doi.org/10.26398/IJAS.0035-012

SPECIAL ISSUE IN MEMORY OF SIMONA BALBI - EDITORIAL

Jörg Blasius
Department of Political Science and Sociology, University of Bonn, Bonn, Germany

Luigi Fabbris
Tolomeo Studi e Ricerche, Padua and Treviso, Italy

Michael Greenacre
Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain

Germana Scepi, Maria Spano
Department of Economics and Statistics, Federico II University of Naples, Naples, Italy

This special issue is dedicated to Simona Balbi, full professor of Statistics at the University Federico II of Naples, who prematurely died in February 2018.

The issue takes its cue from the scientific meeting in memory of Simona Balbi, "Statistics and Data Science", organized in February 2019 by the Department of "Scienze Economiche e Statistiche" at the Federico II University of Naples. This event aimed to bring together people who have known Simona Balbi and shared part of their scientific path with her but also, often, a bond of friendship and affection. The morning was dedicated to speeches by some foreign colleagues (Michael Greenacre, Jörg Blasius, Ludovic Lebart, Mirelle Summa, Gilbert Saporta, Mónica Bécue). The afternoon included a session of personal testimonies, followed by a session of scientific contributions by Italian colleagues who, with Simona Balbi, have developed part of their research paths (Natale Carlo Lauro, Luigi Fabbris, Corrado Crocetta, Furio Camillo, Rosanna Verde, Francesco Palumbo, Michelangelo Misuraca, Roberta Siciliano, Vincenzo Esposito Vinzi, Giuseppe Giordano, Gabriella Grassia, Salvatore Ingrassia, Germana Scepi, and others).

Contributions regarding three main fields of Simona's scientific research were presented at the conference: *a) the social area, and in particular the education system; b) correspondence analysis and related methods;* and, finally*, c) textual data analysis.*

Simona was aware that statistical methodology is a tool to solve real problems, so any problem unsolved by statistics required a methodological improvement. From 2002 to 2012, she participated in various research projects in which the effectiveness of the higher education system was measured through indicators involving the matching between the education received by university graduates and the expectations of the labour market in Italy. In these projects, financed by the Italian Ministry of Education, Research and University, she and the research group at Naples University studied how to fuse databases at various levels of granularity (graduates, courses, universities, local labor markets) for making comparisons feasible.

Moreover, she organized three scientific meetings on higher education effectiveness at the Federico II University of Naples, in which she presented various papers, jointly with colleagues, and was a guest editor of two books (in Italian) on: "Jobs and graduates' competencies" and "Quantitative representation of an effective educational process that may improve best practices". She had the capacity to softly involve colleagues and young scholars in this field of applied research, being able to fit the content purposes with the statistical methods' potentiality.

The first paper of this special issue, "*Counting the Poor in Italy and EU*" by Luigi Fabbris can be placed in the first area of Simona's scientific research. In this paper, the author examines some approaches to poverty and discusses the properties of poverty measures. The paper suggests an interesting correspondence between poverty measurement approaches and intervention purposes. It is highlighted that the choice of a suitable approach to poverty is based on the pertinence of the properties of the statistical measures with respect to the policies and actions to overcome poverty.

Simona wrote many papers and organized a lot of conferences on the second thematic area: correspondence analysis and related methods. The editors remember the World Congress of Sociology in Bielefeld, 18-23 July 1994, and the second of the conferences on correspondence analysis, called Visualization of Categorical Data, which took place in Cologne, May 17-19, 1995. Simona attended all the conferences which came to be called CARME (Correspondence Analysis and Related Methods). She did not only participate, but she also organized the seventh CARME conference, with Jörg Blasius and Michael Greenacre, which took place in Naples from September 20-23, 2015 (http://www.carme-n.org/carme2015). There is a video of the meeting on the YouTube channel, https://www.youtube.com/watch?v=WxQd1eA4fnw

Simona also organized the RC33 conference (Research Committee on Logic and Methodology of the International Sociological Association). This meeting takes place every four years and has about 500 participants. The Naples meeting was held September 1-5, 2008, and was one of the largest ever, with 88 sessions and over 500 papers. As in the previous two and subsequent conferences, the CARME theme was strongly represented with several sessions, including Automatic Textual Analysis (organized by Simona Balbi), Biplots (John Gower), Correspondence Analysis and Related Methods (Jörg Blasius), Geometric Data Analysis (Brigitte Le Roux), and Multidimensional Scaling (Mark de Rooij), to mention just a few.

The paper titled "*From Plain to Sparse Correspondence Analysis: A Generalized SVD Approach"* by Hervé Abdi, Vincent Guillemot, Ruiping Liu, Ndèye Niang, Gilbert Saporta, Ju-Chi Yu can be placed in the Simona's second research area. The authors propose an extension of the sparse correspondence analysis method developed by Liu et al. (2023) by adding a new global algorithm. This algorithm allows us the simultaneous optimization of the dimensionality of the sparsified space and the sparsification parameters of the rows and columns of the data. The paper discusses the properties of the new version of sparse CA estimates. The method is applied to interpret the relationships in a big textual data set—obtained from the Project Gutenberg (Gerlach and Font-Clos, 2020) — compiling common words used in 100 books each from five book categories: Biographies, Love stories, Mystery, Philosophy and Science Fiction. The results show that sparse correspondence analysis simplifies the interpretation of large tables by highlighting important categories and obtaining simple successive dimensions in the spirit of the simple structure of factor analysis.

Finally, the research topic that accompanied Simona throughout her academic career and that probably fascinated her the most was the statistical analysis of textual data.

Simona was a pioneer in this field, producing various scientific contributions focusing on both methodological and applicative aspects of different text mining tasks. Her early work focused on the use of factorial methods, from studying the stability of configurations obtained with non-symmetrical correspondence analysis to the use of Procrustes techniques for the analysis of multilingual corpora. In some of her work, she addressed topics specific to natural language processing, such as word sense disambiguation, proposing the joint use of correspondence analysis and network analysis tools, and the analysis of complex lexical structures through symbolic data analysis methods.

The application areas where she tested her methodological proposals were the most diverse, from classic open-ended survey responses to newspaper articles, annual reports, job postings, and finally to her latest work where her interest shifted to the growing need to analyze data from social media.

Since 2002, she was a member of the permanent scientific committee of JADT, *Journées internationales d'Analyse Statistique des Données Textuelles*, the biennial conference which has constantly gained importance since its first occurrence in 1992, open to all scholars and researchers working in the field of textual data analysis, including natural language processing and lexicography, text mining, information science, computational linguistics, sociolinguistics, analysis of political discourse and content analysis.

In July 2022, the conference was held in Naples, on the proposal of her colleagues Massimo Aria, Giuseppe Giordano, Michelangelo Misuraca, Germana Scepi, and Maria Spano, with the support of the Vadistat association founded by Simona herself in 2008 to spread statistical culture in the scientific, educational, and institutional fields that today continues its activity with the name "Vadistat for Simona Balbi".

Two articles on this issue can be placed in the textual data analysis area: the first paper is titled "Multilingual textual data: an approach through multiple factor analysis" by Belchin Kostov, Ramon Alvarez-Esteban, Mónica Bécue-

Bertaut and Francois Husson. The authors propose a new approach, the multiple factor analysis for generalised aggregate lexical tables (MFA-GALT), for analysing open-ended questions answered in different languages, in the case of multilingual surveys. MFA-GALT is performed in two steps, exactly like a classic MFA. First, each sub-table is analyzed separately while, in the second step, a global factorial analysis is performed on all sets of multiple tables. The properties and the graphical representations of this new approach are also shown with the application to data collected by a railway company on satisfaction of passengers regarding its night trains.

The second paper in this issue deals with textual data is titled "*Visualization of Textual Data: A Complement to Authorship attribution*" by Ludovic Lebart. In textual data analysis, authorship attribution is precisely a leading case of statistical decision. The author analyses a large corpus of 50 French novels of the 20th century, by comparing descriptive (or unsupervised) methods with confirmatory (or supervised) methods. It is shown that additive trees applied to the coordinates of a preliminary correspondence analysis can provide a useful strategy for describing, comparing, understanding this peculiar data and their relationships.

This issue is dedicated by the guest editors to Simona Balbi and, in the conclusion, we report some of the main papers of Simona in the three previously described research areas.

Simona was a dear friend for all of us and to all who met her, with her quiet manner and sense of humor that will always be remembered, not only her academic work and significant contribution to the field of statistics.

The Guest Editors
Jörg Blasius
Luigi Fabbris
Michael Greenacre
Germana Scepi
Maria Spano

MAIN REFERENCES OF SIMONA'S CONTRIBUTIONS

a) Social area:

Lauro, N., Balbi, S., Mola, F., Perna, A., and Scepi, G. (1991). Indagine sui laureati in Economia e Commercio di Napoli negli anni 1986-1989.

Balbi, S., Lauro, N. C., and Scepi, G. (1994). A multiway data analysis technique for comparing surveys. *Methodologica*, 3, 79-90.

Lauro, N., Balbi, S., and Scepi, G. (1994). The analysis of repeated surveys on Italian manufacturing enterprises: a multidimensional approach. In *Technique and Uses of Enterprise Panels-Proceedings of the First Eurostat International Workshop on Techniques of Enterprise Panels. Eurostat.*

Balbi, S., Balzano, S., and Bruzzese, D. (2002). A conceptual apporach to edit and imputation in repeated surveys. In *Compstat* 2002 (Vol. 400).

Balbi, S., and Grassia, M.G. (2003). Meccanismi di accesso al mercato del lavoro degli studenti di Economia a Napoli, Profiling dei laureati attraverso tre indagini ripetute. In *Transizione Università-Lavoro: la definizione delle competenze,* 97-110. Cleup.

Balbi, S., and Grassia, M.G. (2007). Profiling and labour market accessibility for the graduates in economics at Naples University. *Effectiveness of University Education in Italy: Employability, Competences, Human Capital*, 345-356.

Balbi, S., Grassia, M.G., Nappo, D., Tortora, C., and Triunfo, N. (2010). Come misurare l'efficacia di un Corso di Laurea. In *La rappresentazione quantitativa del processo universitario che genera efficacia e attiva il miglioramento*, 17-32. CLEUP.

Aria, M., Balbi, S., and Piscitelli, A. (2017). Profili ideali di laureati per lavorare nel turismo. Indagine sulle strutture alberghiere di Napoli. In *Scienza e coscienza a 1000 euro al mese. Neolaureati e mercato del lavoro.* Università degli Studi di Padova.

b) *Correspondence analysis and related methods:*

Balbi, S. (1992). On stability in non-symmetrical correspondence analysis using bootstrap. *Statistica Applicata*, 4(4), 543-552.

Lauro, N., Balbi, S., and Scepi, G. (1993). Multidimensional data analysis and experimental design. In *Contributed Papers of 49th ISI Session* (Vol. 1).

Lauro, N., Scepi, G., and Balbi, S. (1993). Empirical confidence regions for multidimensional control charts. In *Contributed Papers 49th ISI Session* (Vol. 2).

Balbi, S. (1994). Influence and stability in non symmetrical correspondence analysis. *Metron*, 52, 111-128.

Balbi, S. (1998). Graphical displays in nonsymmetrical correspondence analysis. In *Visualization of Categorical Data*, 297-309. Academic Press.

c) Textual data analysis:

Balbi, S. (1996). Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms. In *Proceedings of the JADT: 3es JADT, Journées Internationales d'Analyse Statistique des Données Textuelles, CISU, Roma*, *2*, 5-12.

Balbi, S. (1998). Textual data analysis for open-questions in repeated surveys. In *Advances in Data Science and Classification: Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98) Università "La Sapienza", Rome,* 449-456. Springer Berlin Heidelberg.

Balbi, S., and Giordano, G. (2001). A factorial technique for analysing textual data with external information. In *Advances in Classification and Data Analysis,* 169-176. Springer Berlin Heidelberg.

Bolasco, S., Verde, R., and Balbi, S. (2002). Outils de text mining pour l'analyse de structures lexicales à éléments variables. In *Proceedings of the JADT: 6es Journes Internationales d'Analyse Statistique des Données Textuelles, St. Malo,* 197-208.

Balbi, S., Bolasco, S., and Verde, R. (2002). Text mining on elementary forms in complex lexical structures. In *Proceedings of the JADT*: *6es Journées Internationales d'Analyse Statistique des Données Textuelles, St. Malo*, 89-100.

Balbi, S., and Di Meglio, E. (2004). A text mining strategy based on local contexts of words. In *Proceedings of the JADT: 7es Journées Internationales d'Analyse Statistique des Données Textuelles, Louvain,* 4, 79-87.

Balbi, S., and Di Meglio, E. (2004). Una strategia di text mining basata su regole di associazione. In *Applicazioni di analisi statistica dei dati testuali*, 29-40. Casa Editrice Università La Sapienza.

Balbi, S., and Di Meglio, E. (2004). Contributions of textual data analysis to text retrieval. In *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago*, 511-520. Berlin, Heidelberg: Springer Berlin Heidelberg.

Balbi, S., and Misuraca, M. (2005). Pesi e metriche nell'analisi dei dati testuali. *Quaderni di Statistica*, 7, 55-68.

Balbi, S., and Misuraca, M. (2006). Procrustes techniques for text mining. *In Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma*, 227-234. Springer Berlin Heidelberg.

Balbi, S., and Misuraca, M. (2010). A doubly projected analysis for lexical tables. *Advances in Data Analysis: Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data, and Neural Networks*, 13-19.

Balbi, S., Infante, G., and Misuraca, M. (2008). Conjoint analysis with textual external information. In *Proceedings of the JADT 2008: 9es Journées Internationales d'Analyse Statistique des Données Textuelles, Lyon,* 1, 129-136.

Balbi, S., Infante, G., and Misuraca, M. (2009). Il text mining per l'individuazione dell'offerta universitaria di competenze nel terzo settore. *formazione e lavoro*, 95-106.

Balbi, S., and Misuraca, M. (2010). A doubly projected analysis for lexical tables. Advances in *Data Analysis: Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data, and Neural Networks*, 13-19.

Balbi, S. (2010). Beyond the curse of multidimensionality: High dimensional clustering in text mining. *Italian Journal of Applied Statistics*, 22(1), 53-63.

Balbi, S., Crocetta, C., Romano, M. F., Zaccarin, S., and Zavarrone, E. (2011). Competences and professional options of the Italian graduates: Results from the textual analysis of the degree course information data. In *Statistical Methods for the Evaluation of University Systems*, 195-207. Physica-Verlag HD.

Balbi, S., Stawinoga, A., and Triunfo, N. (2012). Text mining tools for extracting knowledge from firms annual reports. In *Proceedings of the*

JADT 2012: 11th International Conference on Statistical Analysis of Textual Data, Vol. 2012, p. 11es).

Balbi, S., and Triunfo, N. (2012). Statistical tools in the joint analysis of closed and open-ended questions. In *Survey Data Collection and Integration,* 61-72. Berlin, Heidelberg: Springer Berlin Heidelberg.

Balbi, S., and Stawinoga, A. (2013). Mining the ambiguity: Correspondence and network analysis for discovering word sense. In *Proceedings of the Conference of the Italian Statistical Society*.

Balbi, S., and Stawinoga, A. (2014). Textual data analysis tools for word sense disambiguation. In *Proceedings of the JADT: 12es Journées Internationales d'Analyse Statistique des Données Textuelles, Paris* (1), 57-66. INALCO Sorbonne Nouvelle.

Stawinoga, A., Balbi, S., and Scepi, G. (2016). Network tools for the analysis of brand image. *Italian Journal of Applied Statistics*, 26, 37-48.

Balbi, S., Misuraca, M., and Spano, M. (2016). A cosine-based validation measure for document clustering. In D. Mayaffre, C. Poudat, L. Vanni, V. Magri, P. Follette (eds.), *Statistical Analysis of Textual Data. Proceedings of 13th International Conference* (JADT16), Presses Fac Imprimeur, Nice, vol. 1, 65-74.

Balbi, S., Misuraca, M., and Scepi, G. (2017). A polarity-based strategy for ranking social media reviews. In *SIS 2017. Statistics and Data Science: New Challenges, New Generations. Proceedings of the Conference of the Italian Statistical Society*, 95-102. Firenze University Press.