., & Oliveira, L. (2017). The current state of fake news: challenges and opportunities. *Procedia Computer Science*, **121**, pp. 817-825.

Fletcher, R., Cornia, A., Graves, L., & Nielsen, R. K. (2018). Measuring the reach of "fake news" and online disinformation in Europe. *Australasian Policing*, **10**(2), pp. 25-33.

Kim, S., & Kim, S. (2020). The Crisis of public health and infodemic: Analyzing belief structure of fake news about COVID-19 pandemic. *Sustainability*, **12**(23), 9904, pp.1-23.

Martens, B., Aguiar, L., Gomez-Herrera, E., & Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news—an economic perspective. *JRC Digit. Econ. Work. Pap*, 2, 57.

Quan-Haase, A., Williams, C., Kicevski, M., Elueze, I., & Wellman, B. (2018). Dividing the grey divide: Deconstructing myths about older adults' online activities, skills, and attitudes. *American Behavioral Scientist*, **62**(9), pp. 1207-1228.

Reuter, C., Hartwig, K., Kirchner, J., & Schlegel, N. (2019). Fake news perception in Germany: A Representative Study of People's Attitudes and Approaches to Counteract Disinformation. *Proceedings of the International Conference on Wirtschaftsinformatik*, Siegen, Germany,

Rose, J. (2020). To believe or not to believe: An epistemic exploration of fake news, truth, and the limits of knowing. *Postdigital Science and Education*, **2**(1), pp. 202-216.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, **9**(1), pp. 1-9.

Longitudinal profile of a set of biomarkers in predicting Covid-19 mortality using joint models

Matteo Di Maso, Monica Ferraroni, Pasquale Ferrante, Serena Delbue,
Federico Ambrogi

1. Introduction

In survival analysis, time-varying covariates (i.e., covariates that are repeatedly measured over time) are endogenous when (i) their measurements are directly related to the event status and (ii) when incomplete information occur at random points during the follow-up because subjects may skip schedule visits and dropout from the study (Rizopoulos, 2012). Consequently, the classical time-dependent Cox model (Therneau and Grambsch, 2000) leads to biased estimates.

In order to correctly estimate the association between a time-to-event outcome and endogenous covariates, two approaches become in widespread use. The first is the joint model (JM) for the simultaneously analysis of longitudinal and time-to-event data (Rizopoulos, 2012). In this approach, the survival sub-model (used to predict hazards for a set of time-invariant covariates) and longitudinal sub-model (used to predict time-varying covariates) are interdependent by means of a set of random effects (i.e., shared parameters). Random effects are individual-specific model terms, and their inclusion in JM provides a way of producing overall predictions. The second approach is the landmarking analysis (van Houwelingen and Putter, 2012), a more pragmatic method that avoids modelling the time-varying covariates. In this approach, the estimated effect of the time-varying covariates is based on the value at the landmark time point, after which values of time-varying covariates may change.

During the first wave of Covid-19 pandemic, physicians at Istituto Clinico di Città Studi in Milan collected a set of inflammatory biomarkers in order to understand what might be used as prognostic factors in progression and mortality of Covid-19 disease. Biomarkers were collected repeatedly over the follow-up. Furthermore, particularly in the first epidemic outbreak, physicians did not have standard clinical protocols for management of Covid-19 disease and for this reason, measurements of biomarkers were highly incomplete especially at the baseline.

The aim of the present study is twice. Using data on Covid-19 patients, we firstly evaluate the association of a single biomarker on Covid-19 mortality using JM, landmarking, and time-dependent Cox model in order to compare estimates. Second, we present JM estimates for the whole set of biomarkers collected on Covid-19 patients to evaluate their association on mortality risk.

Matteo Di Maso, University of Milan, Italy, matteo.dimaso@unimi.it, 0000-0002-6481-990X
Monica Ferraroni, University of Milan, Italy, monica.ferraroni@unimi.it, 0000-0002-4542-4996
Pasquale Ferrante, University of Milan, Italy, pasquale.ferrante@unimi.it
Serena Delbue, University of Milan, Italy, serena.delbue@unimi.it, 0000-0002-3199-9369
Federico Ambrogi, University of Milan, Italy, federico.ambrogi@unimi.it, 0000-0001-9358-011X

2. Methods

Theoretical framework of JM

According to the shared parameter approach, the JM consists of two sub-models: one to model the time-to-event outcome (survival sub-model) and the other to model the time-varying covariates (longitudinal sub-model).

The survival sub-model is a typical semi-parametric (or parametric) model for time-to-event outcome. Let T_i^* be the true event time for the i_{th} subject (with $i = 1, \ldots, N$), T_i be the observed event time, defined as the minimum of the potential right-censoring time C_i and T_i^*, i.e., $T_i = \min(T_i^*, C_i)$, and let $\delta_i = I(T_i^* \leq C_i)$ be the event indicator. Furthermore, let $m_i(t)$ be the true and unobserved value of a single time-varying covariate at time t. The (proportional) hazards model is:

$$h_i(t^*) = h_0(t^*) \cdot \exp\{\beta_j X_{ij} + \alpha m_i(t)\}$$

where $h_0(\cdot)$ denotes the baseline risk function, X_{ij} is the set of j time-invariant covariates measured at baseline for the i_{th} subject, β_j is the corresponding vector of regression coefficients, and α is the regression coefficient for the time-varying covariate, quantifying the effect of such variable to the event risk.

The longitudinal sub-model is a typical linear mixed model for longitudinal outcome. As information on the time-varying covariate are collected intermittently and with error at a set of few time points for each subject, the aim of longitudinal sub-model is to predict the complete longitudinal history (also called trajectory) of the time-varying covariate (the outcome of the longitudinal sub-model) for a set of time-invariant covariates. In particular, longitudinal sub-model is:

$$y_i(t) = m_i(t) + \varepsilon_i(t)$$

where $m_i(t) = \gamma_j X_{ij} + g_i Z_i(t)$, with $g_i \sim N(0, D)$ and $\varepsilon_i(t) \sim N(0, \sigma^2)$. The quantity $y_i(t)$ is the observed longitudinal outcome for the i_{th} subject at time t, γ_j denote the estimates for the fixed effects X_{ij} and g_i denote the estimates for the random effects $Z_i(t)$. In the shared parameter approach, the random effects are common for the longitudinal and survival sub-models.

Recently, a Bayesian approach for fitting JM was introduced. In particular, estimation of JM's parameters proceeds using Markov chain Monte Carlo (MCMC) algorithm. The posterior distribution of the model parameters is derived under the assumptions that given the shared parameter, both longitudinal and survival sub-models are assumed independent, and the longitudinal outcomes of each subject are assumed independent. In this approach, non-informative priors can be used for explorative purposes.

Landmarking analysis

The idea behind landmarking analysis is to select, for a given time point $t_{LM} = s$, all subjects alive and under follow-up at time s. In particular, landmarking involves to set s and using the value of the time-varying covariate at s as fixed covariate in a time-dependent Cox model from s onwards, in a subset of subjects at risk at s. For a generic subject i, the objective is to use part of the information of the time-varying covariate of the subject to estimate the conditional probability that the subject is still alive after a predefined time window w. More specifically, at a prediction time point s, the conditional probability that the subject is still alive at time $w + s$ conditionally on being alive at time s and conditional the history of the time-varying covariate up to s is given by:

$$\pi_i (s + w \mid s) = P\{T_i > s + w \mid T_i \geq s, m_i(s)\}$$

with $m_i(s)$ denoting the history of time-varying covariate up to s.

Data collection

Between 21 February and 19 March 2020, a total of 403 Covid-19 patients were admitted at Istituto Clinico Città Studi in Milan. Patients aged 21-100 years and 58.3% were men. Person-time at risk was computed as the time elapsed from the day of hospital admission to the day of Covid-19 death (event time), to the day of hospital discharge, or to the day of moving in other structure (right-censoring time), whichever came first. Baseline characteristics included sex and age of patients, whereas biomarkers measurements included ferritin (ng/ml), lymphocytes count, neutrophil granulocytes count, D-dimer (ng/ml), C-reactive protein (ml/l), glucose (mg/dl) and lactate dehydrogenase (LDH; U/l).

Statistical analysis

In order to compare JM, landmarking, and time-dependent Cox model estimates, ferritin was considered. In particular, logarithm of ferritin (log-ferritin) levels was used to account for the skewedness of the measurements. According to the Bayesian approach, independent and non-informative priors for the fixed effects of the longitudinal and survival sub-models (i.e., age and sex) and for the shared parameter (i.e. subject-specific predicted trajectories of log-ferritin level) were used in the JM. In addition, a natural cubic spline with 2 knots was used to model the subject specific log-ferritin trajectories through time and to model age. Two knots are generally sufficient to detect mild non-linear effects and to avoid over-parametrization of the model considering the available sample size.

In landmarking analysis, a set of landmarking time point for log-ferritin time of measurements was considered. In particular, data were analysed with s running from 3 to 20 days which corresponded to the median and the 75th centile of log-ferritin time of the first and last measurements, respectively. Prediction windows w were set at 7, 14, 21, and 28 days. Age was modelled in the same way as the JM.

In the time-dependent Cox model, observed log-ferritin levels was incorporated as time-varying covariate using a natural cubic spline with 2 knots as well as age.

In order to provide associations between biomarkers and Covid-19 mortality, a JM of each biomarker one at time with the occurrence of Covid-19 death was performed (univariable JM). Logarithmic transformation was considered for ferritin, lymphocytes (log-lymphocytes), neutrophil granulocytes (log-neutrophil granulocytes), D-dimer (log-D-dimer), and C-reactive protein (log-C-reactive protein). In the multivariable JM including all biomarkers (multivariable JM), D-dimer was excluded due to the high number (78; 19%) of patients with missing values. Assumptions for priors, biomarkers trajectories and age were the same as the JM for log-ferritin previously described.

Analyses were performed using JMbayes (Rizopoulos, 2016) and dynpred (Putter, 2015) packages in R Statistical Software, version 4.0.5 (R Core Team 2021).

3. Results

Among 403 Covid-19 patients admitted at Istituto Clinico Città Studi, 140 patients died during the follow-up. Among 263 patients survived, 99 were discharged and 164 were moved in other structures. The median of follow-up was 14 days (range: 0-78 days).

Hazard ratios (HR) and corresponding 95% confidence intervals (CI) from the (biased) time-dependent Cox model and JM for log-ferritin levels (ng/ml) were 2.10 (1.67-2.64) and 1.73 (1.38-2.20), respectively. According to landmarking analysis, the HR was 1.73 (1.25-2.38) for a prediction window of 7 days. With regards to 14, 21, and 28 prediction windows, HRs were 1.86 (1.36-2.54), 1.91 (1.40-2.60), and 1.91 (1.40-2.61), respectively.

The estimates obtained from univariable JM showed decreased level through time for expected log-ferritin according to the negative coefficients for the splines of time at measurements (table 1). Conversely, the expected level of log-ferritin increased with increasing age and men showed higher expected levels than women. The expected log-lymphocytes count increased through time, whereas it decreased with age. No association emerged between log-lymphocytes and sex. The expected log-neutrophil granulocytes count decreased through time, whereas it increased with age and men showed higher levels. Likewise, expected log-D-dimer levels decreased through time, increased with age and men had higher levels. For log-C-reactive protein, expected levels showed a mixed trend through time. In particular, levels initially decreased according to the negative coefficient for the first part of follow-up and increased thereafter. The expected log-C-reactive protein levels increased with age and men showed higher levels than women. Expected levels of glucose and LDH decreased through time, while increasing with age and men had higher levels.

In univariable JM, all biomarkers were significantly associated with Covid-19 mortality. An increase in the levels of biomarkers was associated with an increased in the mortality risk, except for lymphocytes. In particular, doubling of levels for log-lymphocytes count was associated with approximately halving mortality risk (HR=0.58; 95% CI: 0.46-0.73). The strongest associations were observed for log-neutrophil granulocytes (HR=2.87; 95% CI: 2.30-3.51 for doubling of levels), for log-C-reactive protein (HR=2.44; 95% CI: 2.01-2.97 for doubling of levels) and glucose (HR=2.89; 95% CI: 1.92-4.26 for an increase of 100 mg/dl).

The multivariable JM was estimated using data on 320 patients with 96 (30%) events (after exclusion of patients with missing values for D-dimer). For ferritin and lymphocytes there were no more evidence of association with mortality. The strength of the association was attenuated with respect to the univariable JM for log-neutrophil granulocytes (HR=1.78; 95% CI: 1.16-2.69 for doubling of levels), log-C-reactive protein (HR=1.44; 95% CI: 1.13-1.83 for doubling of levels), LDH (HR=1.28; 95% CI: 1.09-1.49 for an increase of 100 UI/l), and glucose (HR of 2.44; 95% CI: 1.28-4.26 for an increase of 100 mg/dl).

However, the strongest effect in both univariable and multivariable JM was observed for age with a HR starting to rapidly increase approximately at 60 years.

4. Conclusion

In the present work, we firstly compared HR estimates of a single time-varying covariate (log-ferritin) using different approaches. The HRs from JM and landmarking approaches were lower than that of the time-dependent Cox model. In addition, landmarking estimate for a 7-day prediction window was similar to the estimate of the JM, but it tended to increase increasing prediction window. However, landmarking estimates were lower than the time-dependent Cox model one.

Finally, the multivariable JM model showed associations between some biomarkers and Covid-19 mortality but the strong association between age and mortality risk persisted after adjusted for biomarkers considered.

References

Rizopoulos D. (2012). *Joint Models for Longitudinal and Time-to-Event Data. With Application in R.* Boca Raton: Chapman & Hall/CRC.

Therneau T., Grambsch P. (2000). *Modeling Survival Data: Extending the Cox Model.* Springer-Verlag, New York (NY).

van Houwelingen HC., Putter H. (2012). *Dynamic Prediction in Clinical Survival Analysis.* Boca Raton: Chapman & Hall/CRC.

Rizopoulos D. (2016). The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data using MCMC. *J Stat Softw.* **72**(7), pp. 1-45.

Putter H. (2015). *dynpred: Companion Package to "Dynamic Prediction in Clinical Survival Analysis".* R package version 0.1.2. https://CRAN.R-project.org/package=dynpred.

Table 1. Univariable and multivariable joint model estimates.

Variables	Univariable model		Multivariable model	
	Effect (95% CI)	p-value	Effect (95% CI)	p-value
Longitudinal process: log-ferritin (ng/ml)				
Intercept	5.18 (4.67, 5.67)	p<0.01	4.88 (4.35, 5.43)	p<0.01
ns(time in days, 2)1	-1.05 (-1.32, -0.77)	p<0.01	-1.02 (-1.70, -0.40)	p<0.01
ns(time in days, 2)2	-1.69 (-2.23, -1.08)	p<0.01	-1.93 (-3.55, -0.54)	p=0.01
Sex (male vs female)	0.53 (0.34, 0.70)	p<0.01	0.66 (0.47, 0.85)	p<0.01
ns(age in years, 2)1	2.13 (1.15, 3.06)	p<0.01	1.70 (0.70, 2.72)	p<0.01
ns(age in years, 2)2	0.20 (-0.18, 0.59)	p=0.31	0.02 (-0.39, 0.44)	p=0.95

Longitudinal process: log-lymphocytes count				
Intercept	0.68 (0.29, 1.05)	p<0.01	0.69 (0.42, 0.95)	p<0.01
ns(time in days, 2)1	1.11 (0.95, 1.27)	p<0.01	1.21 (0.88, 1.56)	p<0.01
ns(time in days, 2)2	0.52 (0.40, 0.65)	p=0.01	0.87 (0.35, 1.42)	p<0.01
Sex (male vs female)	-0.11 (-0.25, 0.03)	p=0.11	-0.15 (-0.25, -0.06)	p<0.01
ns(age in years, 2)1	-1.47 (-2.20, -0.72)	p<0.01	-1.25 (-1.74, -0.76)	p<0.01
ns(age in years, 2)2	-0.69 (-0.97, -0.39)	p<0.01	-0.51 (-0.72, -0.28)	p<0.01
Longitudinal process: log-D-dimer (ng/ml)				
Intercept	4.34 (3.62, 5.01)	p<0.01	-	-
ns(time in days, 2)1	-1.72 (-2.02, -1.39)	p<0.01	-	-
ns(time in days, 2)2	-2.99 (-3.47, -2.57)	p<0.01	-	-
Sex (male vs female)	0.35 (0.10, 0.61)	p=0.01	-	-
ns(age in years, 2)1	4.18 (2.93, 5.52)	p<0.01	-	-
ns(age in years, 2)2	1.57 (1.01, 2.09)	p<0.01	-	-
Longitudinal process: log-neutrophil granulocytes count				
Intercept	0.90 (0.59, 1.22)	p<0.01	0.86 (0.56, 1.15)	p<0.01
ns(time in days, 2)1	-1.10 (-1.41, -0.80)	p<0.01	-0.55 (-1.08, 0.05)	p=0.07
ns(time in days, 2)2	-2.88 (-3.50, -2.25)	p<0.01	-1.28 (-2.31, -0.01)	p=0.05
Sex (male vs female)	0.25 (0.14, 0.38)	p<0.01	0.29 (0.18, 0.40)	p<0.01
ns(age in years, 2)1	1.25 (0.63, 1.86)	p<0.01	0.84 (0.30, 1.39)	p<0.01
ns(age in years, 2)2	0.90 (0.64, 1.14)	p<0.01	0.55 (0.33, 0.78)	p<0.01
Longitudinal process: log-C-reactive protein (ml/l)				
Intercept	-0.18 (-0.70, 0.34)	p=0.50	-0.12 (-0.70, 0.43)	p=0.69
ns(time in days, 2)1	-4.43 (-5.30, -3.56)	p<0.01	-4.10 (-5.59, -2.52)	p<0.01
ns(time in days, 2)2	0.54 (-1.48, 2.58)	p=0.59	2.33 (-1.04, 6.10)	p=0.18
Sex (male vs female)	0.49 (0.31, 0.69)	p<0.01	0.49 (0.28, 0.70)	p<0.01
ns(age in years, 2)1	4.27 (3.23, 5.29)	p<0.01	3.45 (2.42, 4.52)	p<0.01
ns(age in years, 2)2	1.01 (0.60, 1.41)	p<0.01	0.52 (0.02, 1.00)	p=0.04
Longitudinal process: glucose/100 (mg/dl/100)				
Intercept	86.58 (83.00, 90.12)	p<0.01	77.27 (65.91, 88.76)	p<0.01
ns(time in days, 2)1	-19.15 (-27.92, -10.52)	p<0.01	-0.80 (-16.66, 14.61)	p=0.92
ns(time in days, 2)2	-10.09 (-20.68, 0.29)	p=0.06	-1.54 (-20.02, -17.47)	p=0.86
Sex (male vs female)	1.01 (0.95, 1.08)	p<0.01	12.33 (4.57, 19.78)	p<0.01
ns(age in years, 2)1	58.28 (57.94, 58.65)	p<0.01	36.10 (19.21, 53.01)	p<0.01
ns(age in years, 2)2	12.27 (12.13, 12.41)	p<0.01	2.82 (-10.77, 15.95)	p=0.67
Longitudinal process: LDH/100 (UI/100)				
Intercept	106.22 (1.40, 2.77)	p<0.01	97.42 (77.40, 116.33)	p<0.01
ns(time in days, 2)1	-26.77 (-43.74, -9.79)	p<0.01	8.24 (-10.46, 26.40)	p=0.36
ns(time in days, 2)2	-3.40 (-22.74, 16.14)	p=0.72	-7.02 (-27.95, 13.12)	p=0.49
Sex (male vs female)	23.52 (23.29, 23.73)	p<0.01	58.08 (40.70, 75.22)	p<0.01
ns(age in years, 2)1	243.14 (241.65, 243.90)	p<0.01	48.13 (27.32, 70.79)	p<0.01
ns(age in years, 2)2	81.11 (80.73, 81.49)	p<0.01	0.59 (-19.33, 19.95)	p=0.96

Variables	log-hazard (95% CI)	p-value	log-hazard (95% CI)	p-value
Time-to-event process				
Sex (male vs female)	-		0.74 (0.26, 1.20)	P<0.01
ns(age in years, 2)1	-		9.31 (3.35, 15.75)	p<0.01
ns(age in years, 2)2	-		3.82 (2.57, 5.26)	p<0.01
log-ferritin (ng/ml)	0.55 (0.33, 0.79)	p<0.01	-0.13 (-0.47, 0.22)	p=0.48
log-lymphocytes	-0.78 (-1.11, -0.44)	p<0.01	0.04 (-0.43, 0.53)	p=0.89
log-neutrophil granulocytes	1.52 (1.20, 1.81)	p<0.01	0.83 (0.21, 1.43)	p=0.01
log-C-reactive protein (ml/l)	1.29 (1.01, 1.57)	p<0.01	0.53 (0.18, 0.87)	p<0.01
glucose/100 (mg/dl/100)	1.06 (0.65, 1.45)	p<0.01	0.89 (0.25, 1.45)	p=0.01
LDH/100 (UI/l/100)	0.55 (0.46, 0.64)	p<0.01	0.25 (0.09, 0.40)	p<0.01

Assessment of agricultural productivity change at country level: A stochastic frontier approach

Alessandro Magrini

1. Introduction

Productivity growth of agriculture is widely recognized as a key resource to meet food demand of the rapidly increasing world population, thus monitoring agricultural productivity change at country level is of core importance for international decision makers. The United States Department of Agriculture (USDA) represents the reference source for agricultural productivity change estimates at country level, covering almost all countries in the world for a long and updated period (from 1961 to 2016). USDA estimates consist of yearly changes in Total Factor Productivity (TFP) based on the growth accounting method, i.e., they are obtained as the ratio between the aggregated output and the sum of the input quantities weighted by their cost shares (Caves *et al.*, 1982). Growth accounting is a widely adopted methodology to assess TFP change due to several advantages, in particular it does not require assumptions on the characteristics of the production processes and allows to consider one decision making unit at a time. However, input cost shares are often partially available, and thus they should be approximated or imputed based on several different sources, like in the case of USDA estimates (see Fuglie, 2015, Table A2), with uncontrollable consequences on the accuracy of estimates. In addition, the growth accounting method has the limitation of assuming that the decision making units operate at their optimal conditions, thus it may overestimate TFP change in presence of technical inefficiency. Frontier-based methods like Data Envelopment Analysis (DEA, Charnes *et al.*, 1978) and Stochastic Frontier Models (SFMs, Schmidt & Sickles, 1984) represent valid alternatives because, by estimating the production frontier from the sample of decision making units, they can distinguish between change in technology and change in technical efficiency, and do not require input cost shares. The main difference between DEA and SFMs is that DEA does not make any assumption on the production frontier, but it is unable to account for random shocks independent of production and, as a consequence, all the deviations from the frontier are attributed to technical inefficiency. Instead, SFMs can disentangle technical inefficiency from external shocks, but they require parametric assumptions on the production frontier. Despite their appealing properties, SFMs and DEA have been employed only in some scattered studies (see the review in Kryszak *et al.*, 2021) and, as such, the available estimates are not comparable with USDA ones.

In this paper, we apply a SFM with translog specification to the same data on agricultural output and inputs employed by USDA, and exploit the generalized Malmquist index proposed by Orea (2002) to derive country level measures of agricultural TFP change, which are then compared with USDA estimates. Our preference for SFMs over DEA relies in the opportunity to account for external shocks and to assess differences in technology across countries, interactions among inputs and the trend of returns to scale.

2. Data and methodology

In this study, we employ the same data on which USDA estimates of agricultural TFP change are based (USDA, 2019). These data are sourced to Food and Agriculture Organization (FAO)

and to International Labour Organization (ILO), and integrated by modeled estimates. The output variable is the gross agricultural production (Y, thousand US dollars, 2004–2006 average international prices), while the input variables consist of six measures: land use (X_1, rainfed cropland equivalents), labour force (X_2, economically active adults), livestock (X_3, cattle equivalents), machinery stock (X_4, 40 CV tractor equivalents), fertilizer use (X_5, tonnes of nutrients), animal feed (X_6, megacalories of metabolizable energy). The data have annual frequency in the period 1961–2016 and cover the almost totality of countries in the world. In particular, the considered countries account for more than 99.7% of FAO's global gross agricultural output. Some national data have been aggregated to create consistent political units over time (e.g., former Yugoslavia, former Czechoslovakia, Ethiopia plus Eritrea, former Soviet Union) or to avoid very small measurements (e.g., Lesser Antilles, Micronesia), for a total of 170 countries. Futher details and descriptive statistics can be found in Fuglie (2015).

Let $i = 1, \ldots, n$ denote the decision making units (countries) and $t = 1, \ldots, T$ the time points (years). Also, let $y_{i,t}$ be the output level produced by unit i at time t and $x_{i,j,t}$ the level of the j-th input ($j = 1, \ldots, p$) employed by unit i at time t. A Stochastic Frontier Model (SFM) has the following general form (Schmidt & Sickles, 1984):

$$y_{i,t} = f(\boldsymbol{x}_{i,t}; \Theta) \exp(v_{i,t} - u_{i,t}) \qquad i = 1, \ldots, n; \ t = 1, \ldots, T \tag{1}$$

where f is the production frontier, representing the maximum output level technically feasible based on a given combination of the inputs $\boldsymbol{x}_{i,t} = (x_{i,1,t}, \ldots, x_{i,j,t}, \ldots, x_{i,p,t})$ and a given technology Θ, while $v_{i,t} \in \mathbb{R}$ and $u_{i,t} \in \mathbb{R}^+$ are two random errors representing the deviation from the production frontier f due to shocks, respectively, independent of the producer and related to the production. As such, the maximum feasible output may differ from the maximum output level technically feasible due to the occurrence of either favourable or unfavourable events beyond the control of producers. Specifically, the maximum feasible output for unit i at time t is equal to $y_{i,t}^* = f(\boldsymbol{x}_{i,t}; \Theta) \exp(v_{i,t})$, thus technical efficiency is $\text{TE}_{i,t} = y_{i,t}/y_{i,t}^* = \exp(-u_{i,t})$. We employ the following translog specification for f:

$$f(\boldsymbol{x}_{i,t}; \Theta) = \exp\left(\alpha_i + \delta t + \gamma t^2 + \sum_{j=1}^{p} \beta_j \log x_{i,j,t} + \sum_{j=1}^{p} \sum_{k=i}^{p} \beta_{j,k} \log x_{i,j,t} \log x_{i,k,t} + \right.$$
$$\left. + \sum_{j=1}^{p} \lambda_j t \log x_{i,j,t} + \sum_{j=1}^{p} \eta_j t^2 \log x_{i,j,t} \right) \tag{2}$$

This formulation is identical to the most commonly adopted one in the literature (see the review in Laureti, 2006, Chapter 3, and in Magrini, 2021), with the difference that we added parameters $\eta_1, \ldots, \eta_p$ to allow output elasticities to vary in time according to a quadratic trend, rather than to a linear one. The frontier specification in (2) leads to the following SFM:

$$\log y_{i,t} = \alpha_i + \delta t + \gamma t^2 + \sum_{j=1}^{p} \beta_j \log x_{i,j,t} + \sum_{j=1}^{p} \sum_{k=i}^{p} \beta_{j,k} \log x_{i,j,t} \log x_{i,k,t} +$$
$$+ \sum_{j=1}^{p} \lambda_j t \log x_{i,j,t} + \sum_{j=1}^{p} \eta_j t^2 \log x_{i,j,t} + v_{i,t} - u_{i,t} \tag{3}$$

with $\varepsilon_{i,t} = v_{i,t} - u_{i,t}$. We complete the specification of the SFM by assuming:

$$v_{i,t} \sim_{i.i.d.} \text{N}(0, \sigma_V^2)$$
$$u_{i,t} = \phi_i t + \psi_i t^2 + U_i \qquad U_i \sim_{i.i.d.} \text{N}^+(0, \sigma^2) \tag{4}$$
$$\text{Cov}(v_{i,t}, U_i) = 0 \ \forall i, t$$

where parameters ϕ_i and ψ_i regulate the second order polynomial trend of the logarithmic technical efficiency of country i, 'i.i.d.' stands for 'independent and identically distributed', $N(\cdot)$ and $N^+(\cdot)$ denote the Normal and the half Normal distribution, respectively. This specification for $u_{i,t}$ is the same as in Battese & Coelli (1995) with the addition of the quadratic term.

In order to account for technological gaps among countries with different level of development, we specify four separate models according to the WESP 2020 classification (United Nations, 2020): 'industrialized' (28), 'transition' (22), 'developing' (42), 'least developed' (78). Before estimating the parameters, the time variable is coded as the year minus 1961, thus $t = 0, 1, \ldots, 55$, and the input variables are divided by their respective sample mean. This allows first order coefficients $\beta_1, \ldots, \beta_p$ to be interpreted as the output elasticity of each input evaluated at the sample mean and at the first time point (year 1961), and makes the output elasticity of the j-th input evaluated at the sample mean and at year s equal to $\beta_j + \lambda_j (s - 1961) + \eta_j (s - 1961)^2$.

TFP change is assessed through the generalized Malmquist index proposed by Orea (2002), which allows to account for variable returns to scale. Based on this index, the TFP change between two time points s and t ($\text{TFPC}_{s,t}$) is decomposed into technological change ($\text{TC}_{s,t}$), technical efficiency change ($\text{EC}_{s,t}$), and scale change ($\text{SC}_{s,t}$):

$$\text{TFPC}_{s,t} = \text{TC}_{s,t} \cdot \text{EC}_{s,t} \cdot \text{SC}_{s,t} \tag{5}$$

Orea (2002) showed that, under a translog production frontier, these three terms equate to:

$$
\begin{aligned}
\text{TC}_{s,t} &= \exp\left[\frac{1}{2}\left(\frac{\partial \log y_{i,s}}{\partial s} + \frac{\partial \log y_{i,t}}{\partial t}\right)\right] \\
\text{EC}_{s,t} &= \exp[\mathbb{E}(u_{i,s} \mid \varepsilon_{i,s}) - \mathbb{E}(u_{i,t} \mid \varepsilon_{i,t})] \\
\text{SC}_{s,t} &= \exp\left[\frac{1}{2}\sum_{j=1}^{p}\left(\frac{\sum_{j=1}^{p} e_{i,j,s} - 1}{\sum_{j=1}^{p} e_{i,j,s}} e_{i,j,s} + \frac{\sum_{j=1}^{p} e_{i,j,t} - 1}{\sum_{j=1}^{p} e_{i,j,t}} e_{i,j,t}\right)\frac{\log x_{i,j,t}}{\log x_{i,j,s}}\right] \\
e_{i,j,s} &= \frac{\partial \log y_{i,s}}{\partial \log x_{i,j,s}} \qquad e_{i,j,t} = \frac{\partial \log y_{i,t}}{\partial \log x_{i,j,t}}
\end{aligned}
\tag{6}
$$

3. Results

We performed maximum likelihood estimation of model (3) for each group of countries using the R package `frontier` (Coelli & Henningsen, 2020). Parameter estimates imply significant and positive output elasticities at the sample mean for the almost totality of time points in all the four models, suggesting consistency with the economic theory. Also, the quadratic component of the trend of output elasticities (parameters η_j, $j = 1, \ldots, p$) and of logarithmic technical inefficiencies (parameters ψ_i, $i = 1, \ldots, n$) are significant, respectively, for most inputs and countries, supporting the adequacy of our model formulation.

Figure 1 displays the time series of the estimated overall elasticity at the sample mean, equal to the sum of all output elasticities at the sample mean by time point. Since the overall elasticity is almost always significantly lower than one for all groups of countries, we deduce that returns to scale are decreasing (and not constant) in the considered period. Based on this result, the assumption of constant returns to scale made by many authors appears just a simplification and not a real property of the production processes of the various countries.

Based on the estimated models, we computed TFPC and its components (TC, EC and SC) with $s = t - 1$ (chained index numbers) and with $s = 1961$ (index numbers with base year 1961). Table 1 reports average annual percentage changes averaged by group of countries, while Figure 2 displays the time series of index numbers with base year 1961 for a selection of countries. We see that USDA estimates of TFP change are greater in absolute value than ours for most
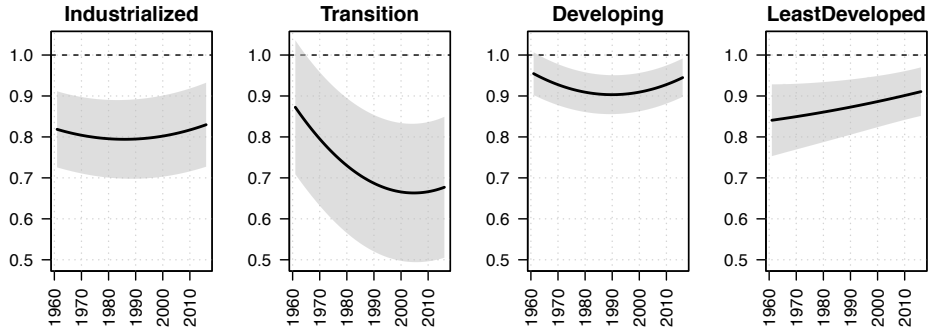
Figure 1: Time series of the estimated overall elasticity at the sample mean. Shaded areas indicate 95% confidence intervals.

Table 1: Average annual percentage variation of our and USDA's estimates of TFP change averaged by geographical region. The region 'Africa, sub-Saharan' does not include South Africa, while the region 'Oceania' does not include Australia and New Zealand.

1961–1975

Region	USDA	TFPC	TC	EC	SC
America, north	+0.56	+1.36	+1.35	+0.17	−0.16
America, central	+1.05	−0.02	+1.48	−1.31	−0.17
America, south	+1.11	+0.22	+1.26	−0.90	−0.13
Europe, north	−0.20	+0.65	+1.38	−0.74	+0.02
Europe, west	+1.14	+0.74	+1.06	−0.07	−0.26
Europe, south	+1.13	+0.26	+1.34	−0.81	−0.26
Europe, east	−0.63	−1.10	+0.93	−0.83	−1.18
Asia, west	+0.07	+0.06	+1.00	−0.72	−0.21
Asia, central	+0.10	+0.16	+1.14	−0.37	−0.59
Asia, east	+1.04	+0.10	+0.71	−0.23	−0.38
Africa, north	+2.92	+0.48	+1.15	−0.45	−0.21
Africa, sub Saharan	−0.41	+0.05	+0.81	−0.61	−0.14
South Africa	−0.65	+0.32	+0.89	−0.34	−0.22
Australia-New Zealand	+1.18	+1.02	+0.98	+0.40	−0.36
Oceania	−1.17	−0.54	+0.99	−1.54	+0.03

1991–2005

Region	USDA	TFPC	TC	EC	SC
America, north	+2.39	+1.48	+1.51	+0.08	−0.11
America, central	+0.93	+0.58	+0.96	−0.29	−0.08
America, south	+0.67	+1.11	+1.05	+0.21	−0.14
Europe, north	+1.07	+0.82	+0.59	+0.14	+0.10
Europe, west	+0.99	+0.81	+0.66	+0.01	+0.15
Europe, south	+1.33	+0.74	+0.43	+0.27	+0.05
Europe, east	+0.56	+2.24	+0.93	−0.07	+1.37
Asia, west	+2.79	+0.69	+0.97	−0.20	−0.07
Asia, central	+1.58	+1.36	+1.29	−0.20	+0.27
Asia, east	+0.90	+0.46	+0.91	−0.27	−0.18
Africa, north	+0.77	+1.49	+0.98	+0.61	−0.11
Africa, sub Saharan	+0.46	+0.61	+1.15	−0.42	−0.10
South Africa	+2.95	+1.62	+1.33	+0.25	+0.04
Australia-New Zealand	+1.77	+1.06	+1.49	−0.14	−0.27
Oceania	+0.27	−0.90	+0.54	−1.32	−0.11

1976–1990

Region	USDA	TFPC	TC	EC	SC
America, north	+1.42	+1.55	+1.41	+0.13	+0.01
America, central	+0.24	+0.24	+1.16	−0.80	−0.11
America, south	+1.17	+0.58	+1.11	−0.35	−0.17
Europe, north	+1.19	+0.88	+1.06	−0.31	+0.13
Europe, west	+1.73	+0.86	+0.91	−0.03	−0.02
Europe, south	+1.97	+0.36	+0.93	−0.27	−0.29
Europe, east	+0.36	−0.35	+0.39	−0.45	−0.28
Asia, west	+2.34	+0.08	+0.99	−0.47	−0.43
Asia, central	+0.30	+0.46	+1.02	−0.29	−0.26
Asia, east	+1.15	+0.33	+0.87	−0.26	−0.27
Africa, north	+1.94	+0.94	+1.06	+0.07	−0.19
Africa, sub Saharan	+0.11	+0.32	+0.95	−0.52	−0.10
South Africa	+2.11	+1.01	+1.10	−0.04	−0.05
Australia-New Zealand	+0.95	+1.29	+1.16	+0.13	+0.00
Oceania	−0.90	−0.86	+0.81	−1.43	−0.22

2006–2016

Region	USDA	TFPC	TC	EC	SC
America, north	+1.18	+1.73	+1.59	+0.05	+0.09
America, central	+1.03	+0.96	+0.82	+0.15	−0.02
America, south	+2.02	+1.65	+0.98	+0.70	−0.03
Europe, north	+1.82	+0.74	+0.17	+0.52	+0.05
Europe, west	+1.47	+0.47	+0.37	+0.04	+0.06
Europe, south	+0.71	+0.62	−0.20	+0.75	+0.08
Europe, east	+1.95	+1.70	+1.52	+0.27	−0.09
Asia, west	−0.07	+0.82	+0.96	+0.04	−0.17
Asia, central	+1.18	+0.77	+1.41	−0.12	−0.51
Asia, east	+1.32	+0.60	+1.08	−0.29	−0.21
Africa, north	+1.66	+2.11	+1.07	+1.09	−0.05
Africa, sub Saharan	+0.21	+0.80	+1.30	−0.33	−0.16
South Africa	+1.55	+1.99	+1.45	+0.51	+0.02
Australia-New Zealand	+0.53	+1.13	+1.60	−0.38	−0.09
Oceania	−1.36	−0.89	+0.43	−1.22	−0.09

geographical regions and periods. Exceptions include North America, Sub-Saharan Africa, East Europe, Central Asia, North Africa and Australia-New Zealand, where our estimates are greater in absolute value than USDA ones, or even discordant, for at least half the periods shown in Table 1. From TFP changes at country level, we note that our and USDA's estimates are in substantial agreement for United States, France, United Kingdom, Australia, South Africa and India, while USDA estimates are very higher than ours for Germany, Italy, Japan, China and Brazil, and moderately higher for Russian Federation and former Yugoslavia. Instead, our estimates are fairly higher than USDA ones for Canada, Afghanistan and Somalia.

The difference between our and USDA's estimates may be due to the presence of techni-
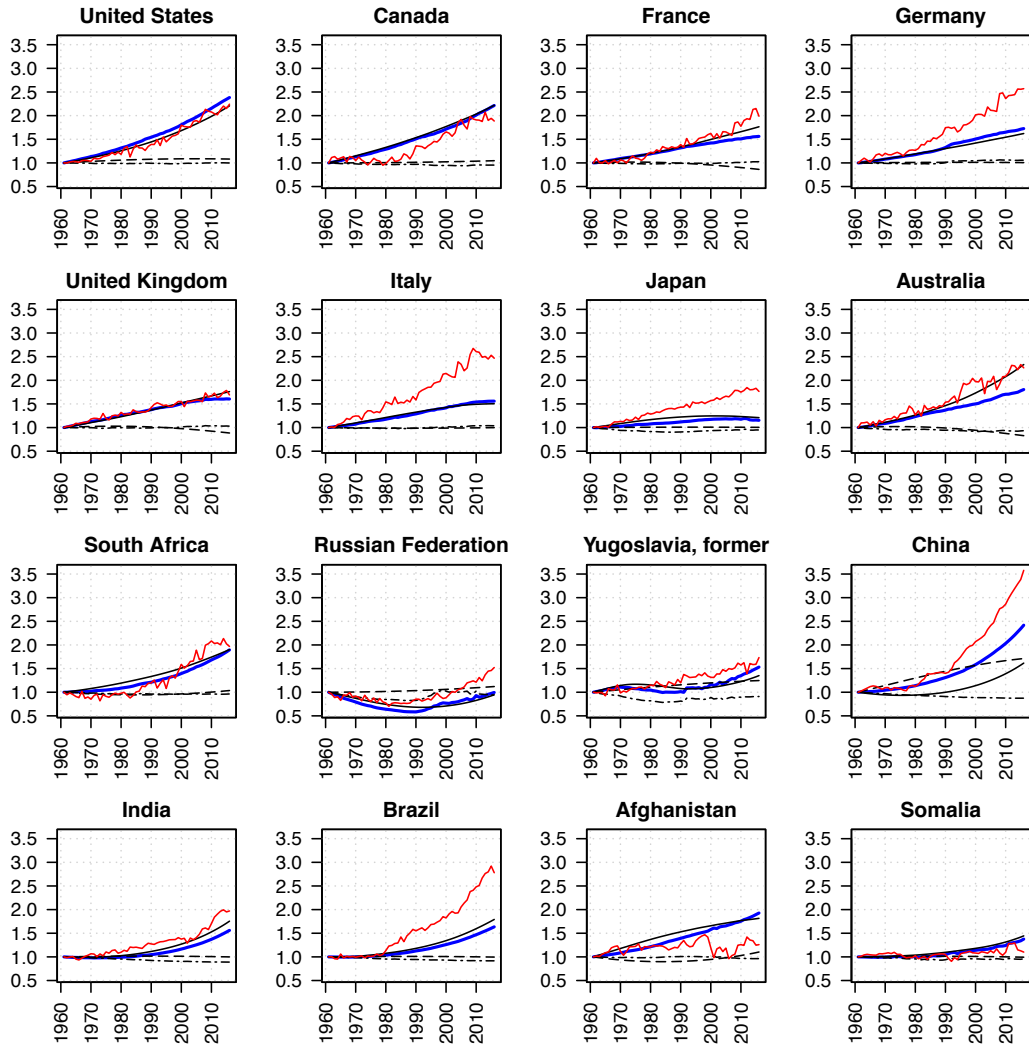
Figure 2: Our and USDA's estimates of TFP change for a selection of countries (indices, 1961=1). The time series of TFPC is shown in blue (TC, EC and SC denoted respectively by straight, dashed and dash-dotted black lines), while USDA estimates are shown in red.

cal inefficiency, that can be taken into account only by stochastic frontier models, but also to inaccuracies in USDA's input cost shares and/or in our model specification. Furthermore, our model is able to detect changes in input use through the term SC, which appears generally non-negligible, coherently with the evidence found in favour of decreasing returns to scale.

To provide an overall assessment on the agreement between our and USDA's estimates, we computed the Person correlation by country and found a median equal to 0.857, with first and third quartile equal to 0.554 and 0.943, respectively. These correlations emphasize that our and USDA's methodology provide different results but in substantial agreement, thus confirming the different theoretical foundations and suggesting the empirical validity of both of them. Full results are available at https://github.com/alessandromagrini/agrTFP.

4. Concluding remarks

We have estimated agricultural TFP change at country level based on the same data employed by the United States Department of Agriculture (USDA) using a stochastic frontier model instead of the growth accounting method. This work has the value to provide, for the first time in the literature, a comparison between agricultural TFP changes estimated with different methodologies, and an additional data source that can be employed in a large variety of longitudinal economic analyses at country level.

Our methodology overcomes the limitation of USDA estimates which rely on approximated and imputed input cost shares, and of the growth accounting method in general, which ignores technical inefficiency. However, the accuracy of estimates based on a stochastic frontier are sensitive to model specification. For this reason, we employed a more flexible specification than those adopted in the literature, but, since it is based on deterministic trends, it may be inadequate for long periods, like the one considered in this study. We also paid attention to account for heterogeneity in technology among the various countries by specifying four separate models based on the level of development.

In the future, we plan to improve our methodology by introducing autoregressive coefficients to represent stochastic trends, and by specifying latent classes to account for heterogeneity in technology among the various countries.

References

Battese, G.E., Coelli, T.J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data *Empirical Economics*, **20**: 325-332.

Charnes, A., Cooper, W.W., Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, **2**(6): 429-444.

Caves, D.W., Christensen, L.R., Diewert, W.E. (1982). The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica*, **50**: 1393-1414.

Coelli, T., Henningsen, A. (2020). frontier: stochastic frontier analysis. R package version 1.1-8. `https://CRAN.R-Project.org/package=frontier`

Fuglie, K.O. (2015). Accounting for growth in global agriculture. *Bio-based and Applied Economics*, **4**(3): 221-254.

Kryszak, L., Swierczynska, K., Staniszewski, J. (2021). Measuring total factor productivity in agriculture: A bibliometric review. *International Journal of Emerging Markets*, **16**(5).

Laureti, T. (2006). L'efficienza rispetto alla frontiera delle possibilità produttive: Modelli teorici ed analisi empiriche. Firenze University Press, Florence, IT.

Magrini, A. (2021). A stochastic frontier model to assess agricultural eco-efficiency of European countries in 1990–2019. *International Journal of Statistics and Probability*, **10**(4): 138-156.

Orea, L. (2002). Parametric decomposition of a generalized Malmquist productivity index. *Journal of Productivity Analysis*, **18**: 5-22.

Schmidt, P., Sickles, R.C. (1984). Production frontiers and panel data. *Journal of Business & Economic Statistics*, **2**(4): 299-326.

United Nations (2020). World economic situation and prospects (WESP 2020). Department of Economic and Social Affairs, United Nations. `https://www.un.org/development/desa/publications/wesp-2020.html`

USDA (2019). International agricultural productivity. Economic Research Service (ERS), United States Department of Agriculture (USDA). `https://www.ers.usda.gov/data-products/international-agricultural-productivity/`

Patient-generated evidence in Epidermolysis Bullosa (EB): Development of a questionnaire to assess the Quality of Life

Laura Benedan, May El Hachem, Carlotta Galeone, Paolo Mariani, Cinzia Pilo, Gianluca Tadini

1. Introduction

Epidermolysis Bullosa (EB) is a genetic disorder characterised by skin fragility and blistering from mild mechanical trauma. There are four major classical EB types: EB simplex, junctional EB, dystrophic EB, and Kindler EB (Has et al., 2020). All types and subtypes of EB are rare. The overall prevalence of inherited EB in the US is about 11 cases per 1 million live births, and the incidence about 20 per 1 million population (Fine, 2016). Similar results have been obtained in some European countries, Italy included (Tadini et al., 2005).

The clinical manifestations and the severity are very heterogeneous. Physical symptoms include fragile skin that blisters easily, causing pain, itch, and odour; dental problems and blisters inside the mouth and throat, dysphagia, and hair loss. This disease may also present muscle, heart, brain, gastrointestinal, bone, or kidney issues. The physical symptoms significantly impact on daily life and everyday activities and are associated with functional limitations and time-consuming medications that can severely affect the Quality of Life (QoL) of patients and their families. Besides, the disfiguring nature of these symptoms causes an additional burden at the psychological and social level, and the overall EB management may have detrimental financial consequences. The rarity of the disease is an additional issue because there is a lack of awareness and understanding by both laypeople and non-specialist healthcare professionals. Dures and colleagues (Dures, Morris, Gleeson, & Rumsey, 2011) underlined how EB patients' unmet needs were above the medical support. Informational needs, self-management, peer support, social skills and one-to-one therapy emerged as critical themes to be improved.

Considering all the implications of living with EB, a valid and reliable scale to assess the QoL of these patients is essential in patient care and management. The most used instrument available to assess EB patients' QoL, which has proven to be valid and reliable, is the QoLEB questionnaire (Frew, Martin, Nijsten, & Murrell, 2009). It was initially developed in English with an Australian sample, and it was successively translated and validated in other languages (Cestari et al., 2016; Dănescu et al., 2019; Frew, Cepeda Valdes, Fortuna, Murrell, & Salas Alanis, 2013; Yuen et al., 2014).

Even though the translation of this existing tool would have been a valid option, in the present study it was decided to conduct a Delphi study to fully understand the patients' point of view, make their voices heard, and capture possible peculiarities of the Italian context.

A three-stage online Delphi consensus procedure was conducted to identify the key domains and specific statements to assess crucial areas of EB patients' QoL.

Laura Benedan, University of Milano-Bicocca, Italy, laura.benedan@unimib.it, 0000-0003-0427-2487
May El Hachem, Bambino Gesù Children's Hospital, Italy, may.elhachem@opbg.net, 0000-0002-8145-4797
Carlotta Galeone, University of Milano-Bicocca, Italy, carlotta.galeone@statinfo.org, 0000-0003-1934-5167
Paolo Mariani, University of Milano-Bicocca, Italy, paolo.mariani@unimib.it, 0000-0002-8848-8893
Cinzia Pilo, REB Onlus Foundation, Italy, cinzia.pilo@fondazionereb.com
Gianluca Tadini, University of Milan, Italy, gtadinicmce@unimi.it, 0000-0003-1164-2802

2. Methodology

The project started from the request of the Italian no-profit association for EB research and the Italian Registry for EB Foundation (REB) to develop for the first time a patient-centred questionnaire to assess the QoL of patients affected by EB. The methodological process to develop the questionnaire consisted of two phases: firstly, a critical review of scientific literature was performed; secondly, an online pseudo-Delphi study was carried out. The Delphi method is an iterative process where several rounds are organised to identify a shared solution, with useful applications also in health research (Trevelyan & Robinson, 2015). It is a flexible method to determine the gist of the discussed problem when it is not entirely known and when it may be challenging to apply a specific statistical model. The Delphi method consists in envisaging one or more topics to a determined group of experts to provide subsequent evaluations in an iterative process aimed to reach a consensus, which will represent the final expression of the group opinion (Marbach, Mazziotta, & Rizzi, 1991). In this case, the Delphi procedure may be considered "Pseudo-Delphi" because, even though each questionnaire was anonymously analysed and summarised to be presented to the group, the discussions were open, and each participant contributed to the group discussion.

A literature review was conducted to understand what was already known about this pathology and what instruments are used at both a national and international level.

After the problem definition, the expert panel was identified. A multidisciplinary panel including patients, caregivers, and clinicians actively participated in round tables.

The team comprised:
- A Delphi master
- A moderator
- Six patients or child patients' caregivers
- Two clinicians with solid expertise in dealing with this pathology and recognised as international key opinion leaders on EB
- A psychologist

Then, a first group meeting was organised to discuss every step of the project, the main topics to cover, and the primary aim to be achieved. Successively, the patients and clinicians were asked to provide a list of spontaneously generated items to describe different areas of the EB patient's QoL. They worked separately, and all the answers were collected in an anonymous way, allowing every person to freely express their opinions and personal state of mind without any social pressure or external influence. As a result, some powerful statements appeared (e.g. *"Sometimes I think it would be better if I died"*). A total of more than 160 items were created. All answers were carefully considered and grouped within a specific domain. Accurate analysis and harmonisation of all the statements were carried out, in a first attempt to summarise the questionnaire, combine the items with the same meaning, and obtain statements that had a clear value generalisable for the entire reference population. The results were presented in the first Delphi table, i.e., a roundtable session to discuss all the implications of daily living with the disease openly. This group meeting was essential to skim the scopes and find the most salient and relevant assessment in daily practice. On this occasion, great care was taken to ensure a comprehensive and accurate understanding of the experts' points of view.

Hence, the first questionnaire (Q1) was created. This questionnaire also included some items from the literature that were not originally reported during the Delphi roundtable. The questionnaire comprised seven core domains (see Table 1) for a total of 80 items. Each participant was asked to read every statement and assess their degree of importance. They were also required to comment on the clarity and specificity of each item and write any missing information that might have been included. Each expert responded anonymously to the questionnaire and returned it to be discussed in the second Delphi round.

Table 1. EB QoL questionnaire domains

Domain	Description	Q1	Q2	Q3
Physical	This section lists relevant aspects in terms of health and physical well-being.	Items: 14 (e.g. *"I suffer from neuropathies"* was removed)	Items: 15 (the item about hands and feet problems were separated)	Items: 15 (the items about hands and feet problems were further modified)
Functional ability and autonomy	It includes statements about self-sufficiency and the ability to perform daily and routine actions.	Items: 15	Items: 13 (some items were moved to another section)	Items: 12
Psychological-emotional	It includes statements related to sensations, emotions, thoughts and feelings that may affect the psycho-emotional well-being.	Items: 13	Items: 13	Items: 14
Family	It includes statements concerning family life, such as the relationship with parents, brothers and sisters, or other people belonging to the family unit, possibly including the partner and children.	Items: 12	Items: 14 (*"some family members make me feel guilty"* was added)	Items: 14
Relational	It includes statements regarding relationships and frequent interactions with people who do not belong to the family (e.g., friends, classmates, colleagues, strangers on the street, etc.).	Items: 9	Items: 10 (*"some friends get in touch only when I'm at the hospital"* was added and later removed)	Items: 9
Work and economic	It includes statements about the work context and the economic implications of the disease.	Items: 11	Items: 13 (e.g. *"I can't work"* was separated into two items to highlight the difference between the physical impossibility and societal barriers)	Items: 13
Medical care and assistance	It includes statements regarding disease-related health care, including medical and nursing assistance.	Items: 6	Items: 8 (two items were inserted about the lack of knowledge of non-specialised clinicians in local and the need for private rooms when hospitalised).	Items: 8
		Total items: 80	Total items: 86	Total items: 85

All the answers were carefully examined, and a ranking was created for every item within each domain according to the degree of importance indicated by the participants. The results of this analysis were discussed in the group, and further refinement of the questionnaire was made. Some items were changed or rephrased for greater clarity; others were merged or removed because of their lesser importance.

A new questionnaire (Q2) was defined, considering all suggestions that emerged from the group meeting. The previously identified core domains remained unchanged, but some new items were suggested and inserted. Overall, Q2 was composed of 86 items. At this stage, each participant was asked to rate both the degree of agreement and the degree of importance of each item on a four-point Likert scale (*"Not at all"*, *"A little"*, *"Quite"*, *"Very"*). This step is necessary to remove some irrelevant statements and evaluate the order in which the items are presented. The agreement and importance measures were constructed as satisfaction-importance measures, in line with the widely used Customer Satisfaction techniques.

In addition to the abovementioned seven domains, some specific questions were inserted about the type of EB diagnosed, some socio-demographic information (e.g., age group, the Italian region of residence, the perceived need for psychological support, the perceived satisfaction of the quality of care, etc.). Finally, an overall QoL satisfaction question was asked ("On a scale from 1 to 10, how do you rate your quality of life?").

The results of this phase were presented to the group to define the questionnaire structure further and prepare the new version (Q3) with 85 items, which each participant anonymously filled in. Only one sentence was removed, and some others were modified to be more easily understandable and clear.

It should be noted that, in some cases, a different view emerged between clinicians and patients, and some information learned by the literature were then rejected or modified to be adapted to the language and the experience of the patients (e.g., the terms used to talk about some physical symptoms).

The final version of the questionnaire will be administered to a larger sample to assess its validity and reliability.

3. Conclusions

The present study is part of a more extensive research project aimed at developing a valid and reliable questionnaire to assess the QoL of EB patients. This tool is meant to grasp the point of view and the patient's subjective experience beyond clinical classifications and take into account the patient's overall experience. Starting from an initial set of areas and through the three-round pseudo-Delphi methodology, a gradual refinement of the statements was carried out, and a list of items was defined to be included in an easy-to-use but meaningful patient-centred questionnaire. Each participant had the opportunity to read and fulfil the questionnaire in private, having anonymity assured, allowing free expression of opinions without any social pressure or compliance effect that may conversely arise during the group discussions. On the other hand, knowing all information gathered from the questionnaires and discussing it in the group offered them the opportunity to critically analyse and re-consider all items and areas composing the questionnaire and achieve a final agreement among participants. From a methodological point of view, this approach is worthy in analysing real-world data pertaining to a subjective topic such as QoL, especially in rare diseases. The final patient-centred questionnaire is thus able to measure the QoL beyond the physical symptoms and the clinical evolution of the disease, encompassing functional autonomy, psycho-emotional state, social relations inside and outside the family context, the working field and several aspects of the medical care and assistance. The experts approved the final version of the questionnaire after three iterations of anonymous online questionnaire completion and related presentation and discussion of results within the group. The future steps of this

research will provide for the assessment of the psychometric properties of the questionnaire to prove its reliability and validity in measuring the QoL of EB patients. This new tool may be a valid aid for clinicians to understand patients better and identify the areas that need more attention; moreover, it may allow them to follow the patients over time and evaluate the impact of any treatments.

Acknowledgements

References

Cestari, T., Prati, C., Menegon, D. B., Prado Oliveira, Z. N., Machado, M. C. R., Dumet, J., … Murrell, D. F. (2016). Translation, cross-cultural adaptation and validation of the Quality of Life Evaluation in Epidermolysis Bullosa instrument in Brazilian Portuguese. *International Journal of Dermatology*, **55**(2), pp. e94–e99.

Dănescu, S., Sălăvăstru, C., Sendrea, A., Tiplica, S., Baican, A., Ungureanu, L., … Cosgarea, R. (2019). Correlation between disease severity and quality of life in patients with epidermolysis bullosa. *Journal of the European Academy of Dermatology and Venereology*, **33**(5), pp. e217–e219.

Dures, E., Morris, M., Gleeson, K., & Rumsey, N. (2011). The psychosocial impact of epidermolysis bullosa. *Qualitative Health Research*, **21**(6), pp.771–782.

Fine, J. D. (2016). Epidemiology of inherited epidermolysis bullosa based on incidence and prevalence estimates from the national epidermolysis bullosa registry. *JAMA Dermatology*, **152**(11), pp. 1231–1238.

Frew, J. W., Cepeda Valdes, R., Fortuna, G., Murrell, D. F., & Salas Alanis, J. (2013). Measuring quality of life in epidermolysis bullosa in Mexico: Cross-cultural validation of the Hispanic version of the Quality of Life in Epidermolysis Bullosa questionnaire. *Journal of the American Academy of Dermatology*, **69**(4), pp. 652–653.

Frew, J. W., Martin, L. K., Nijsten, T., & Murrell, D. F. (2009). Quality of life evaluation in epidermolysis bullosa (EB) through the development of the QOLEB questionnaire: An EB-specific quality of life instrument. *British Journal of Dermatology*, **161**(6), pp. 1323–1330.

Has, C., Bauer, J. W., Bodemer, C., Bolling, M. C., Bruckner-Tuderman, L., Diem, A., … Mellerio, J. E. (2020). Consensus reclassification of inherited epidermolysis bullosa and other disorders with skin fragility. *British Journal of Dermatology*, **183**(4), pp. 614–627.

Marbach, G., Mazziotta, C., & Rizzi, A. (1991). *Le previsioni. Fondamenti logici e basi statistiche*. Milano: edizioni Etaslibri.

Tadini, G., Gualandri, L., Colombi, M., Paradisi, M., Angelo, C., Zambruno, G., … Cainelli, T (2005). The Italian registry of hereditary epidermolysis bullosa. *Giornale Italiano di Dermatologia e Venereologia*, **140**(4), pp. 359-72.

Trevelyan, E. G., & Robinson, N. (2015). Delphi methodology in health research: How to do it? *European Journal of Integrative Medicine*, **7**(4), pp. 423–428.

Yuen, W. Y., Frew, J. W., Veerman, K., van den Heuvel, E. R., Murrell, D. F., & Jonkman, M. F. (2014). Health-related quality of life in epidermolysis bullosa: Validation of the dutch QOLEB questionnaire and assessment in the dutch population. *Acta Dermato-Venereologica*, **94**(4), pp. 442–447.

A Prospective Sustainability Indicator for Pension Systems

Fabrizio Culotta

1. Introduction

Nowadays, population ageing is a central topic in the political agenda of many OECD countries. It is well known that when the demographic structure of population gradually shifts towards older ages, the pressure on the financial sustainability of the welfare system raises. Health care and pension are those public systems affected the most.

Focusing on PAYG (unfunded) public pension systems, i.e. the first of the multi-pillars architecture for pension systems (Holzmann, 2005), the pressure from an older population is twofold. On the payment side, it increases the proportion of recipients and the duration. On the contribution side, instead, it decreases relative portion of contributors. This pressure can be captured by an indicator for old-age dependency ratio tracing the ratio between pension recipients and pension contributors.

From a pension system perspective, the old-age dependency ratio can be conceived as an indicator for the *extensive margin* of financial sustainability since it traces how many individuals are involved in the pension payments and contributions flow. Clearly, solely considering the extensive margin is not exhaustive. In fact, within the set of indicators for *sustainable pensions* (i.e. the second objective of the Pension strand of the Open Method of Coordination), Eurostat considers indicators for two other types of margins. In particular, the extensive margin is combined with the *intensive margin*, i.e. how much workers contribute to and pensioners receive from the public pension system (as a share of GDP). Note that, from the contribution side, the intensive margin can be further decomposed into the product between pensionable wage and pension contribution rate. Thirdly, Eurostat adopts a durational indicator for the length of post-retirement period as well as for the working life. The former traces the effective duration of pension payment, interrupted at individual level because of pensioner's death. The latter proxies the effective duration of pension contribution, often suspended in the case of unemployment (see Bravo and Herce, 2020) or even interrupted in the case of disability. These two indicators are important since they allow to trace a third dimension of the financial sustainability: for how long workers contribute and pensioners benefit. As such, they can identify a *durational margin*.

Considering all three margins not only enriches the set of dimensions to be evaluated, but it also stresses the importance of integration between labour market and pension statistics to analyse more narrowly the sustainability of public pension systems. In this sense, no sustainability indicators for pension systems follows this approach. This work tries to fill this gap by proposing a *Prospective Indicator for the Sustainability of Pension systems* (hereafter, PISP) coherent with the informative system proposed by Eurostat. A pool of European countries is considered as an application, and their ranking assessed. Finally, PISP is compared with an alternative formulation, stressing the contribution of the durational margin, as well as with a benchmark indicator.

2. Data and Methods

Data to construct PISP span over five years, namely the period 2015-2019, and they are extracted from Eurostat database. The pool of European countries is represented by Austria, Germany, Finland, France, Italy, and the Netherlands. The following statistics are selected to construct PISP for each country in the pool (Table 1).

Statistics	Unit of Measure	Area	Margin
Pension Contributions	% GDP	Labour Market	Extensive, Intensive
Pension Payments	% GDP	Pension System	Extensive, Intensive
Duration of Working Life	Years	Labour Market	Durational
Life Expectancy at 65	Years	Pension System	Durational

*Table 1: Statistics used for the construction of ISPS. Source: Eurostat.*

Once statistics are collected, the PISP is constructed as the difference of the product among margins for each flow. Let i and t index country and time, the $PISP_{i,t}$ can be defined as:

$$PISP_{i,t} = PC_{i,t} \cdot WL_{i,t} - PP_{i,t} \cdot LE_{i,t} \qquad (1)$$

where $PC_{i,t}$ and $PP_{i,t}$ refer to the pension contributions and payments, as share of GDP, respectively. $WL_{i,t}$ and $LE_{i,t}$ represent the duration, in years of working life and life expectancy at retirement (age 65) respectively. Values of statistics for each country and year are reported in Appendix A (Table 3). Note that to construct PISP there is no need to explicitly consider country-specific pension parameters of pension systems (e.g. contribution rates, pension formulas).

3. Results

The computed scores of PISP are depicted below and reported for each country (Fig. 1).
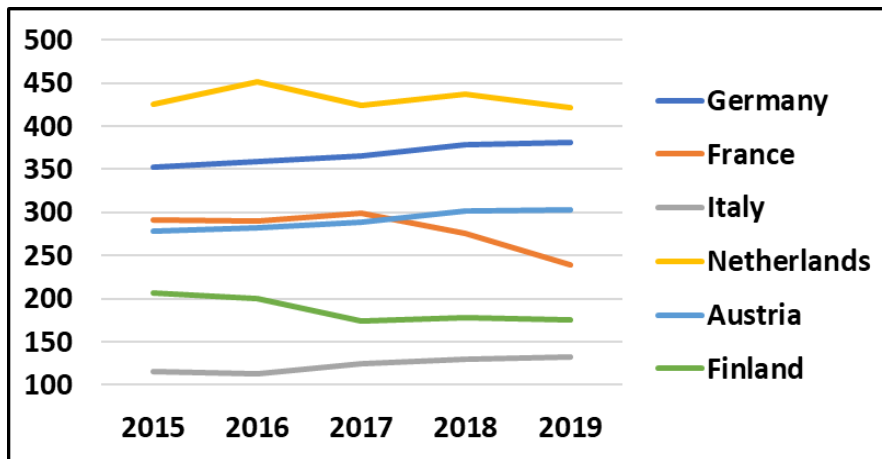
Figure 1: PISP scores across countries and time.
Source: author's own elaborations on Eurostat data.

The Netherlands shows the highest profile in terms of PISP, despite it has decreased in the last years. Germany and Austria report an increasing profile, while those of France and Finland are decreasing. Furthermore, France profile is penalized by a decreasing GDP share of pension contributions ($PC_{i,t}$). Italy shows the lowest profile, despite it slightly increases after 2016.

The PISP, as defined by equation 1, is now compared to an alternative version that excludes the durational margins from PISP (CISP, standing for Current ISP). Thus, CISP is defined as:

$$CISP_{i,t} = PC_{i,t} - PP_{i,t} \qquad\qquad (2)$$

For each country in the pool, the time profile of $CISP_{i,t}$ is depicted in figure 3.
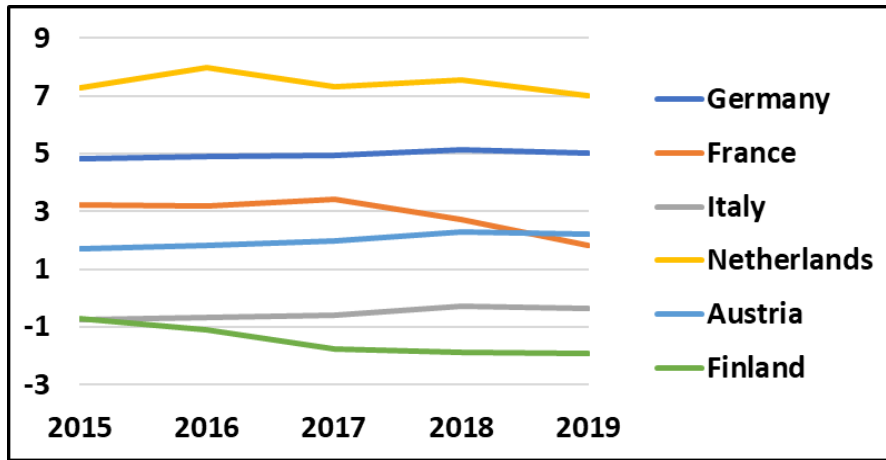
Figure 3: CISP scores across countries and time.
Source: author's own elaborations on Eurostat data.

CISP profiles, being simply defined as the current balance of pension system expressed in terms of share of GDP, show the direct impact of pension contribution and payment flows. All countries but Italy and Finland report positive values, with Finland having the lowest. Overall, three features emerge. Firstly, Italy and Finland maintain the lowest profile across indicators. Secondly, the Netherlands remains at the top positions followed by Germany. Thirdly, from 2017 onwards the financial sustainability of the French public pension system decreases. Explanations rely on the dynamics of each component forming the contribution and payment side.

Finally, the Mercer-Melbourne Global Pension Index - Sustainability ($GPI_{i,t}$) is taken as a benchmark (Mercer-Melbourne, 2015-2019). Countries profiles are reported in Figure 4.
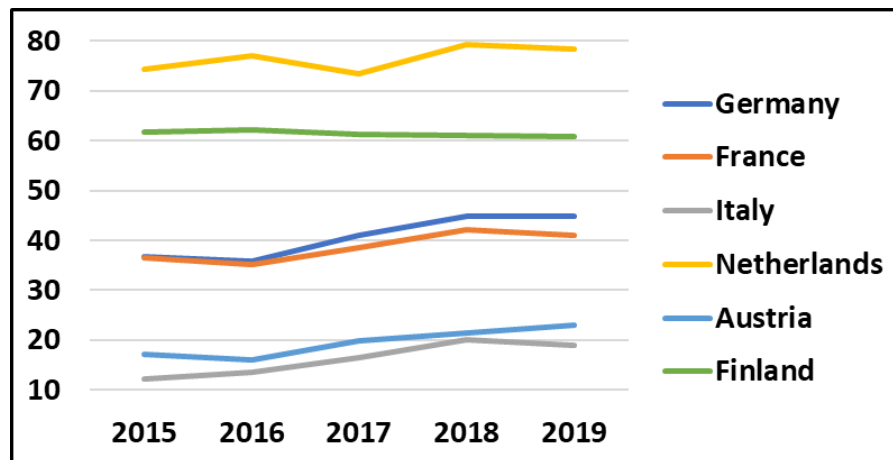
Figure 4: GPI scores across countries and time.
Source: Mercer-Melbourne (2015, 2016, 2017, 2018, 2019).

Despite GPI is an index for the sustainability of the whole (i.e. all pillars of) pension systems, it is possible to note that the Netherlands are confirmed to be ranked first throughout the considered period. Finland, ranked second, shows a stable pattern. Germany and France, placed third and fourth respectively, evolve along the same trend. Lastly, Austria and Italy show a positive trend, with the latter country having the lowest profile for the whole period.

PISP and CISP are then compared each other as well as to the benchmark indicator GPI. For each pair of indicators, their (Pearson) correlation coefficient ρ is measured. Results are reported below (Table 2).

COUNTRY	ρ(PISP, CISP)	ρ(PISP,GPI)	ρ(CISP,GPI)
Germany	0.85	0.96	0.83
France	1.00	-0.60	-0.63
Italy	0.92	0.94	0.98
Netherlands	0.97	0.35	0.15
Austria	0.99	0.94	0.91
Finland	0.98	0.85	0.85
Average	**0.95**	**0.57**	**0.51**
St. Dev.	**0.06**	**0.62**	**0.64**

Table 2: Correlation (Pearson) across pair of indicators: PISP, CISP and GPI.
Source: author's own elaboration.

The first column shows that the correlation between PISP and CISP is quite high for all countries (on average 0.95), with the lowest dispersion across countries (0.06) if compared to the other two columns. This result can be interesting if one considers that countries have different pension regimes, namely different rules determining pensions contribution and payment flows. The last two columns of Table 2 report the correlation between PISP and CISP with the benchmark GPI. Firstly, note that in the case of France it is negative in both cases. This means that GPI may underestimate the impact of current imbalances on the overall sustainability of public pension systems. Secondly, correlation is very high for countries like Germany, Austria, Finland and Italy. On the other side, the proposed indicators are not informative about the overall sustainability of France and the Netherlands as measured by GPI. Overall, PISP reveals some desirable properties. In fact, not only the average correlation with current imbalances is higher for PISP (0.95) than for GPI (0.51), but it is also the least dispersed across countries (0.06 and 0.64, respectively).

4. Conclusions

This work proposes an indicator for the financial sustainability of public pension systems. The novel lies in the consideration of durational margin for pension contribution and payment sides, namely the duration of working life and the life expectancy at retirement. In doing so, it explicitly combines both labour market and pension statistics in a unifying indicator, thus stressing their interplay. These novels are coherent with the recent focus posed by Eurostat within the set of indicators selected to monitor the sustainability of pension systems.

The proposed indicator PISP satisfies some properties which are desirable in the context of financial sustainability of public pension systems: highly correlates with current imbalances and the benchmark GPI and it is the least dispersed across countries. On the contrary, GPI shows a weaker correlation with current imbalances and negatively correlates in the case of France. This

consideration opens the need towards a solid and reliable indicator which meaningfully track the financial performances of pension systems. The structure of PISP can be conceived as a first step to be further developed towards an informative system reliable and comprehensive for policymakers (Whitehouse, 2012). The challenge becomes of utmost importance in a context of an ageing society, where the proportion of retired people increases and accordingly the importance of pensions.

References

Bravo, J. M., & Herce, J. A. (2020). Career breaks, Broken pensions? Long-run effects of early and late-career unemployment spells on pension entitlements. *Journal of Pension Economics & Finance*, 1-27.

Holzmann, R. (2005). Old-age income support in the 21st century: An international perspective on pension systems and reform. *World Bank Publications*.

Mercer, Melbourne (2015). *Melbourne Mercer Global Pension Index*.

Mercer, Melbourne (2016). *Melbourne Mercer Global Pension Index*.

Mercer, Melbourne (2017). *Melbourne Mercer Global Pension Index*.

Mercer, Melbourne (2018). *Melbourne Mercer Global Pension Index*.

Mercer, Melbourne (2019). *Melbourne Mercer Global Pension Index*.

Whitehouse, E. (2012). Pension indicators: reliable statistics to improve pension policymaking (No. 70347, pp. 1-12). *World Bank Publications*.

APPENDIX A: Dataset

COUNTRY	PENSION CONTRIBUTION (% GDP)				
	2015	2016	2017	2018	2019
Germany	14.04	14.21	14.33	14.54	14.70
France	16.72	16.68	16.74	16.03	14.91
Italy	12.95	12.73	12.70	13.00	13.26
Netherlands	13.98	14.67	13.81	13.96	13.49
Austria	14.49	14.51	14.56	14.68	14.82
Finland	12.60	12.70	11.95	11.83	11.76
COUNTRY	WORKING LIFE DURATION (years)				
	2015	2016	2017	2018	2019
Germany	37.9	38.2	38.4	38.7	39.1
France	34.9	35	35.2	35.4	35.4
Italy	30.7	31.3	31.7	31.8	32
Netherlands	39.9	39.9	40.1	40.5	41
Austria	36.7	37.1	37.2	37.5	37.7
Finland	37.7	37.7	38	38.7	38.9
COUNTRY	PENSION EXPENDITURE (% GDP)				
	2015	2016	2017	2018	2019
Germany	9.2	9.3	9.4	9.4	9.7
France	13.5	13.5	13.3	13.3	13.1
Italy	13.7	13.4	13.3	13.3	13.6
Netherlands	6.7	6.7	6.5	6.4	6.5
Austria	12.8	12.7	12.6	12.4	12.6
Finland	13.3	13.8	13.7	13.7	13.7
COUNTRY	LIFE EXPECTANCY AT AGE 65 (years)				
	2015	2016	2017	2018	2019
Germany	19.5	19.8	19.7	19.6	19.9
France	21.6	21.8	21.8	21.9	22
Italy	20.6	21.3	20.9	21.3	21.4
Netherlands	19.8	19.9	20	20	20.3
Austria	19.8	20.2	20.1	20.1	20.3
Finland	20.2	20.2	20.4	20.4	20.6

Table 3: Dataset for the construct PISP and CISP. Source: Eurostat.

Unemployment dynamics in Italy: a counterfactual analysis at Covid time

Illya Bakurov, Fabrizio Culotta

1. Introduction

The experience of the Covid pandemic has revealed the importance of statistical monitoring systems. When the phenomenon of interest is evolving, information about past and future dynamics becomes fundamental to assess both effects and future trajectories. This is true especially during the recovery and post-recovery phases.

The various measures undertaken in each country to contain the spread of the virus moved towards the reduction of physical interaction and, a fortiori, gathering of people. The underlying uncertainty, in the Knightian sense, forced governments of almost all countries to impose harsh remedies. The freedom of movement has been suspended for a while. Undoubtedly, the onset of Covid was an unprecedented and unexpected shock for the world population and, thus, for the whole economy. In this, Italy can be considered a case of study.

Focusing on labour markets, from one side, a substantial drop in unemployment has been observed in Italy during the year 2020 (Fig. 1). The conditions of active search and (immediate) availability to work, whose simultaneous fulfilment identifies an unemployed individual, were not met. Accordingly, a consequent rise in inactivity occurred. From the other side, the Italian government introduced a ban on dismissal operating throughout the year 2020. The goal was to preserve the employment level avoiding firms to fire workers massively. In doing so, the level of employment was forced not to drop. Overall, this was the extraordinary regime under which the observed dynamics of unemployment evolved in Italy during the year 2020.

The aim of this work is to compare the observed dynamics of unemployment during the year 2020 in Italy with a counterfactual outcome to assess the (broad) impact of Covid in Italy in terms of unemployed individuals. In doing so, counterfactual outcomes are generated by Seasonal ARIMA models (SARIMA). Results are presented for the whole population of unemployed individuals and disaggregated by socioeconomic dimension as gender, age, education.
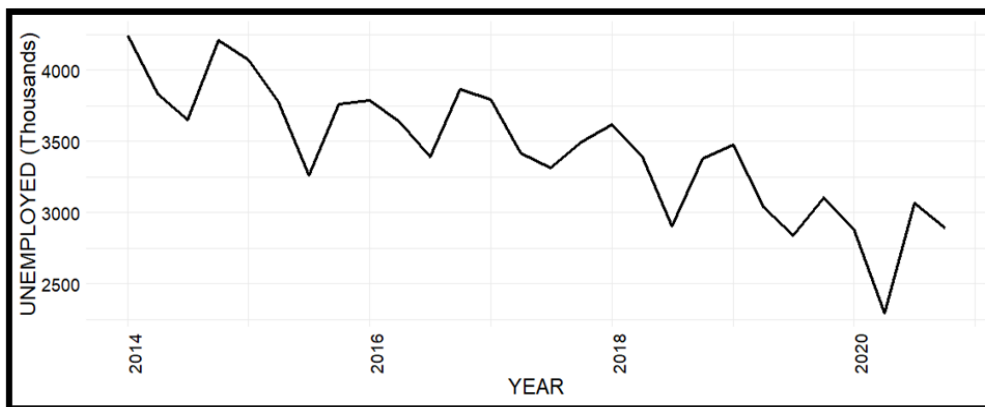
Figure 1: Unemployed individuals (thousands) in Italy over the years 2014-2020 quarters. Source: Italian Labour Force Survey.

2. Data and Methods

This work adopts data from the Italian Labour Force Survey (Rilevazione sulle Forze di Lavoro, hereafter ILFS, for the years 2014-2020 at quarterly frequency focusing on working age population, i.e. aged 15-64 years, is considered. Raw (not smoothed) data covering the period 2014-2019 are used to train SARIMA models to forecast the four quarters of year 2020 (see Hyndman and Athanasopoulos, 2018, as a reference book). It is implicitly assumed that the first quarter of 2020 is the first period affecting unemployment dynamics, that is training data are not affected by the treatment. Estimated projections are then compared with observed values. The causal impact of Covid will be then defined as the difference between what is observed during the 2020 quarters (under the influence of Covid measures) and what would have been observed (in the absence of Covid measures). This empirical exercise is performed not only for the total population but also for eleven socioeconomic groups: two by gender (males and females), five by age (15-24, 25-34, 35-44, 45-54, 55-64), four by educational level (primary, lower and upper secondary, tertiary). The analysis is performed in *R* with help of the package *fpp2* (Hyndman et al., 2020).

Diagnostics analyses are also performed and available upon requests. In particular, trend-cycle decompositions visually suggest that each series exhibits strong seasonality which however is stable in variance over time. The visual inspection suggests that both seasonal and first differencing could take place. Therefore, we run a sequence of Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski et al., 1992) for the null hypothesis of stationarity in the data, and we look for the evidence of rejection. Results from KPSS tests confirm that both seasonal and first differencing should take place. Model orders are selected by inspecting PACF and ACF. The winner model has been selected based on common information criteria AIC and BIC. Estimated models are reported below (Table 1).

Profile	ARIMA (p,d,q)(P,D,Q)$_m$	Drift
Total	ARIMA (0,0,0)(0,1,1)$_4$	Included
Female	ARIMA (0,0,0)(0,1,1)$_4$	Included
Male	ARIMA (0,0,0)(0,1,1)$_4$	Included
Aged 15-24	ARIMA (0,0,0)(0,1,1)$_4$	Included
Aged 25-34	ARIMA (0,0,0)(0,1,1)$_4$	Included
Aged 35-44	ARIMA (0,0,0)(0,1,1)$_4$	Included
Aged 45-54	ARIMA (2,1,0)(0,1,1)$_4$	Excluded
Aged 55-64	ARIMA (0,0,1)(0,1,0)$_4$	Included
Primary	ARIMA (0,0,0)(0,1,0)$_4$	Included
Lower Secondary	ARIMA (0,0,0)(0,1,1)$_4$	Included
Upper Secondary	ARIMA (0,0,0)(0,1,1)$_4$	Included
Tertiary	ARIMA (6,0,0)(0,1,1)$_4$	Included

Table 1: Estimated Seasonal ARIMA (SARIMA) models.
*Drift refers to the time-invariant intercept component.*

The resulting error components are distributed as white noise. Results from Ljung-Box suggest accepting null hypothesis of serially uncorrelated errors. Accordingly, we use such model specifications to produce the counterfactual trajectories for the 2020 quarters.

3. Results

Results are reported in figures 2-5. For sake of graphical clarity, confidence intervals are provided only for the total profile (Fig. 2). For the other profiles, they are available upon requests.
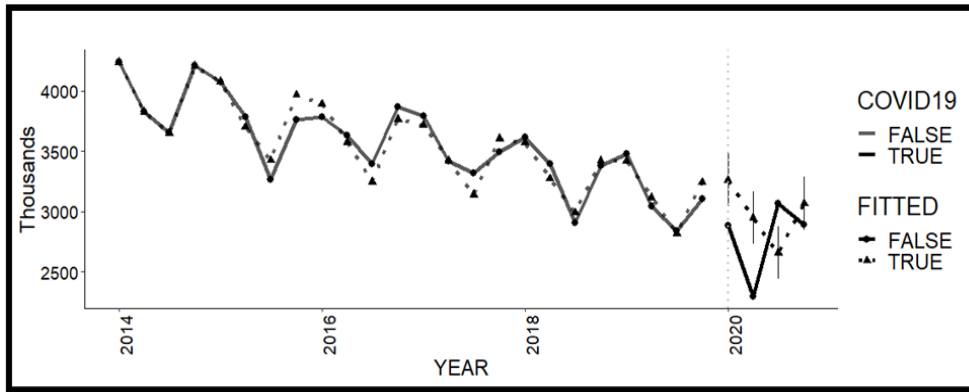
Figure 2: Total profile. Observed (FITTED=F) vs Forecast (FITTED=T) at Covid time (COVID19=T). Source: authors' own elaborations on ILFS data.
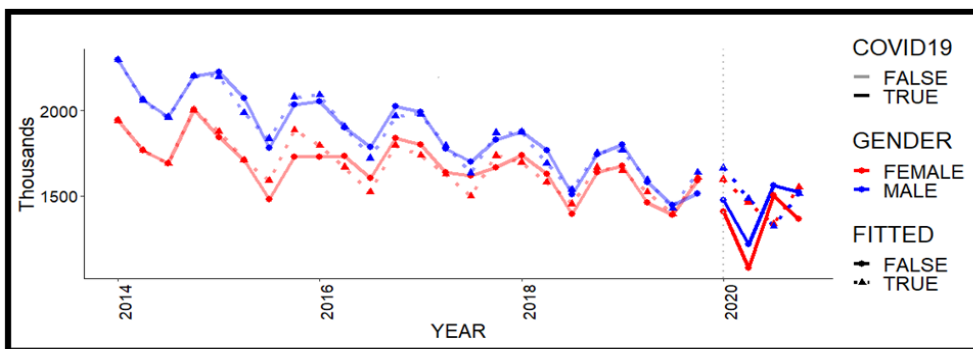
Figure 3: Gender Profiles. Observed (FITTED=F) vs Forecast (FITTED=T) at Covid time (COVID19=T). Source: authors' own elaborations on ILFS data.
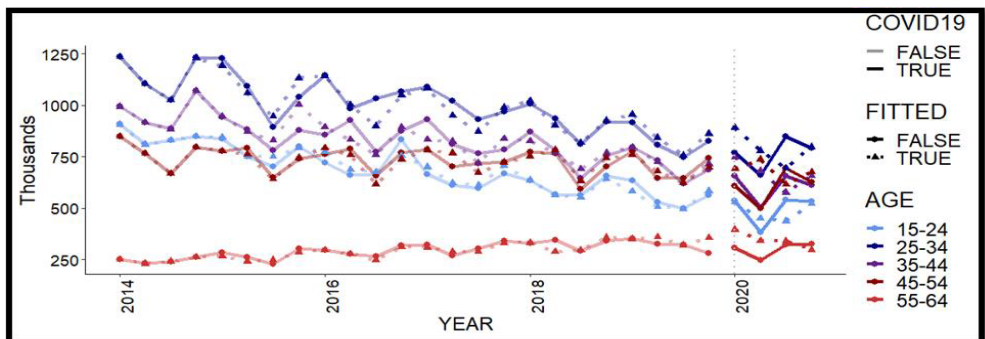
Figure 4: Age profiles. Observed (FITTED=F) vs Forecast (FITTED=T) at Covid time (COVID19=T). Source: authors' own elaborations on ILFS data.

Figure 5: Education profiles. Observed (FITTED=F) vs Forecast (FITTED=T) at Covid time (COVID19=T). Source: authors' own elaborations on ILFS data.

All profiles depicted above share some general features. The difference between observed and forecast values, i.e. the impact of Covi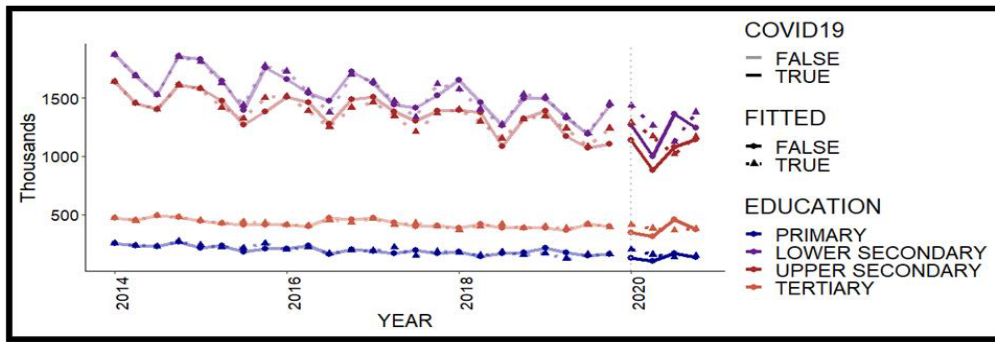d on the number of unemployed, is negative during the first quarter of 2020. The difference is even larger (in absolute value) during the second quarter of 2020, where the impact of Covid in 2020 led a relevant drop in the number of unemployed. The difference is positive at the third quarter, coinciding with the summer season, and becomes negative during the fourth.

Results are also displayed in a tabular format (Table 2). The largest impact of Covid on unemployed workers corresponds to 2020-Q2 (-652000). This drop is concentrated among females (-385000, about 60% of the total drop). Across age classes, 45-54 reports the largest reduction (-237000, around 36%). Whereas unemployed with lower secondary education are the educational group (-295000, 45%). During the third quarter of 2020, the raise in unemployed occurred (410000). Of such a raise, men were the majority (242000, 60%). The age class 25-34 shows the largest share (153000, 37%). Similarly, the largest increase is observed for the upper secondary educational group (238, 58%). Overall quarters, it results that the drop in unemployment caused by Covid during the year 2020 regarded women more than men especially from the second quarter. Among the age groups, Covid had an impact especially on unemployed individuals aged 45-54. Across educational levels, following this logic, individuals without a tertiary education were affected the most.

4. Conclusions

This work studies the impact of Covid pandemic, and related measures, on the number of unemployed workers during the 2020 quarters in Italy. Observed and counterfactual outcomes are compared to identify the causal impact of the onset of Covid since the first quarter of 2020.

In doing so, counterfactual outcomes are produced by means of SARIMA models applied to different socioeconomic groups in the population of unemployed. The causal impact is then measured as difference between observed and forecast values. Results confirms that the drop in unemployment caused by Covid was heterogenous, i.e. not homogenously distributed in the population. Females, individuals aged 45-54 and those with secondary educational levels were those groups associated with the highest drop.

In general, the counterfactual analysis is used as a tool to identify causal mechanism. In the case of this work, the (macro-)econometric model is also offered as a (simple) policy statistical tool. It can be used to identify future patterns and to reason on possible thresholds or rebounds. It can offer an informative, yet statistical, support to face important decisions under uncertainty. Possibly, it can reveal insights for future planning.

PROFILE	Source	2020-Q1	2020-Q2	2020-Q3	2020-Q4
Total	Observed	2883	2295	3067	2891
	Forecast	3261	2947	2657	3065
	Covid Impact	*-378*	*-652*	*410*	*-174*
Female	Observed	1407	1078	1504	1366
	Forecast	1597	1463	1337	1551
	Covid Impact	*-190*	*-385*	*167*	*-185*
Male	Observed	1476	1217	1563	1525
	Forecast	1664	1484	1321	1514
	Covid Impact	*-188*	*-267*	*242*	*11*
Aged 15-24	Observed	539	381	543	532
	Forecast	529	450	437	524
	Covid Impact	*10*	*-69*	*106*	*8*
Aged 25-34	Observed	770	663	849	793
	Forecast	892	780	696	798
	Covid Impact	*-122*	*-117*	*153*	*-5*
Aged 35-44	Observed	658	502	658	613
	Forecast	748	682	576	657
	Covid Impact	*-90*	*-180*	*82*	*-44*
Aged 45-54	Observed	609	500	693	628
	Forecast	690	737	616	675
	Covid Impact	*-81*	*-237*	*77*	*-47*
Aged 55-64	Observed	307	249	324	325
	Forecast	396	341	340	298
	Covid Impact	*-89*	*-92*	*-16*	*27*
Primary Ed.	Observed	129	105	167	132
	Forecast	54	158	137	149
	Covid Impact	*75*	*-53*	*30*	*-17*
Lower Secondary Ed.	Observed	1136	877	1079	1142
	Forecast	1287	1172	1021	1167
	Covid Impact	*-151*	*-295*	*58*	*-25*
Upper Secondary Ed.	Observed	1271	999	1364	1243
	Forecast	1435	1265	1126	1378
	Covid Impact	*-164*	*-266*	*238*	*-135*
Tertiary Ed.	Observed	348	314	457	374
	Forecast	413	381	367	379
	Covid Impact	*-65*	*-67*	*90*	*-5*

Table 2: Counterfactual analysis on unemployment dynamics (thousands of individuals) in Italy at Covid time (2020 quarters). Covid impact is the difference between observed and forecast values. Source: authors' own elaborations on Italian Labour Force Survey data.

References

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, **54**(1-3), 159-178.

Hyndman, R. J., Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., ... , Wang, E. (2020). *Package 'forecast'*. https://cran. r-project.org/web/packages/forecast/forecast.pdf.

TOURISM AND GASTRONOMY

Understanding the sensory characteristics of edible insects to promote entomophagy: A projective sensory experience among consumers

Alfonso Piscitelli, Roberto Fasanelli, Elena Cuomo, Ida Galli

1. Introduction

The world population is continuously growing, and with it, the demand for food increases. Processes such as urbanization and globalization are increasingly influencing dietary change for a considerable part of the population. The result is a constant increase in the need for high biological value proteins, the production of which represents a challenge for the future, especially considering that current production techniques (i.e., animal protein farming) not only have a significant environmental impact but also show a low level of efficiency. These techniques produce high levels of carbon dioxide, consume considerable amounts of water, and involve major waste-disposal problems (Amato et al., 2019).

The European Parliament has indicated that the deficit in protein sources is one of Europe's most critical problems: the Old Continent imported about 80% of its protein from other countries. Insects can be a sustainable alternative to this problem, for their efficient metabolism and their ability to transform organic waste into high-quality protein (Materia and Cavallo, 2015). Western countries' interest in insects as a potential source of food has grown considerably in recent years: the high content of high-quality protein and the sustainability of the production process, compared to traditional sources, primarily meat, have contributed to increasing scientific debate on the topic. The progressive inclusion of insect-based ingredients in the human diet has attracted increasing attention as a valid alternative to overcome the major nutrition challenges the world is facing (Schrögel and Wätjen, 2019). However, a diet based on insects (or their components) entails a radical departure from Western societies' current food traditions. Although recent research shows that consuming insects (raw or processed) provides significant benefits in terms of protein content, social acceptance is, on the contrary, very low in Western societies (Verneau et al., 2016; La Barbera et al., 2018; 2020). However, insects and their derivatives in food products are not entirely new even in the West: products such as jams and fruit juices contain traces of them, for an estimated average per capita consumption of 250 gr/year (Materia and Cavallo, 2015; Sogari and Vantomme, 2014), even if a clear awareness of this is still lacking. Scholars conducted several studies to analyze consumer behavior employing insect-based foods; many of these have identified factors that may positively or negatively influence the degree of acceptance.

Our basic hypothesis is that intention to try insect-based dishes is causally dependent on sensory reasons that are relevant to sensation-seeking. To test this hypothesis, we interviewed a convenience sample of consumers to examine the relationships between their intention to try insect-based dishes and their anticipatory gustative sensations regarding "insects as food". The research data were obtained from a web questionnaire completed by a sample of consumers which was held using social media and via e-mail lists.

The paper is organised as follows. After this introduction, Section 2 describes the sample

Alfonso Piscitelli, University of Naples Federico II, Italy, alfonso.piscitelli@unina.it, 0000-0001-6638-2759
Roberto Fasanelli, University of Naples Federico II, Italy, roberto.fasanelli@unina.it, 0000-0001-8908-3284
Elena Cuomo, University of Naples Federico II, Italy, elena.cuomo@unina.it, 0000-0001-7526-2353
Ida Galli, University of Naples Federico II, Italy, ida.galli@unina.it, 0000-0001-5159-9162

survey and introduces the model for data analysis. Then, Section 3 presents the main results of the statistical analysis of the collected data. Finally, Section 4 interprets the data with reference to sensation seeking in connection with the choice of food.

2. Data and methods

2.1 Consumer survey on entomophagy

A fact-finding/exploratory survey was conducted on a small convenience sample of Italian consumers. The inclusion criterion was represented by having at least heard of the introduction of insects into human nutrition. The survey, conducted between July to December 2020, was carried out by means of a web questionnaire filled out by participants who completely satisfied the inclusion criterion and recruited using social media (e.g., Facebook, Twitter, WhatsApp chats) and via e-mail lists (e.g., University of Naples student lists).

The self-administered questionnaire has been structured into 3 sections each of them with a specific aim of collecting data. In the first section, we asked each participant to answer both semi-structured and structured questions made up starting from the following dimensions: previous knowledge of the "insects as food", informative sources, and opinions about it. As a supplement, three additional items were administered to measure the intention to eat insects in general. This section ended with four items added to ask participants for their willingness to eat specific animals (cow, fish, chicken, and pig) fed with insects. All answers were collected using a 7-point scale from "strongly unlikely" to "strongly likely". A specific section was devoted to what we called projective sensory experience (PSE). To identify respondents' anticipatory gustative sensations regarding "insects as food" we used an ad hoc created two-step tool. In particular, (first step) we asked participants to imagine tasting an insect dish and then rate from 1 imperceptible up to 10 very perceptible the following taste-olfactory sensations inspired by the work of Donadini et al. (2008): Sapidity, Bitter tendency, Acidity, Sweet, Spiciness, Aroma, Greasiness-Unctuosity, Succulence, Sweet, Fatness, Persistence. Furthermore (second step), since a representation is always built from a disturbing object (in positive or negative), at the end of the task, we asked our interviewees to indicate, through a specific checklist, which was the most disturbing and least disturbing imagined taste-olfactory sensation. The goal is to know which kind of "sensory anchoring" participants activate in front of an insect's dish.

The questionnaire ended by collecting the respondents' descriptive characteristics (gender, age, education, living area) as well as their eating habits.

2.2 Analytical model

The model for data analysis includes the intention to try insect-based dishes as a dependent variable, Y, a set of possible regressors, $\mathbf{X}$, and a set of control variables, $\mathbf{Z}$. The relationship may be written as

$$Y = f(X_i \mid Z_i) \, ,$$

where X_i denotes the taste-olfactory sensations. The control variables, which were forced into the model, were gender, age, education, and living area. The Y variable was measured on two levels.

The logistic regression model is written as follows (Agresti, 2002; Bilder and Loughin, 2014):

$$logit \, (\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad ,$$

where $0 < \pi < 1$ and $logit \, (\pi) = log \, [\pi / (1 - \pi)]$, and β_i measures the relation between Y and

X_i when all other variables in the model remain fixed.

$$\pi = \Pr\{Y = 1 \mid X, Z\}$$

where Y = 1 represent the positive intention to try insect-based dishes.

Maximum likelihood estimation is used to estimate the parameters $\beta_0, \ldots, \beta_p$ of the logistic regression model. It will be clear, based on the context, how the probability of success is a function of explanatory variables, being the slope of the logit function hyperplane along each X dependent on the value of the corresponding parameter β.

Statistical analyses were carried out in the R environment (R Core Team, 2021). A logistic regression model to a dichotomous response variable was identified by the *glm* function. Moreover, the *step* function was utilised to perform stepwise model selection with criterion *AIC*.

3. Results

The data were gathered from 154 consumers, of these, about 58% of respondents were female, and about 42% were male. The mean age of respondents was 43 years (SD = 14.12), 53% lived in a highly urbanized area, 39% lived in the urban suburbs, and the remaining 8% lived in a rural area. About 28% of respondents has a very high level of education (superior graduate or PhD), 49% has a bachelor's or master's degree, and the remaining 23% has a level of education of secondary school.

Table 1 summarises the results of the regression analysis reporting the estimates of the regression betas obtained and their significance.

Table 1. Beta estimates, and related standard errors, of the regression model with intention to try insect-based dishes as criterion variable (forward stepwise selection of regressors, *n* = 154; AIC = 181.95; **** 0 < α^{oss} < 0.001; ** 0.001 < α^{oss} < 0.01; * 0.01 < α^{oss} < 0.05; ° 0.05 < α^{oss} < 0.1;* NS= Not significant at 10% threshold; the interactions are labeled by colon between the variables' names)

Regressor	$\hat{\beta}$	se($\hat{\beta}$)	Signific.
Intercept	-0.13239	1.40747	NS
Male	0.24661	0.43939	NS
Age	0.01346	0.01762	NS
AreaMedium urbanization area	-0.14526	0.93184	NS
AreaHigh urbanization area	0.41342	0.90084	NS
EduBachelor's degree	-1.99574	1.54414	NS
EduMaster's degree	-1.70110	1.58013	NS
EduSuperior Graduate or PhD	-2.05785	1.49962	NS
Acidity	-1.08717	0.40715	**
Spiciness	0.70918	0.32008	*
Persistence	-0.15071	0.10155	NS
Sweet_tendency	0.18975	0.10348	°
EduBachelor's degree:Acidity	1.03092	0.46193	*
EduMaster's degree:Acidity	0.80548	0.45596	°
EduSuperior Graduate or PhD:Acidity	0.78677	0.43907	°
EduBachelor's degree:Spiciness	-0.51565	0.37416	NS
EduMaster's degree:Spiciness	-0.42817	0.36845	NS
EduSuperior Graduate or PhD:Spiciness	-0.34148	0.35956	NS

Note: Some regressors are not individually significant but are significant wrt AIC criterion.

Our results highlight a dependence of the willingness to try insect-based dishes by the respondent's fondness to Acidity (indirect), Spiciness (direct) and a mediation effect of education degree on the former. The results of the regression analysis support the following claims:

- imagining a taste-olfactory sensation of "extremely perceptible" acidity reduces the willingness to try insect-based dishes;
- imagining a taste-olfactory sensation of "extremely perceptible" spiciness increases the willingness to try insect-based dishes;
- educational level plays a moderating role because the interaction between educational levels and acidity indicates a differentiating among the participants. The bachelor's degree moderates the effect of acidity and increasing the probability to try insect-based dishes, as do the other two levels of education, but gradually;
- imagining a taste-olfactory sensation of "extremely perceptible" sweet tendency increases the willingness to try insect-based dishes.

4. Discussion and conclusion

We have tested whether the consumption of insects is carried out for sensorial reasons by means of the new tool, ad hoc created, the Projective Sensory Experience (PSE).

As Lammers, Ullmann, and Fiebelkorn (2019) demonstrated, in connection with the choice of food, sensation seeking correlates with liking spicy food and the willingness to try unusual foods also showed that people with a high sensation seeking have a lower food neophobia. From our results, we assume that sensations like "spicy" should be positively related to the willingness to consume insects while "acidity" has a negative role.

In summary, this study confirms the results found in other researches on Italian consumers, who are not completely familiar with the topic of entomophagy. In addition, it was shown that overall, only four of ten of our interviewees would try the experience of eating insects, as well as another study (Sogari et al., 2017) showed that 47% of young Italian "foodies" envisaged insect eating. Moreover, it is important to highlight that the willingness of Italian consumers to adopt insects into their diet seems higher than in other countries of Mediterranean Europe (Mancini et al., 2019). Furthermore, this study identified sensation seeking and especially "sensory anchoring" as additional predictors for the willingness to consume insects as food in addition to the already known influential factors such as gender, educational level, previous insect consumption, food neophobia and food technology neophobia. In our case highest level of education mediate the willingness for spicy insect dishes.

These preliminary results allowed us to identify which aspects are worth focusing on while searching for the multidimensional motivations behind this particular food choice, and also to highlight the moderation role played by some kind of cultural factors. After the screening achieved thanks to this pilot stage of the study, it would be interesting, for example, to understand the role played by environmentalist ideology on the choice to regularly include insects in one's diet. The question arises as to whether the ecology-based food really motivates consumers to eat insects. A more in-depth analysis of the motivation of people already consuming insect products might provide further insights on how insects might be integrated into the Italian diets. The findings of this preliminary analysis are encouraging about the idea of exploring by a "projective" approach the sensory experience related to food, and supports us to continue along this path. Nonetheless, in order to reduce the aversion to insects as food, it is necessary to create opportunities for the Italian population to make their own positive taste experiences, probably by giving to these foods precisely the sensory characteristics that our interviewees have already imagined attractive to their palate.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. Wiley, Hoboken, (NJ).
Amato, M., Fasanelli, R. and Riverso, R. (2019). Emotional Profiling for Segmenting

Consumers: The Case of Household Food Waste. *Calitatea*, **20**(S2), pp. 27–32.

Bilder, C.R., Loughin, T.M. (2014). *Analysis of Categorical Data with R*. CRC Press, Boca Raton, (FL).

Donadini, G., Spigno, G., Fumi, M.D., Pastori, R. (2008). Evaluation of Ideal Everyday Italian Food and Beer Pairings with Regular Consumers and Food and Beverage Experts. *Journal of the Institute of Brewing*, **114**(4), pp. 329-342.

La Barbera, F., Verneau, F., Videbæk, P.N., Amato, M. and Grunert, K.G. (2020). A self-report measure of attitudes toward the eating of insects: Construction and validation of the Entomophagy Attitude Questionnaire. *Food Quality and Preference*, **79**(103757), pp. 1–9.

La Barbera, F., Verneau, F., Amato, M. and Grunert, K. (2018). Understanding Westerners' disgust for the eating of insects: The role of food neophobia and implicit associations. *Food Quality and Preference*, **64**, pp. 120–125.

Lammers, P., Ullmann, L.M., and Fiebelkorn, F. (2019). Acceptance of insects as food in Germany: Is it about sensation seeking, sustainability consciousness, or food disgust?. *Food Quality and Preference*, **77**, pp. 78–88.

Mancini, S., Moruzzo, R., Riccioli, F., and Paci, G. (2019). European consumers' readiness to adopt insects as food. A review. *Food Research International*, **122**, pp. 661–678.

Materia, V.C. and Cavallo, C. (2015). Insetti per l'alimentazione umana: Barriere e drivers per l'accettazione da parte dei consumatori. *Italian Review of Agricultural Economics*, **70**(2), pp. 139–161.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, (AT). Available at: http:/www.R-project.org (accessed 15 June 2021).

Schrögel, P. and Wätjen, W. (2019). Insects for Food and Feed-Safety Aspects Related to Mycotoxins and Metals. *Foods*, **8**(8), pp. 288–315.

Sogari, G. and Vantomme, P. (2014). *A Tavola con Gli Insetti*. Mattioli 1885, Fidenza, (IT).

Sogari, G., Menozzi, D., and Mora, C. (2017). Exploring young foodies′ knowledge and attitude regarding entomophagy: A qualitative study in Italy. *International Journal of Gastronomy and Food Science*, **7**, pp. 16–19.

Verneau, F., La Barbera, F., Kolle, S., Amato, M., Del Giudice, T., Grunert, K. (2016). The effect of communication and implicit associations on consuming insects: An experiment in Denmark and Italy. *Appetite*, **106**, pp. 30–36.

Experience, sensorial skills and personality qualifying a wine consumer as an expert

Luigi Fabbris, Alfonso Piscitelli

1. Introduction

This paper highlights the characteristics of wine consumers that may qualify them as wine experts. In this work, the expertise of wine consumers was measured through various degrees of self-perceived ability. Participants ranged from limited-knowledge consumers to consumers with enough knowledge to perceive wine quality or recognise certain wines and, finally, to professional experts.

Wine is an 'experience good' in that its quality is unknown before consumption. Thus, a wine expert is not only knowledgeable about wine but also practises wine consumption as a continual consumer. In Italy, wine is a cultural product as well, as the consumption habitus depends on consumer taste, which is crucial in choosing products (Bourdieu, 2005). Wine culture is defined as the capacity to harmonise wine and food and conceive of wine as a nutritional, social and health-related means. In this work, the cultural roots of wine were measured through a 'semantic differential' (Osgood et al., 1957) of wine preferences, which was determined with a rating scale designed to measure the assessors' preferences for wine.

Our basic hypothesis is that wine expertise is causally dependent on cognitive and non-cognitive characteristics of the wine experience, sensorial skills that are relevant to wine assessment and wine consumption culture. To test this hypothesis, we evaluated a convenience sample of consumers to examine the relationships between their self-assessment of wine expertise and qualification of their wine-related training and experience (consumption, production, purchase), their sensorial skills (visual, olfactory, gustative), their enogastronomic culture and their approach to evaluating a set of selected wines. The research data were obtained from an evaluation questionnaire completed by a sample of wine assessors at a tasting experiment which was held during a scientific meeting in Pescara, Italy in September 2018. The sample includes both meeting participants and external experts involved in AIS-Abruzzo, the regional association of chartered sommeliers.

The paper is organised as follows. After this introduction, Section 2 describes the methodological aspects of the tasting experience and introduces the model for data analysis. Then, Section 3 presents the main results of the statistical analysis of the collected data. Finally, Section 4 interprets the data with reference to the mainstream literature on wine expertise analysis.

2. Data and methods

2.1 The tasting experience

In September 2018, a sensory evaluation experiment was conducted on 12 white wines originating from six grape varieties (*Trebbiano d'Abruzzo, Pecorino d'Abruzzo, Passerina d'Abruzzo, Pagadebit di Romagna* and *Pignoletto di Romagna*) from two Italian regions, Abruzzo and Romagna. All wines were controlled designation of origin (DOC) products. The pool of tasters included 48 individuals, of whom 30 typically consumed mild amounts of wine

(mild consumers), and 18 were professional sommeliers belonging to the AIS-Abruzzo association. Both mild consumers and sommeliers were selected on the basis of their interest in and availability for the experiment as well as their experience in wine consumption.

The wine characteristics considered in this evaluation experiment were selected through an anonymous paper questionnaire. This questionnaire asked participants to make judgements on 11 intrinsic attributes of appearance, nose and palate for four wines that were randomly selected from the 12 at hand. Subsequently, participants were instructed to provide an overall judgement of each wine. The questionnaire also gathered data on the tasters regarding their background characteristics, their drinking habits, and the relevance of wine in their diet and social life. In this work, we confine the analysis to the characteristics of tasters. The characteristics of the assessed wines enter the analysis only as distributional parameters (mean and variance) of the scores which single assessors assigned to the tasted wines.

2.2 The experiment

The experiment involved a horizontal tasting, as it compared only white wines from the same terroir and of the same vintage. On this basis, it is possible to obtain comparative judgements between the selected wines. In accordance with a fractional factorial experiment, each taster was administered four randomly selected wines from different grapes. The sampling of the administered wines was carried out at the grape-variety level. Only four of the six possible varieties were administered to any taster, and one of the two potential cellars was randomly selected. In this case, the experiment sampled possible choices rather than choosers (Manski and Lerman, 1977).

The sampling design followed a systematic pattern such that each grape variety appeared 8 times every 12 trials. Thus, each wine variety had 32 repetitions once 48 tasters had performed their task; consequently, the number of repetitions of each variety by cellar was 16.

Each taster had five glasses: one for water and four for the wines. The wines were poured in a flight. In the tasting session, the judges received 6 centilitres of each of the four randomly selected wine varieties, which were served at the same cold temperature. The protocol envisaged that tasters could taste and re-taste before concluding preferential judgements, and they would evaluate the intrinsic attributes of each tasted wine.

2.3 Analytical model

The model for data analysis includes the self-evaluation of wine expertise as a dependent variable, Y, a set of possible regressors, $\mathbf{X}$, and a set of control variables, $\mathbf{Z}$. The relationship may be written as

$$Y = f(X_1, X_2, X_3, X_4, X_5 \mid Z),$$

where X_1 denotes wine expertise and learning experience, X_2 represents the descriptors of wine habits, X_3 refers to the sensorial skills, X_4 signifies the descriptors of wine-related attitudes and culture, and X_5 is the evaluation style of the tasted wines. The latter was measured through the mean and standard deviation of the scores for the four tasted wines. The underlying hypothesis was that the evaluations by experts would be more critical and uniform than those of nonexperts. The control variables, which were forced into the model, were gender, age and smoking experience. The Y (ordinal) variable was measured on four levels.

The ordinal logistic regression model is written as follows (Agresti, 2002; Bilder and Loughin, 2014):

$$logit\ [p(Y \leq j)] = \beta_{j0} + \beta_1 X_1 + \cdots + \beta_p X_p \quad (j = 1, \ldots, J\text{-}1),$$

where $logit(p) = ln[(p/(1\text{-}p)]$, and β_i measures the relation between Y and X_i when all other variables in the model remain fixed. We adopted the proportional odds model, which assumes that the logit of the cumulative probabilities changes linearly as the regressors change, and the

slope of the relationship between *Y* and the *X*'s is the same regardless of the category *j* of variable *Y*.

A logistic regression model to an ordered response variable was performed with the *polr* function from MASS package (R Core Team, 2021). After that, the *stepAIC* function was utilised to perform stepwise model selection with criterion *AIC*.

3. Results

Of the 48 assessors, five (10.4%) considered themselves to be wine experts, and eight (16.7%) stated that they were able to recognise some wines but did not consider themselves to be wine experts. The majority of the participating sommeliers classified themselves in the latter category. A larger group of assessors (47.9%) indicated that they possessed sufficient knowledge of wine to adequately understand its quality. Finally, 25% of the assessors admitted that they knew little or very little about wine.

Overall, our sample included a group of experts and a group of nonexperts (each accounting for approximately one-quarter of the tasters) as well as a larger, intermediate category of mildly informed amateurs (about one-half of the tasters). Only 3 of the 48 assessors produced or bottled their own wine; the others bought it occasionally or on a monthly basis either at vineries or in supermarkets or wine shops. A few (8.3%) purchased wine through the internet.

Regarding wine practice, about 56% of assessors had been consuming wine for decades, usually with dinner. The majority (54.2%) had attended a wine-tasting session coordinated by a sommelier. One-half of the tasting sample was female, in which the average age was 47. This group mostly had a college degree (66.7%), worked mainly at a university (81.3%) and did not smoke (41.7% had never smoked, and 29.2% had formerly smoked).

Table 1 summarises the results of the regression analysis and presents the estimates of the regression betas and their significance. We highlight the elevated significance of the statistical analysis: R^2=61.3%. To corroborate the regression results, selected covariates are crossed with the self-perceived expertise of assessors (Table 2).

Table 1. Beta estimates of the regression model with expertise level as criterion variable (forward stepwise selection of regressors, n=48; R^2=61.3%; AIC criterion=76.53;
***$0 < \alpha^{pss} < 0.001$**; **$0.001 < \alpha^{pss} < 0.01$**; *$0.01 < \alpha^{pss} < 0.05$**; °$0.05 < \alpha^{pss} < 0.1$**; **NS**= Not significant

Regressor		$\hat{\beta}$	$se(\hat{\beta})$	Signific.
Intercept:	Little/Enough	36.816	9.960	***
"	Enough/Recognise	43.605	11.291	***
"	Recognise/Expert	47.767	11.988	***
Male		0.662	0.983	NS
Age		0.121	0.043	**
Smoker		3.806	1.311	**
Self-evaluated olfactory skill		1.220	0.373	***
Buys wine on line		5.535	2.127	**
Buys wine at delicatessen/wine shops		1.890	1.045	°
Buys wine from producers, vineries		3.955	1.475	**
Deals with wines at home		4.159	1.378	***
Wine relevant at celebratory meals		1.407	0.516	**
Wine relevant at dinner		-0.575	0.228	**
Dry *vs*. Sweet (semantic differential)		-0.155	0.231	NS
s.d. of visual evaluations		4.270	1.452	**
Mean of global evaluations		1.056	0.541	°

Note: Some regressors are not individually significant but are significant wrt AIC criterion.

The analysis supports the following claims:

- Wine is a relevant aspect of experts' everyday life, and they are particularly interested in consuming wine at home. While 92.3% of experts decided on wine pairings with meals at home, this figure was only 52.2% and 25% amongst amateurs and nonexperts, respectively. While amateurs and nonexperts collaborated in decision-making about wine at home, their choices were agreed upon with other family members. Hence, the results suggest that the leading role of experts in wine selection starts at home.
- Experts typically buy wine through the internet or at specialised shops. Notably, no producer or bottler self-identified as an expert. Individuals who considered themselves to be wine experts purchased only specific bottles that could be traced through the label to guarantee the quality of their contents. At our trial, all assessors who had bought wine through the internet rated themselves as experts. Moreover, experts purchased bottles from exclusive shops or trusted producers.
- Experts attended at least one sommelier-led tasting course. This aspect corroborates the image of experts as people who have refined their skills by completing courses or tasting sessions in which a qualified sommelier guided them in recognising certain intrinsic attributes of wines (Fabbris and Piscitelli, 2021) and developing their olfactory and tasting capacities.
- Olfactory perceptual ability (smelling) is a fundamental skill of wine experts. The experts were aware that smell is the sense that best qualifies their ability to detect the volatile components of a wine. Accordingly, experts assigned higher ratings to their own olfactory skill compared to amateurs and nonexperts (means: 7.69, 6.74 and 5.25 out of 10, respectively). It is well known that even when wine is in a person's mouth, and they are prepared to activate their palate and throat, the notes and flavours that they experience are partly due to aromas that reach the nose. The data reflect that smelling skills specific to experts include the ability to perceive aromas while drinking wine. Scholars have acknowledged a strong correlation between the two skills, which suggests that olfactory-gustatory abilities constitute a joint skill.

Table 2. Some covariates, by the self-perceived expertise of assessors

Regressor	Self-perceived expertise			
	Expert (n=13)	Amateur (n=23)	Non-expert (n=12)	Total (n=48)
Deals with wines at home (%)	92.3	52.2	25.0	56.3
Buys wine on line (%)	30.8	0.0	0.0	8.3
Buys wine at deli/wine shops (%)	53.8	43.5	25.0	41.7
Attended tasting events led by a sommelier (%)	76.9	60.9	16.7	54.2
Wine relevant at celebratory meals (mean)	9.08	9.13	7.92	8.81
Wine relevant at dinner (mean)	6.15	6.04	5.58	5.96
Wine relevant to socialise (mean)	8.08	7.70	7.58	7.77
Olfactory skill (mean)	7.69	6.74	5.25	6.63
Tasting skill (mean)	7.76	6.78	5.67	6.77
Wine visual evaluation (mean)	7.87	7.12	6.85	7.26
Wine olfactory evaluation (mean)	7.17	6.73	6.65	6.83
Wine taste evaluation (mean)	7.19	6.58	6.67	6.77
Wine overall evaluation (mean)	7.13	6.55	6.69	6.74
Wine visual evaluation (s.d.)	0.73	0.93	0.79	0.84
Wine olfactory evaluation (s.d.)	0.93	0.99	1.13	1.01
Wine taste evaluation (s.d.)	0.83	0.96	1.08	0.95
Wine overall evaluation (s.d.)	0.85	1.20	0.94	1.04

Note: The levels "being wine experts" and "being able to recognise some wines" of the self-evaluation of wine expertise, have been merged into the "Expert" category.

- To identify a person as a wine expert, cultural and psychological traits are less relevant than wine consumption habits. However, when all other variables remained fixed, a significant regressor was the relevance which experts attributed to wine in a foreign celebratory meal. This result implies that experts conceive of wine as a professional rather than social means. During an official meal, not only might a refined wine be served, but experts may also have the opportunity to share and improve their expertise with other connoisseurs. Therefore, while many people drink wine to cultivate happiness together, experts tend to try a wine to understand it and possibly foster appreciation for that wine. This element qualifies an expert as an initiate.
- Experts evaluate wines more accurately than amateurs and nonexperts. Experts significantly outperformed amateurs and nonexperts in both their overall assessment of the tasted wines and their visual, olfactory and gustative scores. The experts' scores for the overall judgement of the tasted wines were 6.6% higher than those of nonexperts. In addition, the experts scored 14.9% higher on the visual evaluations and 7.8% higher on the odour and taste evaluations compared to nonexperts. The scores of the intermediate category, amateurs, were likewise intermediate. The multivariate analysis reveals that, *ceteris paribus*, the experts globally evaluated the tasted wines more highly than the others to a significant degree.
- Through analysis of within-category variability, we also checked if all assessors scored wines at the same level of homogeneity. The results indicate that the visual evaluation scores of experts were significantly more homogeneous than those of the other assessors. In fact, all of the scores from experts were more homogeneous than those from other assessors. Notably, amateurs presented the most variance in wine scoring, which evidences that experts and nonexperts are two compact categories, while intermediate expertise implies a relatively heterogeneous knowledge of wines.
- The semantic differential which was applied to unveil people's attitudes toward wine does not highlight significant differences between experts and other wine lovers. This result reflects that wine is well rooted in Italian food culture and enjoys a high degree of acceptance amongst both experts and nonexperts.

4. Discussion and conclusion

This work has aimed to define the characteristics of wine experts. An expert is a person who possesses in-depth knowledge, abundant experience, a proclivity for vivid imagery and a stronger descriptive capacity than that of other people (Parr et al., 2002; Ericsson et al., 2007; Croijmans and Majid, 2016; Croijmans et al., 2020). A wine expert also demonstrates an acute capacity to recognise, classify and evaluate wine characteristics. Notably, skill training can be particularly effective with practice.

The analysis illustrates that experts assumed a leading role in wine selection, which is a habit that they adopted decades ago and had improved over time. The attendance of specific courses and tasting events led by sommeliers or between-peer contests improved their expertise and self-confidence. An expert clearly seeks to train themself through both the exploration of new sensations and products and the intensification of their sensorial skills.

Furthermore, the data analysis indicates that experts perceive themselves as different from producers and bottlers. For both producers and experts, wine is a professional means as well as an important part of their life. However, producers (should) know how to make high-quality wines, whereas experts approach wine from a position which resembles that of an explorer who is constantly seeking out new land to discover. For experts, such exploration targets unfamiliar sources of sensation (e.g. aromas, bouquets, flavours, terroirs, ages, faults) for themselves and possibly for other initiates as well.

Several experiments have reported the tendency of experts to scout out new sensations. For example, in a study by Mariño-Sánchez et al. (2010), Spanish wine tasters perceived more

odours as intense but fewer as irritating compared to the non-trained healthy population. The identification of wine peculiarities, in particular when practicing the olfactory-gustatory skill, involves cognitive skills. Croijmans et al. (2020) have suggested that expertise entails a heightened ability to imagine hidden structures, thus extending the plasticity of cognition which underlies the chemical senses. On this basis, a wine expert can recognise, discriminate and match the peculiarities of a wine much more effectively than a novice. According to Ericsson et al. (2007), a particular kind of practice – a deliberate practice – is imperative to develop expertise. The practice of wine experts essentially concerns revelation, as they strive to identify elements of wine that were previously unknown or difficult for most people to detect independently. Therefore, a wine expert resembles a member of an uncovered sect or, in more politically correct terms, a highly exclusive professional cluster.

In this study, the experts evaluated the tasted wines more highly than the other tasters, which may be considered an indirect compliment to the people who selected the tasted wines. Moreover, the category of experts displayed significantly less divergence in their evaluation scores compared to those of the other tasters. This finding supports the view of experts as a compact cluster and implies that their judgements of wine are generally more reliable than those of other assessors.

Finally, the findings highlight some differences amongst the evaluation styles of experts. In view of this, future research could consider an analysis of between-expert differences.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. Wiley, Hoboken, NJ.

Bilder, C.R., Loughin, T.M. (2014). *Analysis of Categorical Data with R*. CRC Press, Boca Raton, FL.

Bourdieu, P. (2005). *Il senso pratico*. Armando, Roma, IT.

Croijmans, I., Majid, A. (2016). Not all flavor expertise is equal: The language of wine and coffee experts. *PLoS ONE,* **11**(6), pp.1–21.

Croijmans, I., Speed, L.J., Arshamian, A., Majid, A. (2020). Expertise shapes multimodal imagery for wine. *Cognitive Science*, **44**(5), pp.1–16.

Ericsson, K.A., Prietula, J.M., Cokely, T.E. (2007). The making of an expert. *Harvard Business Review*, **85**(7/8), pp. 114–121.

Fabbris, L., Piscitelli, A. (2021). Wine preferences based on intrinsic attributes: A tasting experiment in Alto Adige/Südtirol province. In B., Bertaccini, L., Fabbris, A., Petrucci (eds.) *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the opening conference* (pp. 129–134). Firenze University Press, Firenze, IT.

Manski, C., Lerman, S.R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, **45**(8), pp. 1977–1988.

Mariño-Sánchez, F.S., Alobid, I., Centellas, S., Alberca, C., Guilemany, J.M., Canals, J.M., De Haro, J., Mullol, J. (2010). Smell training increases cognitive smell skills, of wine tasters compared to the general healthy population. The WINECAT Study. *Rhinology*, **48**, pp. 273–276.

Osgood, C.E., Suci, G., Tannenbaum, P. (1957). *The Measurement of Meaning.* University of Illinois Press, Urbana, IL.

Parr, W.P., Heatherbell, D., White, K.G. (2002). Demystifying wine expertise: Olfactory threshold, perceptual skill and semantic memory in expert and novice wine judges. *Chemical Senses*, **27**(8), 747–755.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, AT (http:/www.R-project.org, accessed 15 June 2021).

Prediction of wine sensorial quality: a classification problem

Maurizio Carpita, Silvia Golia

1. Introduction

When dealing with a wine, it is of interest to be able to predict its quality based on chemical and/or sensory variables. There is no agreement on what wine quality means, or how it should be assessed and it is often viewed in intrinsic (physicochemical, sensory) or extrinsic (price, prestige, context) terms (Jackson, 2017). For example, in Golia et al. (2017) it was measured by a global score of quality, ranging from 0 to 100, produced by Altroconsumo, an Italian independent consumer's association, and based on a large set of variables including chemical and sensory variables, as well as variables of context. Cortez et al. (2009) used an indicator, ranging from 0 to 10 with 0 meaning very bad and 10 excellent, obtained from the evaluations of experienced judges who scored the wines.

In this study we started from the Cortez et al. (2009) paper, but we maintained the categorical nature of the variable measuring the wine sensorial quality. The approach to the prediction of this categorical variable followed by Cortez and coauthors makes use of the observed wine quality, but it suffers from the fact that it is necessary to know the wine quality measure. Instead, in this paper we started from the predicted probabilities' record of the categories of the target variable, obtained from the application of the Cumulative Logit Model, and then we applied a classifier in order to predict the final category. This last step is the one of interest for this paper; in fact we will compare the predictive performances of the default method (Bayes Classifier), which assigns a unit to the most likely category, and other two methods (Maximum Difference Classifier and Maximum Ratio Classifier). In order to do that, we will use the data analysed in Cortez et al. (2009) concerning both the white and red variants of the Portuguese "Vinho Verde" wine.

The paper is organized as follows. Section 2 discusses the categorical classifiers used in this study, whereas Section 3 reports the results concerning the prediction of the wine sensorial quality. Conclusions follow in Section 4.

2. The categorical classifiers

As stated in the introduction, the statistical problem of this study refers to the way in which the record of the predicted occurrence probabilities of each of the categories of the categorical target variable is transformed into a single value. The default method is the *Bayes Classifier* (BC), which assigns a unit to the most likely category. BC has the property to minimize, on average, the test error rate (James et al., 2013), so it is the optimal criterion when the accuracy of the classification is the main goal. Nevertheless, BC favors the prevalent category most and when there is not a category of interest but all the categories have the same relevance, it can not be the best choice.

Starting from this observation, in Golia and Carpita (2018, 2020) we have investigated the performances of different categorical classifiers (some of them take into account also the ordinal nature of the target variable) and we have found the so-called *Maximum Difference Classifier* (MDC) promising. In this study we considered MDC and a new classifier denoted as *Maximum*

Ratio Classifier (MRC). Both classifiers are based on the comparison between the predicted probabilities and the sample frequencies and they are defined as follows.

Let pr_i be the predicted probability of the category c_i $(i = 1, 2, \ldots, k)$ of the categorical variable C, and let fr_i be the corresponding frequency computed from observed data. The MDC computes the deviations of pr_i from fr_i and takes the category corresponding to the maximum difference, that is:

$$MDC : \arg \max_{i \in (c_1, c_2, \ldots, c_k)} (pr_i - fr_i).$$

This classifier represents the extension of what proposed by Cramer (1999) for the dichotomous case.

The MRC computes the relative deviations of pr_i from fr_i and takes the category corresponding to the maximum ratio, that is:

$$MRC : \arg \max_{i \in (c_1, c_2, \ldots, c_k)} (pr_i / fr_i).$$

3. The prediction of wine quality

The data under study concern the sensorial quality of the white and red variants of the Portuguese "Vinho Verde" wine (Cortez et al., 2009). The wine quality was measured by a sensory preference variable, from now on denoted as SPV, using a 0-10 scale. For each wine, eleven of the most common physicochemical variables were recorded; they represent the explanatory variables for the SPV, which is the target variable. Table 1 reports the frequencies of SPV scores observed in the white and red wine data sets; not all the available scores were used and some of them own a low frequency.

Table 1: Frequencies of the sensory preferences observed in the white and red wine data sets

Scores	3	4	5	6	7	8	9
White wines	0.004	0.033	0.297	0.449	0.180	0.036	0.001
Red wines	0.006	0.033	0.426	0.399	0.124	0.011	–

The model used to study and predict the occurrence probabilities of each of the categories of the SPV, is the *Cumulative Logit Model* (CLM) (Agresti, 2010), defined as follows. Let Y be a categorical target variable with k ordinal categories $\{1, 2, \ldots, k\}$, and let $\{X_1, \ldots, X_p\}$ be a set of explanatory variables; for the statistical unit s, the CLM has the following form:

$$\text{logit}[P(Y_s \leq i)] = \log \frac{P(Y_s \leq i)}{1 - P(Y_s \leq i)} = \alpha_i + \sum_{m=1}^{p} \beta_m x_{sm}, \quad \text{for } i = 1, 2, \ldots, k - 1.$$

Once estimated the parameters, it is possible to use the model for predictive purposes, so the CLM gives the k predicted probabilities that are passed to the categorical classifier.

In order to evaluate the predictive performance of a classifier, some indicators computed from the confusion matrix can be used. In this study they are: the *Sensitivity* (Sen) of each category, the *Maximum Distance Between Sensitivities* (MDBSen), the *Overall Accuracy* (OvAc), the *Macro Average F1 score* (MAF1) and the *Kappa statistic* (Kappa) (Raschka and Mirjalili, 2019). Sen_i expresses how well the classifier recognizes a unit belonging to the category c_i. MDBSen, defined as:

$$\text{MDBSen} = \max_{i \neq j} |\text{Sen}_i - \text{Sen}_j|,$$

highlights the balanced or unbalanced ability of the classifier to assign a unit to the right category, the lower the MDBSen, the more balanced the classification. The OvAc is the rate of correct classification and it is the indicator maximized by BC. The MAF1 is another indicator to measure the accuracy of the classifier and it is obtained as the average of the F1 scores class-by-class. The choice of MAF1 instead of the weighted average F1 score, is linked to the will to attribute the same relevance to all classes. Kappa is used to measure the agreement between the actual and the predicted classifications of a dataset, while correcting for agreement occurred by chance.

Table 2 reports the value of these statistics computed on the base of the in-sample prediction of the SPV of all the available wines. For the sake of clarity, we added the percentage variation of OvAc, MAF1 and Kappa with respect to the value obtained applying BC in the last three rows.

Table 2: Performance Indicators for white and red wines

	White wines			Red wines		
	BC	MDC	MRC	BC	MDC	MRC
Sen_3	0.000	0.000	0.200	0.000	0.000	0.600
Sen_4	0.031	0.049	0.258	0.000	0.019	0.208
Sen_5	0.493	0.668	0.386	0.739	0.720	0.355
Sen_6	0.757	0.444	0.342	0.614	0.527	0.392
Sen_7	0.222	0.522	0.301	0.271	0.482	0.357
Sen_8	0.000	0.000	0.091	0.000	0.000	0.556
Sen_9	0.000	0.000	0.800			
MDBSen	0.757	0.668	0.709	0.739	0.720	0.392
OvAc	0.527	0.493	0.336	0.593	0.577	0.369
MAF1	0.215	0.229	0.200	0.272	0.285	0.253
Kappa	0.230	0.252	0.131	0.329	0.327	0.161
Var. OvAc	–	-6.429	-36.251	–	-2.740	-37.829
Var. MAF1	–	6.499	-6.912	–	4.781	-6.997
Var. Kappa	–	9.617	-43.184	–	-0.363	-50.891

In the face of an expected but limited reduction in OvAc (6.4% for white wines and 2.7% for red wines), MDC performs better than BC with respect to MAF1 and Kappa and shows more balanced values of the sensitivities, especially for the white wines. MRC outperforms both BC and MDC in terms of balancing the sensitivities, but loses a lot in terms of OvAc and Kappa.

Given that the lowest and highest sensory preferences have low frequency, we merged the first two and the last two categories for both the two varieties of wine, obtaining a SPV on a 5-category ordinal scale for white wine and on a 4-category ordinal scale for red wines.

Table 3 reports the indicators of Table 2 with the exception of the sensitivities of the single categories. The results show the same behaviour observed in Table 2. It is of interest to note that also in this case there are some categories with a low frequency and others that absorb the majority of the statistical units.

4. Conclusions

In this paper we investigated the impact of different classifiers in the capability to predict the wine sensorial quality of the Portuguese "Vinho Verde" wine. We have studied this variable applying the CLM for prediction purposes. We have transformed the prediction of the occurrence probabilities of each of its categories into a single sensory preference through three

Table 3: Performance Indicators for white and red wines after merging some categories

	White wines (5 categories)			Red wines (4 categories)		
	BC	MDC	MRC	BC	MDC	MRC
MDBSen	0.757	0.650	0.157	0.740	0.686	0.370
OvAc	0.528	0.494	0.358	0.602	0.585	0.432
MAF1	0.302	0.332	0.304	0.422	0.445	0.400
Kappa	0.232	0.254	0.153	0.343	0.340	0.225
Var.OvAc	–	-6.381	-32.173	–	-2.703	-28.170
Var.MAF1	–	9.940	0.659	–	5.394	-5.242
Var.Kappa	–	9.798	-33.956	–	-0.969	-34.502

different classifiers, the BD, the MDC and the MRC. The results have shown that, despite an expected but limited reduction of the overall accuracy, the MDC seems to be the suitable categorical classifier in an unbalanced context (that is when some categories absorb almost all the statistical units) and when all the categories have equal importance (i.e. different types of mis-classification do not involve different costs).

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, 2nd ed*. John Wiley & Sons, Hoboken, New Jersey.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, **47**, pp. 547–5.33.

Cramer, J.S. (1999). Predictive performance of the binary logit model in unbalanced samples. *The Statistician*, **48**(1), pp. 85–94.

Golia, S., Brentari, E., Carpita, M. (2017). Causal reasoning applied to sensory analysis: The case of the Italian wine. *Food Quality and Preference*, **59**, pp. 97–108.

Golia, S., Carpita, M. (2018). On classifiers to predict soccer match results, in *ASMOD 2018: Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data*, eds. S. Capecchi, F. Di Iorio and R. Simone, FedOAPress, pp. 125–132.

Golia, S., Carpita, M. (2020). Comparing classifiers for ordinal variables, in *Book of short papers SIS 2020*, eds. A. Pollice, N. Salvati and F. Schirripa Spagnolo, Pearson, pp. 1160–1165.

Jackson R.S. (2017). *Wine Tasting, 3rd ed*. Academic Press.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer, New York.

Raschka, S., Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing, Birmingham.

Tourism of Italians in Italy through crisis and development: the last 15 years, region by region

Fabrizio Antolini, Antonio Giusti

1. Introduction

Tourism is a very important economic activity for many nations and Italy is among those that particularly benefit from it. However, the economic effects determined by tourist flows depend also on their composition. For example, the distinction between international and domestic tourism is important to understand the pattern of the performed expenditure. Similarly, the distinction between tourism within the region and tourism from outside the region is important to improve the programming of certain services (in particular transport) or to evaluate the attractiveness of the area as a tourist destination.

In recent years, many believe that local tourism has increased; however, a precise estimate of the phenomenon is not easy to make. An indirect estimate, using statistical official sources, can be made using the tourist flows within the region, or by observing the trend of hikers as an indirect measure of the phenomenon. Finally, it would be useful to use new information sources (big data), even if their quality does not yet seem to be able to guarantee an adequate representation of the phenomenon.

However, the analysis of external tourist flows is also relevant, since they express the attractiveness of the territories, (in this case regions) as a tourist destination, which can increase or decrease over time, also due to the policies implemented at the territorial level.

The paper examines the tourism of Italians in Italy, in the various regions, from 2006 to 2020, using the origin-destination matrix produced by ISTAT, distinguishing external flows from those within the region. Particular attention will be given to the year 2020, for which economic information will be provided, showing that overall, the tourism sector in Italy has continued to play an important role, despite the pandemic crisis. The choice of arrivals, instead of night-spent, reduces the influence of the specific type of tourism in each region. The initial results appear interesting and have also been summarised using correspondence analysis.

2. The territory and tourist flow among regions: different approaches

The analysis of the measurement of tourism trends presents several problems since it can have very different objectives. On the other hand, whatever the variable considered, at territorial level the tourism phenomenon almost always presents a high degree of variability, even in contiguous areas. This inevitably raises the question of the real usefulness of an aggregate analysis of the tourism phenomenon. While the territorial variable is important, the choice of territorial detail is often conditioned by the availability of data and the sectoral policy competencies of the territories. In our analysis, the detail used is regional, since it is at this level that the prevalence of public policies on tourism is decided, but also because of the greater availability of data, especially economic and social data. On the other hand, there is still a lack of evaluation models at territorial level that can also be used in the monitoring phase of the various measures, using a simplified input-output scheme, in which the two subjects of the hypothetical function can be identified as public tourism expenditure (input) and arrivals or nights-spent (output). The latter two indicators, often used interchangeably, have a very different descriptive capacity (Antolini et al., 2017). For example, a territory that increases its nights-spent more than its arrivals is a territory that succeeds in retaining tourists. This may be because the services offered are competitively priced, or because the territory

Fabrizio Antolini, Università di Teramo, Italy, fantolini@unite.it, 0000-0002-3112-524X
Antonio Giusti, University of Florence, Italy, antonio.giusti@unifi.it, 0000-0001-9804-4578

has been able to exploit its many tourist destinations. On the other hand, as far as policy is concerned, two different approaches can be followed: one aimed at carrying out many small actions appropriately coordinated in different sectors; the other aimed at carrying out a small number of actions concentrated in sectors considered to be a priority on the scale of possible actions. Whichever approach is used, which must of course be identified only after an overall analysis of the territory's strengths and weaknesses, an analytical knowledge of tourist flows cannot be omitted.

If, as described above, the distinction between arrivals and nights-spent is important, for the purposes of analysing tourism flows it is very useful to distinguish between internal and external tourism flows. Reference is often made, especially when describing the impact of Covid-19 on the tourism sector, to proximity tourism. It is defined as that form of tourism characterised by trips relatively close to the place of residence of the visitors. One way to measure proximity tourism is to break down the analysis of regional tourism flows into internal and external. As we shall see, the two flows at regional level in the period considered (2006-2020) have not always shown similar trends. Another important tool for the measurement of movements is represented by big data that in the Covid-19 period represented a powerful tool for the analysis of people's movements on the territory. The reference is to Google Map (https://www.google.com/covid19/mobility/) as it provides information to assess movements in a specific territorial area. For example, in the metropolitan area of Rome in the period July 15th - August 26th, using these data we know that the attendance in the parks increased by 4% compared to the reference period. The use of data for the analysis of tourist flows still presents several critical issues, which could be resolved by better targeting the usability of the data. In fact, these data do not yet allow a comparison between geographical areas with certain tourist vocations (mountain, seaside, cultural, …).

3. The data employed and the descriptive analysis

The breakdown of flows into internal and external uses the data contained in the origin-destination matrix of arrivals at a regional level as recorded in the survey on the movement of accommodation facilities (ISTAT, 2020). This survey, as it is known, is a census survey (Petrei and Manente, 2018; Antolini and Grassini, 2020) which produces information down to the level of detail of individual municipality. The availability of the origin-destination matrix is only at the regional level, while it would be important for it to be available down to the provincial level. In this way, tourism within the region would be mapped very precisely and could also support regional transport planning. The analysis is carried out considering tourist arrivals, as we want to analyse tourist flows from origin to destination and highlight connectivity between regions. This analysis will be carried out at a later stage, however, compared the preliminary analysis which aims to break down regional tourist flows into external and internal.

In Figure 1, we have considered arrivals because our objective is initially to verify whether the two flows always record the same trend at the regional level and, subsequently, through correspondence analysis, to have a measure of the level of similarity existing between each region (R_{ij}). Over the years, the first consideration that can be made is that the tourist flows within the region are always lower than those from other regions, except for Sicily from 2016 to 2017 and Piedmont from 2008 to 2012. However, the difference between internal and external tourism remains the simplest index of attractiveness which in fact is particularly high in some of the regions shown in the figure above (Friuli, Trentino, Tuscany, Umbria, Aosta Valley, Marche).

Overall, internal tourism is less variable than external tourism, partly because it is a more habitual flow of tourists. There is also a statistical measurement problem that affects the level of internal tourism (but not its performance). In fact, only rented houses managed in an entrepreneurial way or those rented with a registered contract for tourism purposes are recorded, while second-owned houses escape the survey. Finally, the level of internal tourism is also conditioned by the large number of the resident population. The combination of these aspects explains both the lower level and the lower variability of the internal tourist flow compared to the external one.

However, some regions show variability regarding internal flows, for example, Lazio,

Campania, Apulia, Calabria (only from 2016) and Sardinia. Among these, Campania is the only region to have an extremely irregular trend as regards internal flows, in particular decreasing from 2008 to 2016 and increasing from 2016 to 2019, but in any case not far from the level of flows recorded in 2006 and 2007. The trend can be explained by the economic crisis that was particularly incisive at a regional level, and in fact the trend of internal tourist flows takes on the same trend as GDP, particularly over the years 2012-2013 where GDP fell by 0.8 and 1.3 percent. Sicily is the only region where internal tourism exceeds external tourism in 2015 and 2016. This is due to the decrease in external flows while the internal flow, although slightly growing, remains substantially stable. The dynamics of external flows appear to be conditioned by the lack of infrastructure, which makes it difficult to reach and physically move around the territory.

Figure 1 –Arrivals of residents in Italy (from the region or from other regions)

Source: Our processing on ISTAT data.

In addition, the small and sometimes negative difference (2016-2017) between flows from outside and inside the region indicates Sicily's low attractiveness. Finally, it should be noted that in years when there is a decrease in the external tourist flow in Sicily (e.g., 2016-2017), those from neighbouring regions increase in Basilicata, Apulia, and Calabria. Finally, the decrease in external flows in 2012 can be traced back to the country's recessionary crisis at that time.

A special analysis deserves the year 2020 where, due to the pandemic, there was a change in the levels and composition of tourist flows. If we analyse the summer period (July-September 2020), the number of visitors to establishments decreased in trend terms by 36.1 per cent. This was due to the sharp drop in foreign visitors, who fell by 39.7 per cent, even though the overall flow

was made up of the lower outflow of residents who poured into the country. Italian tourists accounted for 86.2 per cent of the total (ISTAT, 2020). Moreover, again analysing the tourist flows, with reference to the overall data just mentioned, it must be remembered that the motivation is almost entirely due to leisure travel, while the missing part is due to business travel, although in the July-September period it usually has a lower impact. Regarding the impact of Covid-19 on the productive sectors, an important proxy is the indicator related to the opening of VAT numbers. Compared to 2019 in the section of economic activity relating to accommodation and catering and sports and entertainment activities, the contraction was 34.1 per cent and 33.5 per cent respectively. As is well known, both activities represent an important part of the tourism industry. The situation remains difficult in 2021 and in the period January-March, considering the same economic activities, the contraction was 25.3% and 4.7% respectively (Ministry of Economy and Finance, 2021).

## 5. Some first results with the correspondence analysis

To get a synthetic picture of the dynamics we are considering, we decided to use the correspondence analysis (Benzécri, 1973), a technique of multivariate statistical analysis of an exploratory nature, which allows us to analyze the existence of association patterns between qualitative variables. As is known, this technique considers each modality of the qualitative variables as an element of analysis; therefore, we will have 20 regions as destination and 20 regions as origin. We have only analyzed four years: 2008 which reflects the recessive economic crisis, 2014 which is the moment of recovery after the second economic crisis occurred in 2012, 2019 because it is an extremely positive year for Italian tourism, and 2020 which represents the time of the pandemic.

In figure 2, we have the projection of the 20 Italian regions in a graph with two factorial axes. In particular, the regions as origins, in blue with the indication of the year, and the region as destination, in red. In this first reading we will only look at the regions as destination.

In 2008, a year of relative crisis for tourism, the graph shows three alignments, which allow us to identify three different groups of regions, which, limiting ourselves to the destinations, are: 1) Calabria, Basilicata, Campania, Molise, Apulia, Abruzzo, Umbria, and Lazio; 2) Piedmont, Aosta Valley and Liguria; 3) Friuli-Venezia Giulia, Trentino Alto Adige and Veneto. Marche, Tuscany and Lombardy are close to the center of gravity.

In 2014, a year of recovery after the previous crises, the graph is more compact. Campania moves away from the center of gravity, followed at a distance by Calabria, Molise, Basilicata, and Apulia; while Piedmont, Aosta Valley and Liguria show another alignment.

Figure 2 – Some results of the correspondence analysis

2008

2014

243

| | 2019 | 2020 |
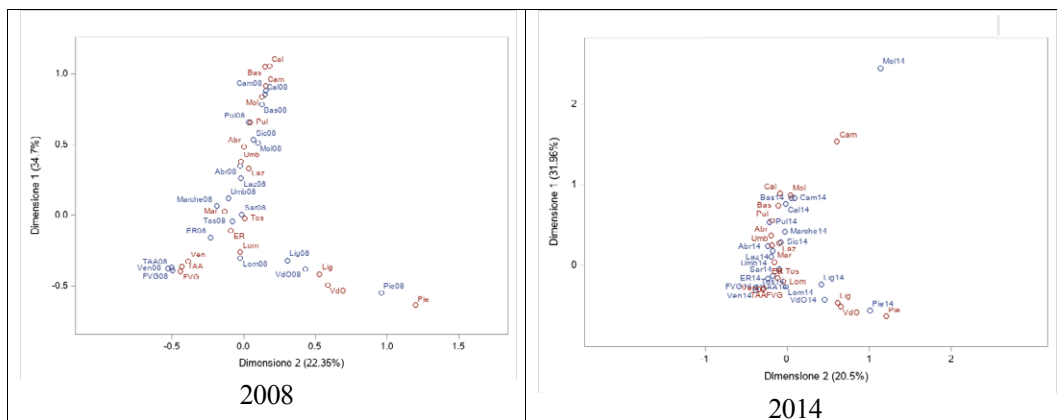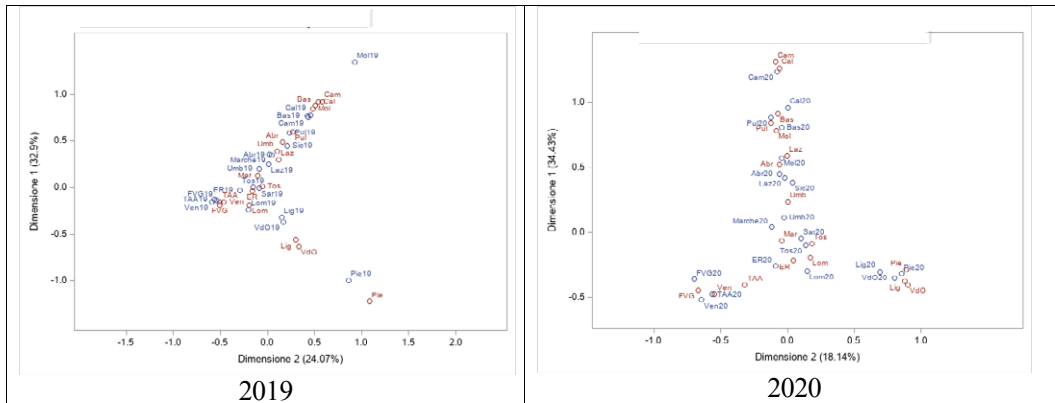
Source: Our processing on ISTAT data.

We can consider 2019 a year of development that could mark a real growth of our tourism. Campania, Calabria, Basilicata, Molise, with Apulia, Abruzzo, Umbria, and Lazio reconstitute an alignment already seen, as do Piedmont, Aosta Valley and Liguria.

Finally, 2020, with the health crisis, brings us back to a situation like that of 2008 with three alignments, made up almost as much as then. On this aspect, the analysis is still in progress and will require evaluating the significance of the factorial axes.

6. A final remark

The work is still in progress and will require a careful analysis of the reasons that have led to a different role for the Italian regions in the context of domestic tourism.

As we have mentioned, the choice of domestic tourism was made considering it more stable than international tourism, which is more easily affected by positive or negative conditions (health, economic, political, etc.).

The choice of arrivals, rather than nights-spent, was done to make the analysis less dependent on the type of tourism, considering the variability of the average stay based on the type of destination (sea, mountains, countryside, tourist cities, etc.).

References

Antolini F., Grassini L. (2020). Issues in Tourism Statistics: A Critical Review. *Social Indicators Research*, **150**, 1021-1042.

Antolini, F., Giusti, A., Grassini, L. (2017). L'importanza di una corretta programmazione territoriale. Attrattività dei territori e dei flussi turistici. *Turistica Italian Journal of Tourism*, **4**: 5-35.

Benzécri, J.P. (1973). *L'Analyse des Données. Volume II. L'Analyse des Correspondances.* Dunod. Paris.

ISTAT, (2020). Movimento Turistico in Italia, *Statistiche Report*, 29/12/2020.

Ministry of Economy and Finance (2021) *Osservatorio delle Partite IVA, sintesi dei dati del 2020.* https://www1.finanze.gov.it/finanze3/osiva/contenuti/Sintesi_annuale_dati_2020.pdf?d=1613137800.

Petrei, F., Manente, M. (2018). Il turismo in Italia nel 2017 e nel 2018: risultati, tendenze e previsioni, 2018. In Becheri, E., Micera R., Morvillo A. (a cura di), *Rapporto sul turismo italiano XXI edizione*, Rogiosi Editore (ISBN 978-88-6950-225-5).

Assessment of visitors' perceptions in protected areas through a model-based clustering

Annalina Sarra, Adelia Evangelista,Tonio Di Battista

1. Introduction

Protected areas are well-defined geographical spaces that, in view of their recognized, natural, ecological or cultural values, receive protection. They have the twofold mandate of protection of natural resources and cultural values and providing a space for nature-based tourism activities, including, among others, mountain hiking, nature photography, bird and animal watching. In the last years, the nature-based tourism is experiencing a positive and sustainable growth worldwide, making it an important sector of the tourism activity, with substantial impacts on the environment, economy and local communities. As broadly highlighted in literature, visitors' experience can be deemed a complex interaction between people and their internal states, the activity they are undertaking, and the social and natural environment in which they find themselves (Leung et al., 2008). Understanding the value attached by visitors to their destination and know their assessment on various activities in which they are engaged during their stay is a key element in shaping their satisfaction. A number of studies have shown that visitor's satisfaction is essential to boost demand, since it increases intention to revisit and recommend the destination to other people (see, among others,Sangpikul (2018); Su et al. (2016)). Besides, a greater knowledge of needs and perception of different visitor-groups should lead to improved management and marketing strategies and to more targeted provision of facilities. In this study, we focus on analyzing the perceived value of visitors who had a specific experience in the Majella National Park, located in the Abruzzo region (Italy). The research data were collected by means of a structured questionnaire administrated to people who visited the sites of the protected area during the last three summer months of 2020. A total of 151 valid questionnaires were obtained and form the base of the data analysis. Our aim is to assess the views and preferences of visitors of that protected natural space in relation to specific profiles. To this end, through a Bayesian model-based clustering, better known as Bayesian Profile Regression, we partitioned visitors into clusters, characterized by similar profiles in terms of their demographic characteristics (age, gender, education attainment, professional activity, origin area), as well as, in terms of the features of their travel behaviour (accommodation, length of stay, past visitation experience). The benefit of the followed approach lies in the ability of that Bayesian technique of simultaneously estimating the contribute of all covariates to the outcome of interest. In our context, we explore the association of detected groups with visitors' satisfaction. In the survey, the global quality of tourism service is segmented into single features and respondents were asked to give their level of appreciation on a five-point Likert satisfaction scale. To estimate the latent trait measured by the items and related to the overall satisfaction, we followed an IRT modelling. The rest of the paper is organized as follows. In Section 2 we describe the study area and the data collected. Section 3 is devoted to illustrate the methodology adopted. Finally, in Section 4 are presented the main results.

2. Study case and data

Majella National Park (MNP) is a protected area located in the provinces of Chieti, Pescara and L'Aquila, in the region of Abruzzo, Italy. It was established in 1991 and it extends over an area of about 74,000 hectares, comprising the mountains of Majella and Morrone. The mountains dominate the territory of this national park: as a matter of fact, the 55% of it is over the 2,000 meters. It includes wide lands with particular wilderness aspects, the rarest and most precious part of the biodiversity national heritage. The diversity of the environments, the richness of nature, the evidences left by the human presence make Majella protected area attractive for visitors, potentially involved in different activities, ranging from visits to hamlets and hermitages, climbing and trekking excursions to participation to traditional festivals. In this study, a sample of visitors was intercepted through a non-probabilistic design. At the end of the survey period, a total of 151 valid questionnaires were returned and form the basis of the data analysis reported herein. The questionnaire, other than to include a few questions with respect to respondents' background (age, gender, education, professional activity, origin area), is organized in two sections devoted to investigate different aspects. Part 1 controls for travel behaviour characteristics of respondents (accommodation type, length of stay, past visitation experience, daily average expenditure) and their expectations. Since visitors are increasingly demand high quality recreational opportunities and the service that support them, the second section of the questionnaire deals with the satisfaction. The satisfaction scale contains 23 items, corresponding to different aspects of tourism experience (staff, food, excursions, outdoor activities, accommodation, information services, naturalistic and historical heritages, hospitality of local population, safety and sustainability, sanitation). Respondents were asked to indicate the degree to which they agree with each item on a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). In our sample, the majority of the visitors surveyed were men (55%), from outside region (68%), aged between 36 and 45 (40%). Half of them had a university education. Regarding their professional activity the respondents were mainly employees (29%). The main reasons that encourage people to visit the Majella National Park are relax and contact with nature, they are in fact curious to carry out guided tours, and they make this choice thanks to the testimony of their acquaintances (word of mouth). The tourist offices to which they contacted for information are principally those of the province of Pescara. Visitors are encouraged to return to the places of the Majella National Park; in fact about half of those interviewed have already been there more than 5 times, and there spent more days (at least 1-3 nights). However, the average daily expenditure that tourists expect to expend is between 10-30 euros.

3. Methodology

3.1 IRT model for polytomous response data: the Graded Response Model   Item response theory (IRT), initially developed in the 1960s, comprises a versatile family of measurement models concerned with the measurement of an individual's latent trait assessed indirectly by a group of items (de Ayala, 2009). The basic idea behind IRT is that the structure in the manifest responses is explained by assuming the existence of one or more latent traits (θ). The mathematical characteristics of IRT models allow a transformation from binary or ordinal answer pattern, e.g. Likert type data, into measure on an equal-interval scale. In this study, the parametric Graded Response Model (GRM; Samejima (1969)) was applied to analyze the ordered response categories of the satisfaction scale included in the questionnaire administered to tourists. In the GRM, items are described by a discrimination parameter (λ) and two or more location parameters (γ). Graded-response model item parameters are easily interpretable: the location parameters locate the point at the latent trait continuum where is a 50% chance of scor-

ing at or above category c_k of item k whereas the discrimination parameter reflects the degree to which the item is related to the underlying latent trait and can differentiate among persons with different trait levels. Specifically, in the GRM the probability that a person's response falls at or above a particular ordered category c ($c = 1, \ldots, C_\kappa$), given the latent trait θ, may be expressed as follows:

$$Pr(Y_{ij} = c | \lambda_k, \theta_i, \gamma_k) = \Phi(\lambda_k \theta_i - \gamma_{k,c-1}) - \Phi(\lambda_k \theta_i - \gamma_{k,c}) \tag{1}$$

Eq. 1 describe the normal ogive formulation for the unidimensional two-parameter GRM model, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution functions and c represent the k ordered categories. GRM cannot be identified because it is overparameterized: for each item a set of $k-1$ location parameters along k item slope (λ) are to be estimated. To overcome this issue some restrictions on parameters are necessary.

3.2 Model-based clustering: the BPR   In this work, we focused on tracing the profile of the tourists who visited the Majella National Park, considering in the analysis, as covariates, the socio-demographic background, vacation habits, typically activities at destinations. To this end, we opted for a cluster-based method, better known as Bayesian Profile Regression (BPR)(Molitor et al., 2010). Profile regression is a Bayesian cluster method that, by capturing the heterogeneity among covariates, allows both identifying specific covariate profiles that are representative of a population (i.e. cluster) and associating them with the outcome of interest (in our case tourists'satisfaction) via a regression model. The Bayesian aspect of this clustering process has some advantages over traditional clustering approaches (e.g. Latent Class Analysis) in that the number of clusters has not to be fixed in advance but it is informed by the data and the model is fitted as a unit, allowing that an individual's outcome potentially influences cluster membership, so that the outcome and the clusters mutually inform each other. Additionally, BPR provides a unified procedure in which the uncertainty associated with clustering is naturally propagated into the regression model and incorporated into posterior inference via MCMC algorithms. Formally, for each individual $i = 1, \ldots, N$, Y_i denotes the outcome of interest and $\mathbf{X}_i = (X_{i_1}, \ldots, X_{i_P})$ represents the covariate profile. The joint probability model for the outcome Y_i and the predictors $\mathbf{X}_i$ can be written as an infinite mixture model (Molitor et al., 2010; Hastie et al., 2013)

$$f(Y_i, \mathbf{X_i} | \mathbf{\Theta}) = \sum_{c=1}^{\infty} \psi_c Pr(\mathbf{X}_i | \mathbf{\Theta}_c, \mathbf{\Theta}_0) Pr(Y_i | \mathbf{\Theta}_c, \mathbf{\Theta}_0). \tag{2}$$

The probability model in equation (2) consists of two sub-models:
the *profile sub-model*, $Pr(\mathbf{X}_i | Z_i, \mathbf{\Theta}_{Z_i}, \mathbf{\Theta}_0)$, and the *response sub-model*, $Pr(Y_i | Z_i, \mathbf{\Theta}_{Z_i}, \mathbf{\Theta}_0)$. For each mixture component, the probability models for the outcome Y_i and the profile $\mathbf{X}_i$ are independent conditionally on some component specific parameters $\mathbf{\Theta}_c$ and some global parameters $\mathbf{\Theta}_0$. In the BPR approach, in order to make inference, an additional allocation parameter Z_i is introduced such that $Z_i = c$ indicates that individual i is assigned to component c and $Pr(Z_i = c) = \psi_c$, with ψ_c the mixture component weight. As pointed out in Molitor et al. (2010), it is possible to approximate the infinite mixture model with a finite one, by specifying a maximum number C of components. The mixture weights, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_C)$, are modelled according to a stick-breaking prior (Ishwaran and James, 2001).

Table 1: Two-parameter GRM estimates

Item	Discrimination parameter λ_j	Thresholds γ_1	γ_2	γ_3	γ_4
Cleaning	2.20	-1.90	-1.19	-0.37	0.53
Signposting	2.92	-1.63	-1.09	-0.35	0.53
Places accessibility	3.14	-1.83	-1.11	-0.41	0.56
Web site	3.52	-1.81	-1.07	-0.36	0.46
Park info	3.92	-2.13	-1.10	-0.31	0.33
Guided tours	3.90	-1.71	-1.06	-0.29	0.40
Public transport services	2.27	-1.09	-0.15	0.42	1.29
Cultural events	2.25	-2.12	-1.00	0.03	1.01
Catering quality	1.85	-3.08	-1.90	-0.95	0.28
Food & wine products	2.17	-2.12	-0.97	-0.57	0.05
Children services	2.88	-1.73	-0.79	0.08	0.90
Path maintenance	3.38	-1.66	-0.98	-0.12	0.54
Accommodation facilities	2.80	-2.25	-1.27	-0.54	0.50
Hospitality of local population	2.07	-2.84	-1.89	-0.89	-0.01
Naturalistic heritage	4.08	-1.86	-1.31	-0.59	0.38
Historical cultural heritage	3.33	-1.82	-1.31	-0.43	0.56
Environmental education center	3.23	-1.82	-1.03	-0.11	0.66
Sanitation	2.84	-1.66	-0.95	-0.19	0.92
Reception center	5.09	-2.05	-1.25	-0.56	0.26
Promotion park activities	4.15	-1.69	-1.17	-0.43	0.39
Park staff	2.69	-2.05	-1.55	-0.96	-0.12
Park staff's knowledge of foreign languages	2.25	-2.57	-1.59	-0.47	0.42
Information material area	2.92	-2.31	-1.61	-0.91	-0.07

4. Results and conclusion

In this section, we first present the parameter estimates obtained by fitting the GRM to the items of visitors' satisfaction scale. All data were analyzed in the R programming environment (R Core Team, 2020) with mirt package (Chalmers, 2012). Table 1 shows the estimates for discrimination and threshold parameters. Discrimination estimates for the items ranged from 1.85 to 5.09, indicating that all items discriminate well between low and high levels of satisfaction: higher values indicate better discrimination. Specifically, the inspection of discrimination parameter estimates suggests that the key indicators in distinguishing visitors with different satisfaction levels are those related to appropriate promotion and guided tours. Also, play a fundamental role in discriminating visitors scoring high and low on the latent satisfaction trait, the item ascertaining the appreciation of the park naturalistic heritage. Likewise, the items that had a smaller discriminative power are those linked to the quality of catering, food and wine products, the hospitality of local population, park staff's knowledge of foreign languages. As for threshold parameters, it is worth noting that they reflects the cut-points between the five item categories. Each of them, mirrors the probability of scoring above or below a given cut-point. In other terms, the thresholds can be thought of as being on the same scale as the z-scale, where a normal distribution is centered at zero with a unit standard deviation metric. By comparing thresholds values across items, we see that, for example, item related to "catering quality" has the lowest initial threshold value of -3.08 and item referring to "pubblic transport services" has the largest initial threshold value of -1.09. This result indicated that fewer people endorsed the first item response category for the item related to "catering quality" compared to the item concerning "pubblic transport services". After estimating the visitors' satisfaction by means of IRT modelling, the next step in our data analysis was the identification of specific visitor-groups. The BPR was fitted using the R package PReMiuM (Liverani et al., 2015). The BPR has pro-

Figure 1: Characterization of visitors'satisfaction profiles associated to each cluster

Legend: modalities code of each categorical variable included in the profile sub-model

Area of origin (0= Abruzzo; 1= Other Regions); Gender (0= Male; 1= Female); Age (0= less than 25 years, 1= 25-45 years; 2= 46-65 years; 3= more than 65 years); Education (0= upper secondary school; 1= Degree); Job (0= Self-employed; 1= Public-employee; 2= Other professions); Reasons (0= Contact with nature; 1= Visit to the historical, artistic and cultural heritage; 2= Relax); Expectation (0=; 1=; 2=); Knowing (0= Personal reccomandation; 1=Park website; 2=Guidebook); Tourist office (0= Chieti; 1= Pescara; 2= L'Aquila); Number of previous visits (0= Once; 1= 2-5 times; 2= more than 5 times); Number of overnight stays (0= None; 1= 1-3 overnights; 2= 4-7 overnights; 3= more than 7 overnights); Average daily expenditures (0= less than 30€, 1= 30-50€; 2= more than 50€).

duced a partition of visitors into 3 clusters: each of them is characterized by similar covariates profile as well as the same satisfaction level. In order to delineate the visitors' specific profile we can refer to the characterization of each cluster in terms of covariates, as illustrated in Figure 1. On the left panel of each figure, the MCMC posterior draws of the satisfaction level of the identified clusters are given; conversely, for categorical variables such those considered in this study, the right panel of each figure shows the posterior distributions of the probability that an explanatory variable appears with one of the discrete categories across the identified groups. Note that each column corresponds to one covariate and cluster labels are specified on horizontal axis. The different colours of box-plots (blue, green and red) refer to a 95% credible interval respectively under, within or upper the global average on all visitors (whatever the cluster). The order of cluster representation follows the order of the associated estimated visitors'satisfaction level of each cluster. A close analysis of Figure 1 reveals that in the typical profile of cluster associated with the highest satisfaction level there is a prevalence of visitors coming from other regions, aged 25-45 years, who had never been before in the Majella National Park area and who have decided to stay overnight from 4 to 7 days. On the other hand, among the visitors who

exhibited lower level of appreciation of the natural area, we find a greater number of Abruzzo resident people, for whom the word of mouth has had a key role in their decision making process to choose that tourist destination. Additionally, both the number of previous visits (more than 5) and the overnight stays (more than seven) have contributed to negatively shape the visitors'satisfaction. The results of this study might have practical implications for managers of protected areas giving them useful insights on how elaborate programs according to visitors' profile. To our knowledge this research represents the first attempt of identifying clusters of visitors with similar covariate profiles through a Bayesian approach based on Dirichlet modeling mixture techniques. Along this benefit, it is important to stress the major limitation of this work concerning the selection of sample units, intercepted through a non-probabilistic design. As a result, we are not able to infer the actual visitors' flows over all the different seasons of the year.

References

Leung, Y., Marion, J., Farrell,T. (2008). Recreation ecology in sustainable tourism and ecotourism: A strengthening role, in *Tourism, Recreation and Sustainability: Linking Culture and the Environment, 2nd Edition*.Wallingford, UK, pp. 19–37.

Sangpikul, A. (2018). The effects of travel experience dimensions on tourist satisfaction and destination loyalty: the case of an island destination, *International Journal of Culture, Tourism and Hospitality Research*, **12** (1), pp. 106–123.

Su, L., Swanson, S. R., Chen, X. (2016). The effects of perceived service quality on repurchase intentions and subjective well-being of Chinese tourists: The mediating role of relationship quality, *Tourism Management*, **52**, pp. 82–95.

de Ayala, R. J. (2009). *The theory and practice ofitem response theory*.New York: Guilford Press.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores, *Psychometrika Monograph Supplement*, **34** (17), pp. 100-114.

Molitor, J., Papathomas, M., Jerrett, M., Richardson, S. (2010). Bayesian profile regression with an application to the national survey of children's health, *Biostatistics*, **11**, pp. 484–498.

Hastie, D., Liverani, S., Azizi, L., Richardson, S., Stücker, I. (2013). A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer, *BMC Medical Research Methodology*, **13** (129).

Ishwaran, H., James, L. (2001). Gibbs Sampling Methods for Stick-Breaking Priors, *Journal of the American Statistical Association*, **96** (453), pp. 161–173.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: http://www.R-project.org/

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment, *Joournal of Statistical Software*, **48** (6), pp. 1–29.

Liverani, S., Hastie, D., Papathomas, M., Richardson, S. (2015). PReMiuM: An R Package for Profile Pegression Mixture Models Using Dirichlet Processes, *Journal of Statistical Software*, **64** (7), pp. 1–30.

This book includes 40 peer-reviewed short papers submitted to the Scientific Conference titled *Statistics and Information Systems for Policy Evaluation*, aimed at promoting new statistical methods and applications for the evaluation of policies and organized by the Association for Applied Statistics (ASA) and the Dept. of Statistics, Computer Science, Applications DiSIA "G. Parenti" of the University of Florence, jointly with the partners AICQ (Italian Association for Quality Culture), AICQ-CN (Italian Association for Quality Culture North and Centre of Italy), AISS (Italian Academy for Six Sigma), ASSIRM (Italian Association for Marketing, Social and Opinion Research), Comune di Firenze, the SIS – Italian Statistical Society, Regione Toscana and Valmon – Evaluation & Monitoring.

Bruno Bertaccini is Associate Professor of Statistics at the "G. Parenti" Department of Statistics, Computer Science and Applications of the University of Florence. After earning his PhD in Applied Statistics, he became an expert in the effectiveness assessment of public policies, with particular attention to the evaluation of the quality of higher education and academic policies.

Luigi Fabbris is a freelance researcher in the field of statistics and social science; President of the ASA, (Italian) Association for Applied Statistics; formerly Professor of Social Statistics at the University of Padua; he has authored or co-authored more than 400 papers or books.

Alessandra Petrucci is Professor of Social Statistics and she holds a PhD in Applied Statistics. Her research interests include survey sampling methods, spatial statistics, multivariate statistical analysis applied to social and environmental issues. She is the author and co-author of numerous scientific papers published in national and international journals, and she has been a member of research projects and scientific committees for several national and international conferences.