CLADAG 2023

BOOK OF ABSTRACTS AND SHORT PAPERS

14th Scientific Meeting of the Classification and Data Analysis Group
Salerno, September 11-13, 2023

edited by

Pietro Coretto
Giuseppe Giordano
Michele La Rocca
Maria Lucia Parrella
Carla Rampichini

Pearson

SCIENTIFIC PROGRAM COMMITTEE

Carla Rampichini (chair, University of Florence - Italy)

Claudio Agostinelli (University of Trento - Italy)
Michela Battauz (University of Udine - Italy)
Antonio Canale (University of Padua - Italy)
Carlo Cavicchia (Erasmus University Rotterdam - Netherlands)
Claudio Conversano (University of Cagliari - Italy)
Eustasio del Barrio (University of Valladolid - Spain)
Roberto Di Mari (University of Catania - Italy)
Stefania Fensore (University of "G. d'Annunzio" - Italy)
Nial Friel (University College Dublin - Ireland)
Maria Giovanna Ranalli (University of Perugia - Italy)
Leonardo Grilli (University of Firenze - Italy)
Luigi Grossi (University of Padua - Italy)
Christian Hennig (University of Bologna - Italy)
Mia Hubert (KU Leuven - Belgium)
Alfonso Iodice D'Enza (University of Naples "Federico II" - Italy)
Julien Jacques (University of Lyon - France)
José Joaquim Dias Curto (ISCTE-Instituto Universitário de Lisboa- Portugal)
Michele La Rocca (University of Salerno - Italy)
Silvia Montagna (University of Turin - Italy)
Barbara Pawelek (University of Cracow - Poland)
Fulvia Pennoni (University of Milano-Bicocca - Italy)
Mario Rosario Guarracino (University of Cassino - Italy)
Katrijn Van Deun (University of Tilburg - Netherlands)
Simone Vantini (Politecnico di Milano - Italy)
Donatella Vicari (Sapienza University of Rome - Italy)
Helga Wagner (Johannes Kepler University Linz - Austria)
Hiroshi Yadohisa (Doshisha University - Japan)

LOCAL PROGRAM COMMITTEE

Michele La Rocca (chair, University of Salerno - Italy)

Pietro Coretto (University of Salerno - Italy)
Giuseppe Giordano (University of Salerno - Italy)
Paolo Rocca Comite Mascambruno (University of Salerno - Italy)
Marcella Niglio (University of Salerno - Italy)
Maria Lucia Parrella (University of Salerno - Italy)
Marialuisa Restaino (University of Salerno - Italy)
Domenico Vistocco (University of Naples "Federico II" - Italy)
Maria Prosperina Vitale (University of Salerno - Italy)

CLADAG 2023 BOOK OF ABSTRACTS AND SHORT PAPERS:
14th Scientific Meeting of the Classification and Data Analysis Group, Salerno, September 11-13, 2023
edited by Carla Rampichini, Michele La Rocca, Pietro Coretto, Giuseppe Giordano, Maria Lucia Parrella

Front cover: Genome sequence map, chromosome architecture and genetic sequencing chart abstract data,
© Tartila / Shutterstock

Published by Pearson Education Resources, Italia
www.pearson.it
ISBN: 9788891935632

INDEX

Preface

This book collects the abstracts and short papers presented at CLADAG 2023, the 14th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS). The meeting has been organized by the Department of Economics and Statistics of the University of Salerno, under the auspices of the University of Salerno, the SIS and the International Federation of Classification Societies (IFCS).

CLADAG is a member of the IFCS, a federation of national, regional, and linguistically-based classification societies. It is a non-profit, non-political scientific organization, whose aims are to further classification research. Every two years, CLADAG organizes a scientific meeting, devoted to the presentation of theoretical and applied papers on classification and related methods of data analysis in the broad sense. This includes advanced methodological research in multivariate statistics, mathematical and statistical investigations, survey papers on the state of the art, real case studies, papers on numerical and algorithmic aspects, applications in special fields of interest, and the interface between classification and data science. The conference aims at encouraging the interchange of ideas in the above-mentioned fields of research, as well as the dissemination of new findings. CLADAG conferences, initiated in 1997 in Pescara (Italy), were soon considered as an attractive information exchange market and became an important meeting point for people interested in classification and data analysis. A selection of the presented papers is regularly published in (post-conference) proceedings, typically by Springer Verlag.

The Scientific Committee of CLADAG 2023 conceived the Keynote Sessions to provide a fresh perspective on the state of the art of knowledge and research in the field. The scientific program of CLADAG 2023 is particularly rich. All in all, it comprises 5 Keynote Lectures, 31 Invited Sessions promoted by the members of the Scientific Program Committee, and 27 Contributed Sessions. We thank all the session organizers for inviting renowned speakers, coming from many different countries. We are greatly indebted to the referees, for the time spent in a careful review of the abstracts and short papers collected in this book. Special thanks are finally due to the members of the Local Organizing Committee and all the people who collaborated for CLADAG 2023. Last but not least, we thank all the authors and participants, without whom the conference would not have been possible.

Pietro Coretto
Giuseppe Giordano
Michele La Rocca
Maria Lucia Parrella
Carla Rampichini

Salerno, September 2023

Plenary Session

REDUCING SELECTION BIAS IN NON-PROBABILITY SAMPLE BY SMALL AREA ESTIMATION

Francesco Schirripa Spagnolo [1], Gaia Bertarelli[2], Nicola Salvati [1], Donato Summa [3], Monica Scannapieco [3], Stefano Marchetti [1] and Monica Pratesi [13]

[1] Department of Economics and Management, University of Pisa, (e-mail: francesco.schirripa@unipi.it, nicola.salvati2@unipi.it, stefano.marchetti@unipi.it, monica.pratesi@unipi.it)

[2] Department of Economics, Ca' Foscari University of Venice (e-mail: gaia.bertarelli@unive.it)

[3] ISTAT (e-mail: summa@istat.it, scannapieco@istat.it, pratesi@istat.it)

ABSTRACT: Nowadays, the availability of a huge amount of data produced by a wide range of new technologies is increasing. However, data obtainable from these sources are often the result of a non-probability sampling process. We propose a method to reduce the selection bias associated with the big data in the context of Small Area Estimation. Our approach is based on data integration and it combines a big data sample and a probability sample. Real data examples are considered in the context of Italian enterprises sensitiveness towards Sustainable Development Goals and e-commerce.

KEYWORDS: official statistics, big data, data integration, SDGs, e-commerce.

1 Introduction

For many decades probability surveys have been the standard for producing official statistics. However, the decline in response rates in probability surveys associated with the increasing cost of data collection have become big issues for producing official statistics. Due to technological innovations, over the past decade, there has been an unprecedented increase in the volume of "new" data, called *big data*, which are often the results of non-probability sampling processes but, at the same time, they offer very rich data sets. Anyway the "nature" itself of the data, as collected without a probability scheme, opens the door to possible selection bias, even at domain level.
Although, there is a trend to modernize official statistics through a more extensive use of big data, making reliable inferences from a non-probability sample alone is very challenging and a naive use of these data can lead to

biased estimates as affected by selection bias and measurement error. The Italian National Statistical Institute has a strategic program of investments on the use of these new data sources to complement and enrich official statistics. In this context a roadmap document, named "Roadmap for Trusted Smart Statistics"(RTSS), has been released. This work must be laid in the methodological action of the RTSS related to quality improvement by reducing non-representativeness of Big Data sources at survey unplanned domain level.

2 Notation

We consider a population U of size N divided into m non-overlapping subsets U_i of size N_i, $i = 1,\ldots,m$. Let y_{ij} denote the value of the target variable for the unit j belonging to the area i. A non-probability sample B is available for the target population, with $B \subset U$. We assume that the non-probability sample is available in each area of interest: B_i is the non-probability sample in the area i, $B_i \subset U_i$. We denote the inclusion indicator in B_i as δ_{ij}; in other words, $\delta_{ij} = 1$ if $j \in B_i$, $\delta_{ij} = 0$ otherwise; therefore $N_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$. The study variable y_{ij} is observed only when $\delta_{ij} = 1$. The non-probability sample contains other auxiliary variables, denoted by $\mathbf{x}$.

A survey data of size n, denoted by A, is also available; $A_i \in U_i$ drawn randomly. The survey data do not contain the variable of interest but contain only the auxiliary variables $\mathbf{x}$. The area-specific samples A_i are available in each area, but the number of sample units in each area, $n_i > 0$, is limited. Therefore, the areas of interest can be denoted as "small areas". In general, an area is regarded as "small" if the domain-specific sample size is not large enough to obtain direct estimates with acceptable statistical significance. In these cases, SAE techniques need to be employed.

In summary, the available data can be denoted by $\{(y_{ij}, x_{ij}), i \in B\}$ and $\{(x_{ij}), i \in A\}$, and the quantities of interest are the area means $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_{ij}, i = 1,\ldots,m$. By using B we can estimate $\bar{Y}_i$ by:

$$\bar{Y}_{B_i} = N_{B_i}^{-1} \sum_{j \in B_i} y_{ij},$$

where $N_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$ and y_{ij} is the jth observation in the area i. Because of the selection bias and the measurement error, the sample mean $\bar{Y}_{B_i}$ from the non-probability sample is biased, and it does not represent the target population (Kim & Wang, 2019). Therefore, we propose a techniques in order to make valid inference from big data sources when the aim is to provide reliable estimates at small area level.

3 Reducing selection bias in big data: a data integration approach using SAE methods

We consider a data integration method for combining probability and non-probability samples in order to reduce the bias which is assisted by unit level small area model, following the approach of Kim and Wan (2019). We consider the case in which the survey data and the big data are available in each small area of interest. We also assume that the selection mechanism for the big data is non-informative :

$$P(\delta_{ij} = 1|\mathbf{x}_{ij}, y_{ij}; u_i) = P(\delta_{ij} = 1|\mathbf{x}_{ij}; u_i)$$

where u_i is an area-specific random effect characterizing the between-area differences in the distribution of y_{ij} given the covariates $\mathbf{x}_{ij}$.

Moreover, we can observe δ_{ij}, the big data sample inclusion indicator, from the sample A. We can use the data $\{(\delta_{ij}, \mathbf{x}_{ij})\} \in A_i$ to fit a model for the for the propensity scores $P(\delta_{ij} = 1|\mathbf{x}_{ij}) = p(\mathbf{x}, \lambda)$ in sample B based on the missing at random. Usually, a logistic regression model for the binary variable δ_{ij} can be used in order to obtain estimators $\hat{p}_{ij}$ in sample B.

In order to take into account the hierarchical structure of the data, we consider the following generalized linear random intercept model for the propensity scores:

$$\hat{p}_{ij}(\hat{\lambda}, \hat{u}_i) = g^{-1}(\mathbf{x}_{ij}^T \hat{\lambda} + \hat{u}_i),$$

where $g(\cdot)$ is a logit link function; $\hat{\lambda}$ and $\hat{u}_i$ are the ML estimates of λ and u_i.

To develop our estimator we suppose that the following working population model holds for sample B:

$$E[y_{ij}|\mathbf{x}_{ij}, \gamma_i] = \mu_{ij} = h^{-1}\left(\mathbf{x}_{ij}^T \beta + \gamma_i\right), \tag{1}$$

where $h(\cdot)$ is the link function, assumed to be known and invertible, γ_i is the area-specific random effect for area i characterizing the between-area differences in the distribution of y_{ij} given the covariates $\mathbf{x}_{ij}$. It should be noted that the covariates used here could be different from those used to fit the propensity model. Model in equation (1) includes three important special cases: the linear model obtained with $h(\cdot)$ equal to the identity function and y_{ij} is a continuous variable; logistic generalized linear random intercept model, where $h(\cdot)$ is the logistic link function and the outcome variable is binomial; the Poisson-log generalized linear random intercept model where $h(\cdot)$ is the log link function

and the individual y_{ij} values are taken to be independent Poisson random variable. Using data from the big data sample B, assuming the model is correctly specified, we obtain an estimator of $\hat{\beta}$ which is consistent for β (Rao, 2021). Then a doubly robust (DR) estimator of the mean is given by:

$$\hat{\theta}_{i;DR}^{EBLUP} = \frac{1}{N_i}\left\{\sum_{j\in B_i}\frac{1}{\hat{p}_{ij}(\hat{\lambda},\hat{u}_i)}(y_{ij}-\hat{\mu}_{ij}) + \frac{N_i}{n_i}\sum_{j\in A_i}\hat{\mu}_{ij}\right\}, \tag{2}$$

where $\hat{\mu}_{ij} = h^{-1}\left(\mathbf{x}_{ij}\hat{\beta}+\hat{\gamma}_i\right)$ and $\hat{\beta}$ and $\hat{\gamma}_i$ are respectively the estimated regression coefficients and the random effects based on the big data sample.

The estimator in Eq. (2) is DR in the sense that it is consistent if both the model for propensity scores and the model for the study variable are correctly specified (Kim & Wang, 2019, Rao, 2021).

4 Real data examples

The proposed methodology has been applied to estimate the proportion of enterprises sensitive to Sustainable Development Goals (SDGs) of the 2030 Agenda at the provincial level in Italy. The Big Data sample is represented by the enterprises' websites accessed due to a web scraping procedure. The probabilistic sample dataset is a sub-sample of the survey "*Situazione e prospettive delle imprese nell'emergenza sanitaria Covid-19"* (2020). The target variable is a binary indicator computed for each enterprise and represents if the enterprise is sensitive or not to SDGs. This indicator has been computed through machine learning methods by analyzing the big data sample and looking for a set of pre-defined SDGs-related words on each website. Furthermore, an application related to the diffusion of e-commerce in Italian companies, using the same data of the application on sustainability, will be considered.

References

KIM, JAE KWANG, & WANG, ZHONGLEI. 2019. Sampling techniques for big data analysis. *International Statistical Review*, **87**, S177–S191.

RAO, JNK. 2021. On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, **83**, 242–272.

SPARSE AND ROBUST ESTIMATORS FOR OUTLIER DETECTION IN DISTRIBUTIONAL DATA

Pedro Duarte Silva [1], Peter Filzmoser [2] and Paula Brito [3]

[1] Católica Porto Business School & CEGE, Universidade Católica Portuguesa, Porto, Portugal, (e-mail: psilva@ucp.pt)

[2] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria, (e-mail: peter.filzmoser@tuwien.ac.at)

[3] Faculdade de Economia, Universidade do Porto & LIAAD-INESC TEC, Porto, Portugal, (e-mail: mpbrito@fep.up.pt)

ABSTRACT: The classical data representation model is too restrictive when the data to be analysed are not real numbers but comprise variability. In this talk, we are interested in numerical distributional data, where units are described by histogram or interval-valued variables. We consider parametric probabilistic models, which are based on the representation of each distribution by a location measure and interquantile ranges. A multivariate outlier detection method is proposed that makes use of restricted configurations for the covariance matrix, and is based on a sparse robust estimator of its inverse. The computations rely on an efficient adaptation of the graphical lasso algorithm. A simulation study puts in evidence the usefulness of the robust estimates for outlier detection.

KEYWORDS: outliers, robust statistics, distributional data, Mahalanobis distance, graphical lasso

1 Introduction

Multivariate datasets often include atypical data points known as *outliers*, i.e. points that deviate from the main pattern. Outlier detection is important because outlying data points may reveal nonconforming phenomena and the results of usual multivariate methods can be heavily influenced by them.

In this paper we address the problem of outlier detection in multivariate distributional data. Distributional data may result from the aggregation of large amounts of open/collected/generated data, or may be directly available in a structured or unstructured form, describing the variability of some features. In recent years, different approaches have been investigated and methods proposed for the analysis of such data. However, most existing methods rely on non-parametric descriptive approaches.

A common approach for multivariate outlier detection measures outlyingness by Mahalanobis distances. Given a sample of n observations, a point i is considered an outlier if its distance $D^2_{\hat{\mu},\hat{\Theta}}(i)$ from an appropriate mean estimate, $\hat{\mu}$, is above a relevant threshold. Here, $\hat{\Theta}$ is an estimate of the precision matrix, $\Theta = \Sigma^{-1}$, and Σ denotes the population covariance. However, if $\hat{\mu}$ and $\hat{\Theta}$ are chosen to be the classical sample mean vector and inverse covariance matrix, S^{-1}, this procedure is not reliable, as $D^2_{\hat{\mu},\hat{\Theta}}(i)$ may be strongly affected by atypical observations. Furthermore, S^{-1} has a large sample variability when its dimension, d, is close to n, and it is is not even computable when $d > n$. To address these issues Öllerer and Croux (Öllerer & Croux, 2015), proposed sparse precision matrix estimators based on the GLASSO L_1-penalized log-likelihood function (Friedman *et al.* , 2008).

In this paper we address the problem of outlier detection in distributional data, combining Öllerer and Croux estimators with a parametric modelling of distributional data, along the lines of Brito & Duarte Silva, 2012, and Duarte Silva *et al.* , 2018.

2 Distributional Variables

Let $S = \{s_1, \ldots, s_n\}$, be the set of n units under analysis. We consider that for each unit, the descriptive variables are (in general) not constant, but present variability.

We represent the "values" of a numerical distributional variable by an ordered vector of quantiles, always including the minimum and the maximum. Formally, a numerical distributional variable is defined by an application

$$\begin{gathered} Y : S \to T \\ s_i \to Y(s_i) = (Min_i, \psi_{1i}, \ldots, \psi_{qi}, Max_i) \end{gathered}$$

Let $Y_1, \ldots, Y_p$ be the p numerical distributional variables, defined on S. Here we assume that all variables are represented by the same set of $q+2$ quantiles, and that $Min_{ij} < \psi_{1ij} < \ldots < \psi_{qij} < Max_{ij}, 1 \leq i \leq n, 1 \leq j \leq p$ (strict inequalities).

The model consists in representing $Y_j(s_i)$ by

- a central statistic C_{ij}, typically the Median Med_{ij} or the MidPoint $\frac{Max_{ij}+Min_{ij}}{2}$
- the $[Min, \psi_1[$ range: $R_{1ij} = \psi_{1ij} - Min_{ij}$
- the $[\psi_1, \psi_2[$ range: $R_{2ij} = \psi_{2ij} - \psi_{1ij}$
- ...

- the $[\psi_q, Max[$ range: $R_{mij} = Max_{ij} - \psi_{qij}$

Typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles.

The proposed model consists in assuming that the joint distribution of the central statistic C and the logarithms of the ranges $R^*_\ell, \ell = 1, \ldots, m$, is Gaussian:

$(C, R^*_1, \ldots, R^*_m) \sim N_{(m+1)p}(\mu, \Sigma)$

In the most general formulation (configuration 1) we allow for non-zero correlations among all central statistics and log-ranges; for distributional variables there are however other cases of interest: the distributional-valued variables Y_j are non-correlated, but for each variable, the central statistic and all its log-ranges may be correlated among themselves (configuration 2); central statistics (respectively, log-ranges) of different variables may be correlated, but no correlation between central statistics and log-ranges is allowed (configuration 3); central statistics (respectively, each log-range) of different variables may be correlated, but no correlation between central statistics and log-ranges or between non-corresponding log-ranges is allowed (configuration 4); and, finally, all central statistics and log-ranges are non-correlated (configuration 5).

3 Outlier Detection of Distributional Data

Let $X_i = \left[C_i^t, R_{1i}^{*\,t}, \ldots, R_{mi}^{*\,t}\right]^t$ be the $d = (m+1)p$ dimensional column vector comprising all central statistics and log-ranges for s_i, $i = 1, \ldots, n$.

The identification of outliers is based on robust Mahalanobis distances, $D^2_{\hat{\mu},\hat{\Theta}}(i) = (x_i - \hat{\mu})^t \hat{\Theta} (x_i - \hat{\mu})$ from each data point to a robust location vector, $\hat{\mu}$, which are then compared with the 97.5% quantile of a chi-squared distribution with d-degrees of freedom. In our approach we choose as location vector, the L_1 median (Fritz *et al.* , 2012), which has a break-down point of 0.5 and, given our Gaussian assumption, is a robust estimator of μ.

Following Öllerer and Croux (2015) we estimate $\Theta = \Sigma^{-1}$ by

$$\hat{\Theta} = argmax_{\Theta \in \vartheta} \log\det(\Theta) - tr(\hat{\Sigma}\Theta) - \rho \sum_{j,k=1}^{d} |(\Theta)_{jk}| \qquad (1)$$

where $\vartheta := \{\Theta \in \mathbb{R}^{d \times d} : \Theta \succ 0\}$ is the space of d-dimensional positive-definite matrices, $\hat{\Sigma}$ is a robust covariance estimate, and ρ a regularization parameter.

For each covariance configuration, we set the null elements of Σ to zero in its initial $\hat{\Sigma}$ estimate, and for the remaining elements we use the formula