Adv Data Anal Classif (2015) 9:1–4
DOI 10.1007/s11634-015-0202-1

EDITORIAL

Editorial for Special Issue on Symbolic Data Analysis

Preface by the guest editors

Published online: 28 February 2015

This Special Issue of the journal *ADAC* is devoted to the domain of Symbolic Data Analysis (SDA) and presents recent work at both theoretical and applied levels. SDA is concerned with representing and analyzing data presenting intrinsic variability. When the data set under analysis is not composed of single elements, but of groups gathered on the basis of some given criteria or abstract concepts, then data generally do not present one single value for each variable as in classical Statistics and Data Mining. Instead, and in order to take account of the inherent intra-group variability, new variable types are introduced whose realizations are not single real values or categories, but sets, intervals, or, more generally, distributions over a given domain.

Introduced by E. Diday in the 1980's, SDA has considerably developed, various different approaches have been investigated, and many special methods were proposed in this context. Interval data is by far the mostly considered case, three papers in this Special Issue are concerned with such data. On the other hand, the analysis of distributional data is getting increased attention and new research is emerging for this type of data, as exemplified by three other papers published in this *ADAC* Special Issue. From the large catalogue of SDA methods, the six articles in this Special Issue provide an interesting sample, including clustering methods, regression models, dimensionality reduction methods and forecasting approaches. They resulted from a list of 20 manuscript submissions.

SDA methods have traditionally proven useful in simulation contexts and well-known academic data sets, but more effort should be done in real-life applications where the potential of symbolic data can be fully exploited. This is one of the challenges of the discipline. This Special Issue provides nice examples for situations where the consideration of symbolic data proves useful in real life applications. For example:

- The use of interval data in marketing where variables such as age and time spent are usually expressed by intervals.
- High–low intervals of financial prices for representing volatility whose dynamics are analyzed by regime switching models.

- The analysis of the Science Citation Index (SCI) using the journal category to reveal the principal components in journal metrics and the behavior of journal categories.
- Modelling the distribution of ozone concentration using, as predictors, the distributions of the temperature, solar radiation and wind speed at several places of the United States.
- The use of the symbolic approach in epidemiology to describe demographic and environmental aspects of the disease and to assess the relative effectiveness of different treatment strategies.

Below we provide a short description of the accepted articles:

- The article "LASSO-constrained regression analysis for interval-valued data" by *Paolo Giordani* proposes a new approach of linear regression analysis for interval-valued data which uses LASSO constraints. Two linear regression models are considered, one for the midpoints and one for the radii of the intervals. However, and unlike previous similar approaches (namely CCRM), the vector of coefficients of the radii model, $\mathbf{b_R}$, is written as $\mathbf{b_M} + \mathbf{b_A}$, where $\mathbf{b_M}$ is the vector of midpoint coefficients and $\mathbf{b_A}$ is a vector of the additive coefficients that indicates how much the coefficients for the midpoints $\mathbf{b_M}$ differ from the corresponding ones for the radii. LASSO constraints are introduced on the magnitude of the elements of $\mathbf{b_A}$. The authors stress the flexibility of the model and the improved prediction accuracy gained by the use of the LASSO constraints on the regression coefficients. The method is evaluated and compared with CCRM on the basis of a simulation study, considering different data configurations. Three applications are presented, one concerning mushrooms descriptions and the two other ones cardiological data.
- The article by *Pierpaolo d'Urso*, *Livia de Giovanni* and *Riccardo Massari* on "Trimmed fuzzy clustering for interval-valued data" introduces a clustering model for interval data that comprises several remarkable features: first, it is a fuzzy approach that makes possible the partial membership to clusters and also provides information about the membership to other clusters. Second, the cluster prototypes are representative objects suitably selected from the dataset. Third, it comprises a trimming procedure to ignore interval outliers in the clustering process. The effectiveness of the method is demonstrated by a simulation study. Its usefulness in real-life context is illustrated by two real-life applications.
- The article "Modeling and forecasting interval time series with threshold models" written by *Paulo M. M. Rodrigues* and *Nazarii Salish* is the first work so far that considers regime switching threshold models for interval time series. Intervals are characterized by their center and radius and the resulting model is in fact a two-dimensional vector time series model. The models are applied to forecast the S&P500 index interval returns. The results suggest that nonlinear threshold-type models are able to improve forecast performance in comparison with vector autoregressive models and k-nearest neighbour methods. Interestingly, the proposed model also makes possible to map the regimes into high- and low-volatility periods of the returns.

- The paper "Principal component analysis for probabilistic symbolic data: a more generic and accurate algorithm" by *Meiling Chen, Huiwen Wang* and *Zhongfeng Qin* addresses the problem of generalizing the method of Principal Component Analysis to distribution-valued data. The principal components are represented using the convolution of the initial distributions according to the weights of the PCA. Further, the paper defines a covariance matrix for probabilistic symbolic data and presents a method based on this variance-covariance structure. The effectiveness of the proposed method is illustrated by a simulated numerical experiment and two real-life cases dealing with clustering of oils and fats data and the evaluation of indexed journals of SCI.
- In the paper "Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein distance" the authors *Antonio Irpino* and *Rosanna Verde* present a new linear regression model for distributional symbolic variables, i.e., variables whose realizations are histograms, empirical distributions or empirical estimates of parametric distributions. The model is defined on the basis of the quantile functions associated with the observed distributions. The proposed method uses the L_2 Wasserstein distance to measure the error between the observed and the predicted distributions of the dependent variable. The resulting model can be decomposed in two parts, one modelling the averages (i.e. the central tendency) of the distributions, and another modelling the variability of the centred distributions. Model parameters are obtained by solving a constrained least squares problem. Measures of goodness-of-fit are also proposed and discussed. Finally, the method is evaluated both on simulated data and on two real-world datasets.
- The paper "Strategies evaluation in environmental conditions by SDA: application in medicine and epidemiology to trachoma" written by *Christiane Guinot, Denis Malvy, Jean-François Schémann, Filipe Afonso, Raja Haddad* and *Edwin Diday* presents an application of SDA to an interventional study on trachoma conducted in Mali. Trachoma, caused by repeated ocular infections with *Chlamydia trachomatis* whose vector is a fly, is an important cause of blindness in the world. First, the authors build classes according to the evolution of the disease during the study. Then they transform the quantitative variables into qualitative ones, bar charts. The method chooses the strata bounds that give the most discriminant bar charts for the classes. After that, they apply an extension of PCA to symbolic data. The obtained results are compared to those previously obtained by multiple logistic regression analysis. SDA provides here a new perspective and suggests that some demographic, economic and environmental parameters are related to the disease and its evolution during the treatment, whatever the strategy. Also, the proposed methodology allowed assessing the relative effectiveness of the different strategies.

The papers in this Special Issue constitute a selection of recent research and applications in SDA. However, much remains to be done in this relatively recent and dynamic field of research, challenging problems lay ahead. We hope the reader finds the papers in this Special Issue intellectually stimulating and motivating for future work in this domain.

The Editors gratefully acknowledge the invaluable assistance of the experts and colleagues in the process of reviewing the manuscripts that were submitted for this Special Issue.

Paula Brito (Porto, Portugal)
Monique Noirhomme-Fraiture (Namur, Belgium)
Javier Arroyo (Madrid, Spain)