EDITORIAL

In 1991, the first international conference ever held on the topic of correspondence analysis was held in Cologne, Germany, hosted by the University of Cologne. This conference was organized by Michael Greenacre and Jörg Blasius, and led to a series of conferences that became known as CARME, which stands for Correspondence Analysis and Related MEthods (for more details, see www.carme-n.org). Since then the conference has taken place every four years, the seventh and most recent in Naples (Italy), September 2015, organised by Simona Balbi (University of Naples Federico II), Jörg Blasius (University of Bonn) and Michael Greenacre (Universitat Pompeu Fabra, Barcelona)

The objective of this conference was to spotlight the very latest research in correspondence analysis and related techniques of multidimensional visualization, as well as to discuss future developments. In general, the aim of CARME is to bring together theoretical and applied researchers in all the areas where correspondence analysis and related methods are currently being used, notably sociology, psychology, education, ecology, archaeology, geology, linguistics, philosophy, genetics, biomedical research, health economics, marketing and management. The conference attracted approximately 100 researchers from 16 countries, and we, the editors, invited several participants to contribute to a special issue of the Italian Journal of Applied Statistics. For all the contributions received, we requested external reviews and, in addition, refereed the papers ourselves. As a result of this reviewing process, we accepted 13 papers for publication.

The first paper of the special issue, given by Jean-Luc Durand (Université Paris 13, France) and Brigitte Le Roux (Université Paris Descartes, France), shows that the strength of the relationship between variables is linked to the variance of the eigenvalues in principal component analysis (PCA) and in multiple correspondence analysis (MCA). Using two historical data sets, the authors examine the association between the eigenvalue variance and the correlations between the variables in PCA or the mean-square contingency coefficients between questions in MCA.

The paper from Jörg Blasius (University of Bonn, Germany), Oleg Nenadić (University of Göttingen, Germany) and Victor Thiessen (Dalhousie University, Canada) shows the applicability of the dirty data index (DDI) for assessing the quality of survey data in international comparisons. The DDI is based on the quantifications from categorical principal component analysis, which works on item batteries with ordered response categories. As an example the authors use data from the European Social Survey in 2012, which includes 36 countries and more than 56,000 cases.

A new development in the analysis of sparse contingency tables (i.e. large tables of counts with a preponderance of zero entries) is given by Vartan Choulakian (Université de Moncton, Canada) in his paper on taxicab correspondence analysis (TCA). This is an alternative geometric method that is based on the weighted L1 norm (sum of absolute values) rather than the weighted sum of squared values as used in CA. This approach has novel properties that enable it to cope with highly sparse data sets. Several examples are given, which highlight the similarities and differences between the two methods.

Analysing several contingency tables having a common set of rows at the same time, one analytical and visualization approach is called simultaneous analysis (SA) of contingency tables. This is the subject of the paper by Enzo Fenoglio (Cisco Systems, France), where he shows how SA can lead to quantitative scores of demographic profiles of TV viewers. These scores provide the essential quantitative inputs to the following stages of the analysis, which involve sequentially logistic regression, mixture modelling and finally audience prediction.

In a free sorting exercise, each subject partitions a set of objects, for example consumer products, into groups of similar ones. El Mostafa Qannari and Evelyne Vigneau (Nantes Atlantic National College of Veterinary Medicine, Food Science and Engineering, France) summarize existing approaches to analysing free sorting data and propose a new association model that admits explanatory variables such as gender or experimental factors and allows hypothesis testing on the results. They illustrate this method using a data set of consumers evaluating 12 luxury perfumes.

Angelos Markos (Democritus University of Thrace, Greece) and Alfonso Iodice d’Enza (University of Cassino and Southern Lazio, Italy) treat the computation of MCA solutions in the area of big data analysis. They face the challenge of adjusting the numerical solution sequentially for the dominant eigenvalues and eigenvectors as additional data become available. This incremental form of adjustment, already studied for principal component analysis, is carried over to the correspondence analysis context of categorical data and tested on artificial data for speed and accuracy.

The variation in a table of frequencies, based on a sample of respondents, is measured in correspondence analysis by the total inertia. Michael Greenacre (Universitat Pompeu Fabra, Catalonia) and Tor Korneliussen (Nord University, Norway) consider the decomposition of this measure in the presence of a factorial structure defined on the sample. Parts of inertia can be attributed to fixed factors such as gender, nationality and age group, as well as their interactions, and these sources of variation can be visualized separately in CA-type maps. Their approach is applied to the information search behaviour of European tourists, using data from the Flash Eurobarometer survey on tourism in 2009.

Fionn Murtagh (University of Derby and University of London, UK) discusses the key role of analytics in correspondence analysis, concentrating on the use of aggregation and different resolution scales to aid interpretation of structures and relationships in massive data sets. A low resolution scale can improve computational time, while supporting high resolution mapping of the subjects and attribute categories. He demonstrates these ideas using a Twitter data set of 12 million tweets, and a small example of national funding agency data.

Mainly thanks to data available on the Web, there is the challenge of jointly analysing numerical and textual data. Simona Balbi (University of Naples Federico II, Italy), Michelangelo Misuraca (University of Calabria, Italy) and Maria Spano (University of Naples Federico II, Italy) focus their attention on tools offered by methods related to CA, when verbal expressions are considered dependent on numerical features. Two methods are compared, PCA onto a reference subspace and canonical correspondence analysis, in analysing firm performance indices and management commentaries for major Italian firms.

In his ongoing work of textual data analysis, Ludovic Lebart (Télécom Paris Tech, France) discusses some inferential issues in text mining. A corpus of texts comprises several parts, and a fragmentation of the text provides an unsupervised variant of the analysis of the global lexical table, in the present case a table of parts by words. He discusses internal and external validation procedures that allow for a critical use of the methods and provide an assessment of the results, using variants of bootstrap techniques applied to textual data.

John Gower (Open University, UK), Sugnet Gardner-Lubbe and Niël Le Roux (Stellenbosch University, South Africa) discuss basic features and problems that are linked to data analysis when describing models, distances, norms, measures of approximation and algorithms used in the practice of data analysis. The authors emphasize the importance of the actual data, including how they are collected, the types of variables and the structure of the data, for example, if they are organized in arrays or multiway tables.

Symbolic data analysis is concerned with complex forms of data such as histograms and density functions; for example, a single datum might be the distribution of sales of a product over 24 hours on a website. Rosanna Verde and Antonio Irpino (University of Campania "Luigi Vanvitelli", Italy) show how such data can be coded as blocks of variables that can be analysed using multiple factor analysis, an approach similar to simultaneous analysis discussed by Fenoglio (see summary above). In their application to distributions as symbolic objects, this approach is related to the Wasserstein distance that measures similarity between distributions in terms of their location, scale and shape.

Solène Bienaise and Brigitte Le Roux (University Paris Descartes, France) discuss a method of statistical inference for geometric data analysis (GDA) that is based on a combinatorial framework to highlight the role of permutation tests. The method is applicable to any individuals by variables data table, with structuring factors on individuals, and numerical variables possibly produced by a GDA method such as PCA or MCA. The authors demonstrate several features of their method using data from medical research on Parkinson's disease.

For comparing the 13 papers we prepared a table of word counts of all terms that are related to correspondence analysis and related methods, resulting in a 13-by-198 matrix of term frequencies. Figure 1 shows the CA biplot of this frequency table, where terms are shown according to their contribution to inertia in the two-dimensional solution, that is, the longer the length of the vector, the higher the contribution of the respective term. For a better readability of the figure, we excluded all terms that have lower than average contribution and for this reason not helpful for differentiating the papers.

The CA biplot shows three separate clusters of papers. The more CA-related papers are shown in a group of nine papers lower down, associated with terms such as CA, table, map, profile, and contingency. Then there are two pairs of papers that separate out at upper left and right.

At upper right are the papers from Durand/Le Roux and Blasius/Nenadiĉ/Thiessen, which can be described by their use of terms such as category, categorical, contribution, axis and correlation. At upper left are the papers by Bienaise/Le Roux and Verde/Irpino, which can be best differentiated by terms such as cloud, test, distribution and sample. Notice the gradient of papers on the left, from Choulakian up to Bienaise/Le Roux, with Gower/Lubbe/Le Roux closest to average position, and with the bunches of terms pointing in opposing directions characterizing the opposition along this gradient. This CA solution was also the basis for our ordering of the papers in this special issue, going from the two papers on the right and then moving upwards along the gradient on the left.

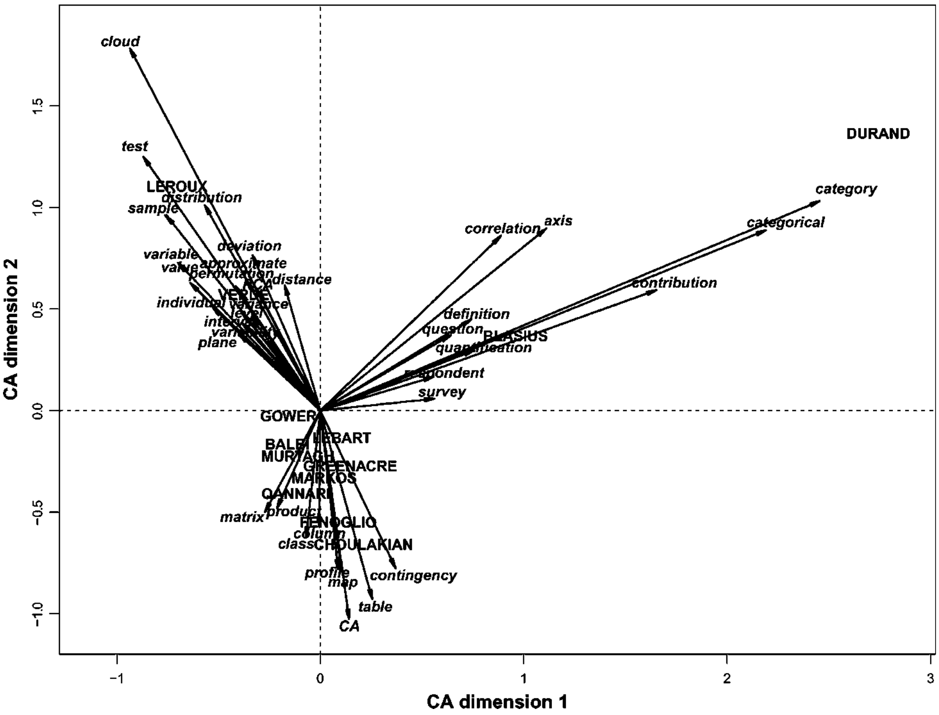

Figure 1: Correspondence analysis biplot of the words used by the authors of the 13 chapters in this special issue. Names of first authors only are shown and, out of the 198 most frequent words, only those 36 words that contribute more than average to the two dimensional solution are represented.

So, by this simple illustration, we have shown how correspondence analysis can be used to classify these articles written on the theme of correspondence analysis itself!

This special issue constitutes another milestone in the CARME project, which has already produced four edited books and another special issue of a journal before this one. We, the guest editors, are pleased to thank Luigi Fabbris, the Editor-in-Chief of the Italian Journal of Applied Statistics, for his support in publishing these articles, and the Editorial Manager, Rossella Berni, for her patient assistance.

Simona Balbi, Jörg Blasius and *Michael Greenacre*