IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment

BOOK OF SHORT PAPERS

Editors: Rosaria Lombardo, Ida Camminatiello and Violetta Simonacci

Book of Short papers
10th International Conference **IES 2022**
Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment

Department of Economics, University of Campania "L. Vanvitelli",
January 27th - 28th 2022

Scientific Committee of the group of the Italian Statistical Society on Statistics for the Evaluation and Quality of Services – SVQS

Pietro Amenta -University of Sannio
Matilde Bini -European University of Roma
Luigi D'Ambra -University of Naples "Federico II"
Maurizio Carpita -University of Brescia
Paolo Mariani -University of Milan "Bicocca"
Marica Manisera -University of Brescia
Monica Palma -University of Salento
Pasquale Sarnacchiaro -University of Rome Unitelma Sapienza

Program Committee of the conference IES 2022

Chair: Rosaria Lombardo -University of Campania "L. Vanvitelli"
Fabio Bacchini -ISTAT
Laura Baraldi -University of Campania "L. Vanvitelli"
Eric Beh -University of Newcastle, Australia
Wicher Bergsma -The London School of Economic and Political Science, UK
Enrico Bonetti -University of Campania "L. Vanvitelli"
Eugenio Brentari -University of Brescia
Clelia Buccico -University of Campania "L. Vanvitelli"
Rosalia Castellano -University of Naples "Parthenope"
Ida Camminatiello -University of Campania "L. Vanvitelli"
Carlo Cavicchia -University of Rotterdam, The Netherlands
Enrico Ciavolino -University of Salento
Corrado Crocetta -University of Foggia
Claudio Conversano -University of Cagliari
Antonello D'Ambra -University of Campania "L. Vanvitelli"
Antonio D'Ambrosio -University of Naples "Federico II"
Alfonso Iodice D'Enza -University of Naples "Federico II"
Tonio Di Battista -University of Chieti "G. D'Annunzio"
Michele Gallo -University of Naples "L'Orientale"
Francesco Gangi -University of Campania "L. Vanvitelli"
Michele La Rocca -University of Salerno
Amedeo Lepore -University of Campania "L. Vanvitelli"
Riccardo Macchioni -University of Campania "L. Vanvitelli"
Filomena Maggino -University of Rome "La Sapienza"
Angelos Markos -Demokritus University of Thrace, Greece
Lucio Masserini -University of Pisa
Stefania Mignani -University of Bologna
Nicola Moscariello -University of Campania "L. Vanvitelli"
Francesco Palumbo -University of Naples "Federico II"
Alessandra Petrucci -University of Florence
Alessio Pollice -University of Bari

Donato Posa -University of Salento
Luca Secondi -University of Tuscia
Amalia Vanacore -University of Naples "Federico II"
Michel van de Velden -University of Rotterdam, The Netherlands
Rosanna Verde -University of Campania "L. Vanvitelli"
Donatella Vicari -University of Rome "La Sapienza"
Grazia Vicario -Polytechnic University of Turin
Maurizio Vichi -University of Rome "La Sapienza"

Organizing Committee

Ida Camminatiello -University of Campania "L. Vanvitelli"
Antonello D'Ambra -University of Campania "L. Vanvitelli"
Rosaria Lombardo -University of Campania "L. Vanvitelli"
Elvira Romano -University of Campania "L. Vanvitelli"
Luca Rossi -University Cusano
Violetta Simonacci -University of Naples "L'Orientale"

Editors

Rosaria Lombardo - University of Campania “L. Vanvitelli”, Italy
Ida Camminatiello - University of Campania “L. Vanvitelli”, Italy
Violetta Simonacci - University of Naples “L’Orientale”, Italy

PKE - Professional Knowledge Empowerment s.r.l.
Sede legale: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Sede operativa: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Ufficio Di Rappresentanza: Via Giacomo Peroni, 400 - 00131 Roma (RM)
CF / P.I. 03167830920 — www.pke.it; e-mail info@pke.it — Privacy

February 2022 PKE s.r.l.
ISBN 978-88-94593-35-8 on print
ISBN 978-88-94593-36-5 online

PKE Publisher

Multiple factor analysis with external information on PISA survey data

Analisi Fattoriale Multipla con Informazioni Esterne sui dati dell'indagine PISA

Violetta Simonacci, Marina Marino, Maria Gabriella Grassia and Michele Gallo

Abstract OECD-PISA survey data include performance measurements, expressed as tables of plausible values, and a variety of socio-biographical information. Such data, if properly modeled, can provide useful insights on the causes of low performing students. A methodology which deals with multivariate sets of plausible values and investigates the effects of context variables without assumptions is required. Here Multiple Factor Analysis with External Information is proposed. Specifically, after defining context variable groupings, a partitioning of the variability structure of the data tables is carried out using projection operators than a simplified Multiple Factor Analysis with bootstrap is performed.

Abstract *I dati delle indagini OCSE-PISA includono misurazioni della performance, espresse come insiemi di valori plausibili, e una varietà di informazioni socio-biografiche. Tali dati, se opportunamente modellati, possono fornire utili spunti sulle cause dello scarso rendimento degli studenti. Si rende necessaria dunque una metodologia che si occupi di insiemi multivariati di valori plausibili e indaghi gli effetti delle variabili di contesto. Qui è proposta l'Analisi Fattoriale Multipla con Informazioni Esterne. In particolare, dopo aver definito i raggruppamenti delle variabili di contesto, è eseguito un partizionamento della struttura di variabilità delle tabelle mediante operatori di proiezione, poi è eseguita un'Analisi Fattoriale Multipla semplificata con bootstrap.*

Violetta Simonacci
Dept. of Social Science, University "Federico II", Naples, Italy e-mail: violetta.simonacci@unina.it

Marina Marino
Dept. of Social Science, University "Federico II", Naples, Italy e-mail: marina.marino@unina.it

Maria Gabriella Grassia
Dept. of Social Science, University "Federico II", Naples, Italy e-mail: mgrassia@unina.it

Michele Gallo
Dept. of Human and Social Sciences, University "L'Orientale", Naples, Italy e-mail: mgallo@unior.it

Key words: bootstrap replicates, context variables, MFA, multiset data, plausible values

1 Introduction

The periodic assessment of 15 year old students carried out with the OECD's Programme for International Student Assessment (PISA) reveals a considerable performance disparity among Italian regions. The Campania Region Administration has been actively promoting research programs to investigate the causes behind its low performance.With this purpose, their database of PISA results was made available for exploratory analysis.

The data display a complex structure which includes performance levels in different domains expressed as sets of plausible values and socio-economic, attitude and cultural measures. To provide a useful insight into students performance, a fitting methodology is outlined. Modeling tools were chosen to address two specific research goals: i) studying the multivariate structure of performance by correctly dealing with tables of plausible values; and ii) assessing group-level differences on the basis of context variables, without major assumptions.

To avoid bias estimation of group-level differences PISA performance measurements are not provided as single point estimates for individuals, but rather as ten plausible values randomly drawn for the posterior distributions. Randomly extracted instances are thus provided for all individuals in each domain, yielding ten fully crossed "*student by domain*" tables. Modeling single or averaged plausible values does not take into account uncertainty of sampling or testing unreliability [6]. Any analysis on such data should be carried out on each table separately and, only afterwards, results can be aggregated.

Given these considerations, a bilinear exploratory method which decomposes each table separately and then provides a compromise solution would be a good choice. In this perspective an adaptation of Multiple Factor Analysis (MFA) is proposed. Bootstrap replicates are included in the model to deal with sampling variance.

After choosing an appropriate exploratory model, the second research goal can be addressed. To understand the causes of the performance gap on the basis of the background variables, group differences should be evaluated within the model. This is achieved by extending the methodology of Principal Component Analysis (PCA) with External Information [5] to MFA. This method consists in segmenting the total information of the data into two structures of variability, one explained by the selected background variable(s) (external analysis) and a residual one (internal analysis). Separate PCA are carried out on each variability structure of interest. The adaptation of this method to MFA is straight forward.

To sum up, the aim of this work is to assess the effects of background variables on the multivariate structure of Campania Region PISA results. To do so, an MFA with bootstrap resampling and External Information is carried out. In Section 2 the

database in described in detail; in Section 3 the methodology is outlined and in Section 4 some conclusive remarks and preliminary results are presented.

2 Performance Data: PISA 2018 in Campania

The Campania Region Administration provided a large dataset of PISA responses and measures referring to the 2018 survey. The evaluation was carried out on a sample of 1670 individuals. After excluding students with multiple missing entries, the sample is reduced to 1548 units. The data can be subdivided into performance variables and background information.

Let us focus on performance measures first. In 2018, five domains were assessed: "Problem solving" (Pr _ Sol), "Financial Literacy" (Fin _ L) and the three core domains, "Reading" (*RDN*), "Mathematics" (*MAT*), "Science" (*SCI*). Ten plausible values were imputed for each domain and included in the dataset. In 2018 all domains were scaled in the same way and each set of plausible values was drawn at the same time.

Such data can be arranged into ten tables each of size (1548×5), yielding a two-way matrix for each plausible value imputation. An output similar to a repeated measures design is generated. It is clear that a suitable analysis of such data requires a tool which models performance while taking into account sampling uncertainty by considering all sets of plausible values in the solution. A methodology based on MFA is proposed here and briefly introduced in the following section.

In addition to performance measurements, numerous context variables are also provided for each student. The first step in selecting background variables of interest is to eliminate all the measures with a large amount of missing values and redundant information (items already included in other estimates). A total of 30 variables is selected. Given the large amount of information, after a quick first assessment of significance, only the most relevant results will be presented.

Most of these variables were built as latent constructs based on the responses to multiple items and expressed as Weighted Likelihood Estimates (WLE) on an interval-scale. In the External Information analysis, these quantitative measurements are transformed into dummies to build homogeneous student groupings. Seven groups are identified for each WLE variable based on cut-off values of the index score. The following groups are constructed: "very low score", "low score", "score below the average", "average", "score above the average", "high score", "very high score". Such classification is possible because the variables are standardized with respect to the distribution of OECD countries to have a mean of 0 and a standard deviation of 1. Consequently, a negative score does not mean that a student has answered negatively to a question (or set of questions) but simply that they answered less positively than the average student in OECD countries.

3 Methodology

3.1 MFA for plausible values tables

Multiple Factor Analysis is a technique based on singular value decomposition (SVD) designed for the analysis of $[1,\dots,k,\dots,K]$ tables $\mathbf{X}_k$ which collect sets of variables on the same $[1,\dots,i,\dots,I]$ observations [2, 3]. Generally the K groups of variables differ from one another and each $\mathbf{X}_k$ stores its own $[1,\dots,j_k,\dots,J_K]$ variables. In some instances, however, all $\mathbf{X}_k$ may refer to the same variables $[1,\dots,j,\dots,J]$ measured under different conditions (repeated measures). In the PISA dataset this latter simplified version of MFA is considered but, instead of repeated measures, the tables contain different plausible values imputations.

The procedure is outlined as follows. First, each table is decomposed by (truncated) SVD and scaled by dividing its elements by the first singular value. These scaled tables are then joined in a single wide matrix which is also subsequently analyzed by SVD. The results include common scores and loadings, generally known as compromise or consensus, and partial factor scores for each of the K tables. Formally, these subsequent steps are executed.

1. Truncated SVD is performed on each $\mathbf{X}_k$ to retrieve the first singular value

$$\text{SVD}(\mathbf{X}_k) = \mathbf{u}_k \sigma_k \mathbf{v}_k' \qquad with \quad 1,\dots,k,\dots,K \tag{1}$$

   σ_k is the first singular value of $\mathbf{X}_k$; $\mathbf{u}_k(I\times 1)$ and $\mathbf{v}_k(J\times 1)$ are the first left and right singular vectors, respectively.

2. The singular values $\sigma_1,\dots,\sigma_k,\dots,\sigma_K$ are used to build the matrix of weights $\mathbf{A}(K\times K)$:

$$\mathbf{A} = diag(\alpha_1,\dots,\alpha_k,\dots,\alpha_K) \qquad with \quad \alpha_k = \frac{1}{\sigma_k^2} = \sigma_k^{-2} \tag{2}$$

3. Given $\mathbf{X} = [\mathbf{X}_1 \mid \dots \mid \mathbf{X}_k \mid \dots \mid \mathbf{X}_K]$, a weighted wide matrix $\tilde{\mathbf{X}}$ of juxtaposed K tables can be formulated:

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{A} = [\alpha_1\mathbf{X}_1 \mid \dots \mid \alpha_k\mathbf{X}_k \mid \dots \mid \alpha_K\mathbf{X}_K] \tag{3}$$

4. The wide matrix $\tilde{\mathbf{X}}$ is decomposed in R factors by SVD:

$$\text{SVD}(\tilde{\mathbf{X}}) = \tilde{\mathbf{U}}\tilde{\Delta}\tilde{\mathbf{V}}' \tag{4}$$

   where $\tilde{\Delta}(R\times R)$ is the diagonal matrix of singular values while $\tilde{\mathbf{U}}(I\times R)$ and $\tilde{\mathbf{V}}(J\cdot K\times R)$ are the left and right singular vector matrices respectively. The matrix $\tilde{\mathbf{V}}$ can be expressed as:

$$\tilde{\mathbf{V}} = \left[\tilde{\mathbf{V}}_1' | \ldots | \tilde{\mathbf{V}}_k' | \ldots | \tilde{\mathbf{V}}_K'\right] \tag{5}$$

where the generic matrix $\tilde{\mathbf{V}}_k(J \times r)$ stores the right singular vectors of the corresponding matrix $\tilde{\mathbf{X}}_k$.

At this point the consensus solution can be explored. Compromise scores for individuals are easily found by $\mathbf{F} = \tilde{\mathbf{U}}\tilde{\Delta}(I \times R)$. Compromise loadings can be computed only in the special case of repeated measures. Each right singular vector matrix $\tilde{\mathbf{V}}_k$ in eq.5 is scaled back to its original variability structure so that K rescaled $\mathbf{Q}_k$ matrices are yielded with $\mathbf{Q}_k = \frac{1}{\alpha_k}\tilde{\mathbf{V}}_k$.

Now the compromise factor loading matrix $\bar{\mathbf{Q}}(J \times R)$ with element $\bar{q}_{jr} = \frac{1}{K}\sum_{k=1}^{K} q_{jkr}$ can be defined as the barycenter of the partial factor loadings.

Bootstrap resampling is incorporated into the procedure to provide information on sample variability, following PISA technical repots indications [1, 4].

3.2 Adding external information

To study the impact of various context variables on the structure yielded by MFA, the External Information methodology can be added to the model in the following manner. First, for each external variable a generic matrix $\mathbf{G}$ of dimension $(I \times m)$ with $(m < I)$ can be built, where m is the number of groups defined within the variable. The columns of the matrix are dummies which indicate if a subject belongs to a certain group or not.

Successively the variability structure within each generic k-th table can be decomposed as follows in order to study the effect of $\mathbf{G}$:

$$\mathbf{X}_k = \mathbf{G}\mathbf{B}_k + \mathbf{E}_k \tag{6}$$

where $\mathbf{E}_k$ is the residual matrix of dimension $(I \times J_k)$ which includes internal variability (not explained by $\mathbf{G}$) and $\mathbf{B}_k$ contains the coefficient to estimate by minimizing the sum of squares of residuals $SS(\mathbf{E}_k) = tr(\mathbf{E}_k'\mathbf{E}_k)$. Thus, we have $\hat{\mathbf{B}}_k = \mathbf{P}_G + \mathbf{X}_k$ were $\mathbf{P}_G = \mathbf{G}(\mathbf{G}^t\mathbf{G})^{-}\mathbf{G}^t$ is the orthogonal projection operator generated by $\mathbf{G}$. The decomposition can then be rewritten as:

$$\mathbf{X}_k = (\mathbf{P}_G + \mathbf{P}_G^{\perp})\mathbf{X}_k = \mathbf{P}_G\mathbf{X}_k + \mathbf{P}_G^{\perp}\mathbf{X}_k \tag{7}$$

where $\mathbf{P}_G^{\perp}$ is the orthogonal complement of $\mathbf{P}_G$. It is clear that the first term of the equation contains the external information explained by $\mathbf{G}$ and the second term only the residual one.

In light of these steps, eq. 1 and eq. 3 of the MFA procedure can be modified to focus on external information. Specifically all $\mathbf{X}_k$ tables in the formula must be pre-multiplied by $\mathbf{P}_G$. MFA steps can then be carried out as specified in the previous section.

4 Preliminary considerations

PISA surveys data represent an enormous source of information not only on students' performance, but also on their background with respect to school, family and society. Identifying modeling tools suitable for the complexity of their structure is an ongoing challenge.

The proposed methodology, based on MFA with external information, has been chosen to adequately deal with plausible values tables and to evaluate students' results in relation to their context.

Preliminary results have highlighted interesting effects of the selected external variables on the performance of students. Some of the most interesting context variables are "Sense of anxiety", "Sense of belonging", "Emotional support of parents" and "Family wealth". In reference to the latter measure, for example, it was observed how the first axis explains most of the external variability (more than 80%). Groups of students with low scores record lower levels of performance in all domains. As the "Family wealth" score increases the level of performance also increases. However, this linear trend disappears at the highest level.

Results will be displayed with the support of the symmetrical MFA biplot representation with confidence interval ellipses. Some methodological remarks will also be discussed on the use of projection operators in MFA, specifically on how the choice of the procedure step in which the partition of variability is introduced impacts results.

References

1. Babamoradi, H., van den Berg, F., Rinnan, Å.: Bootstrap based confidence limits in principal component analysis—a case study. Chemometrics and Intelligent Laboratory Systems **120**, 97–105 (2013)
2. Escofier, B., Pages, J.: Multiple factor analysis (afmult package). Computational statistics & data analysis **18**(1), 121–140 (1994)
3. Pagès, J.: Multiple factor analysis by example using R. CRC Press (2014)
4. Pagés, J., Husson, F.: Multiple factor analysis with confidence ellipses: a methodology to study the relationships between sensory and instrumental data. Journal of Chemometrics: A Journal of the Chemometrics Society **19**(3), 138–144 (2005)
5. Takane, Y., Shibayama, T.: Principal component analysis with external information on both subjects and variables. Psychometrika **56**(1), 97–120 (1991)
6. Von Davier, M., Gonzalez, E., Mislevy, R.: What are plausible values and why are they useful. IERI monograph series **2**, 9–36 (2009)