Springer Proceedings in Mathematics & Statistics

Alessandra Petrucci
Filomena Racioppi
Rosanna Verde *Editors*

New Statistical Developments in Data Science

SIS 2017, Florence, Italy, June 28–30

SIS Società Italiana di Statistica

Springer

Springer Proceedings in Mathematics & Statistics

Volume 288

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Alessandra Petrucci · Filomena Racioppi ·
Rosanna Verde
Editors

New Statistical Developments in Data Science

SIS 2017, Florence, Italy, June 28–30

Springer

Editors
Alessandra Petrucci
Dipartimento di Statistica, Informatica,
Applicazioni 'G. Parenti' (DiSIA)
Università degli Studi di Firenze
Florence, Italy

Filomena Racioppi
Dipartimento Scienze Statistiche
Sapienza Università di Roma
Rome, Italy

Rosanna Verde
Dipartimento di Matematica e Fisica
Università della Campania 'Luigi Vanvitelli'
Caserta, Italy

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-21157-8 ISBN 978-3-030-21158-5 (eBook)
https://doi.org/10.1007/978-3-030-21158-5

Mathematics Subject Classification (2010): 60-XX, 60Gxx, 97K50, 37A50, 62-XX

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface: Statistics and Data Science Today

Nowadays, it seems almost impossible to find a shared definition of Data Science. Some believe that Data Science is still a buzzword, and that it is not really a new domain of science or a new scientific discipline. The truth of such a claim is only partial, however, since Data Science is developing rapidly in response to the need to gain knowledge from huge amounts of data of various types and deriving from different sources. Data are often updating in real time; they are usually complex and can be structured or unstructured—here, we are referring to social network data, textual data, micro array, sensor data stream, and so on.

The debate has centered on the framework of reference for data management and processing and has involved specialists and scientists in various ways. Statistics, rather than being considered a *science*, has been defined as a *discipline* since it does not have a single field of application but rather contributes across fields, as and when required. Statistics has its own paradigms, namely *synthesis of information* and *research into the new*. Beyond the capture of data and their management, storage, cleaning, and processing, Statistics provides descriptive tools to represent, in a synthetic way, phenomena, causal relationships among variables, models for the analysis of temporal data in order to monitor evolving phenomena, dimensionality reduction techniques to synthesize data, forecasting models, and so on.

Since the end of the 1970s, when the spread of computing systems began, the application of *Data Analysis* within Statistics has resulted in a new field of research, *Computational Statistics*, which in turn has laid the foundations for Data Analytics. Possibly, John W. Tuckey predicted this development as long ago as 1962, when he wrote "The Future of Data Analysis".[1]

It is undeniable that Statistics covers many of the skills essential in Data Science —hence the important role played by Statistics, as well as other disciplines within the field of Computer Science, such as Machine Learning and Artificial Intelligence, that address data processing and learning from data.

[1] John W. Tuckey "The Future of Data Analysis" Ann. Math. Statist., 33, 1 (1962), 1–67.

Data Science aims to transform data into useful knowledge that enables prediction and supports and validates decisions. Statistics may be regarded as one of the two "parents" of Data Science, representing its logic, while the other, Computer Science, is its language.

In this great revolution brought about by data production and dissemination, the boundaries of the skills in different fields have become ill-defined. The skills required by Data Science are strongly interdisciplinary: they bring together Computer Science, Statistics, and Machine Learning, with the purpose of giving meaning to the data. Moreover, within this framework the expertise domain constitutes the connector which allows the transformation of data into knowledge.

The tradition of Statistics, through its different areas, offers a potential role to Data Science in exploiting and interpreting data results and improving links between Statistics and Information Technology. Statistical thinking is fundamental for the provision of detailed descriptions of data structures and interpretation of complex phenomena, for explanation of causal relations, and for forecasting trends and monitoring data evolution.

There is a common belief that Data Science is merely a fashionable topic among scholars, rather than a scientific mentality or way of scientific thinking. But its potential is high and its "toolbox" is broad, with specific methods and techniques related to the specific application areas where data processing is required. So, it seems more appropriate to refer to several *Data Sciences*, one for each domain (Financial Data Science, Economics Data Science, Environmental Data Science, Social Data Science, etc.) where expert knowledge plays the main role.

This reference to the domains of data surely offers an interpretative key to Data Science(s). The contribution of Statistics is not compromised by the multidomain nature of Data Science: our discipline addresses *collective phenomena*, and these collective phenomena are specific within the different knowledge domains and fields of application.

In recent years, especially through the Italian Statistical Society (SIS), our scientific community has been asked to define its role and involvement in Data Science. This led to the constitution of an Italian SIS Group for Statistics and Data Science and to the SIS Conference on "Statistics and Data Science: new challenges, new generations," held in Florence on June 28–30, 2017.[2] This volume contains a selection of papers that are extended versions of some of the contributions presented at this conference. They provide some examples of recent developments in statistical methods that are of relevance for the challenges posed by Data Science.

The volume is organized into six parts, which to some extent reflect the transition from Data Science to Data Sciences referred to above. These parts cover topics including strong statistical methodologies, Bayesian approaches, applications in population and social studies, studies in economics and finance, and techniques

[2] At the conference, a round table entitled "Let's talk about Data Science," chaired by Carlo Lauro with four panelists and several planned interventions from the floor, covered many considerations referred to in this Preface.

of the sample design. The volume concludes with several contributions in mathematical statistics.

The authors approach all of these topics strictly in terms of Data Science. For example, in the first part, "Complex data analytics," the eight contributions underline the development of new techniques for dimensional reduction of big and/or high-frequency data. As is widely recognized, the high data dimension and the complexity of the new data increasingly require a suitable setting of traditional techniques of data analysis for classical *individuals* × *variables* matrices.

The second part, "Knowledge based methods," containing five contributions, explores Bayesian approaches for the estimation of causal effects, mixture models, and count models able to analyze high-dimensional data or structured data, such as the ERG models for networks affected by missing data. The third part, "Sampling design for Big Data exploration," also comprising five contributions, highlights novelties in sampling schemes and modeling of big data, empirical approaches for small area predictors, large survey strategies, and data quality monitoring.

The fourth and fifth parts, "Data Science methods for social and population studies" and "Applying Data Science in economics and labour market," include seven and five contributions, respectively. The focus here is, for example, on applications in the broad field of socio-demographic studies, including tweets analysis in the context of teacher evaluation; applications in population studies, including with respect to the behavior of young people; and applications in economic studies, such as studies offering insights into the labor market, e.g., on wage distribution inequality, or involving the development of a scoring system for lending platforms. The sixth part, "Mathematical statistics for Data Science," comprises four contributions that stress mathematical approaches in Data Science. Specific topics to be addressed are the application of a hidden Markov model for the suitable transformation of time series, representation methods for extreme value distributions, an algorithm for clustered equations, and dimensional reduction of arrays of a factorial design in an experimental framework.

This volume is addressed primarily at scholars and researchers interested in new frontiers in Statistics and Data Analysis, but it also provides useful supportive supplementary material for students in the disciplines covered.

We trust that the volume will be appreciated by both statisticians and data scientists and that they will recognize each other's fundamental role.

Florence, Italy — Alessandra Petrucci
Rome, Italy — Filomena Racioppi
Caserta, Italy — Rosanna Verde
March 2019

Contents

Mathematical Statistics for Data Science

About the Editors

Alessandra Petrucci is Professor of Social Statistics at the University of Florence, where she is Head of the Department of Statistics, Computer Science, and Applications and a member of the Academic Senate. Dr. Petrucci holds a Ph.D. in Applied Statistics. Her research interests include methods for survey sampling, spatial statistics, environmental statistics, multivariate analysis, and teaching evaluation. She has published a considerable number of papers in national and international journals and has been a member of many research projects and scientific committees for several national and international conferences. Dr. Petrucci is an elected member of International Statistical Institute (ISI) and a fellow of the Italian Statistical Society (SIS). She is a member of the Centro Camilo Dagum, Tuscan Interuniversity Centre—Advanced Statistics for Equitable and Sustainable Development.

Filomena Racioppi is Associate Professor of Demography in the Department of Statistics at Sapienza University of Rome, where she is on the Board of the Ph.D. program in Statistical Sciences. She is a member of the Italian Association for Population Studies, the Italian Statistical Society (SIS), the European Association for Population Studies (EAPS), and the International Union for Scientific Studies of Population (IUSSP). She has been a member and/or coordinator of several national and international research projects. She was executive editor of Genus from 1999 to 2009 and a member of the scientific advisory board for Population Research and Policy Review. Currently, she is on the scientific board of Genus and serves as a reviewer for several demographic international journals. Dr. Racioppi's research areas have included, among others, reproductive behaviors, applied and business demography, active aging and age management, and gender issues during the life course.

Rosanna Verde is Professor of Statistics in the Department of Mathematics and Physics at the University of Campania Luigi Vanvitelli, where she is coordinator of the Bachelor's degree in Data Analytics. She is also on the Board of the Ph.D. program in Social and Statistical Sciences at the University of Naples. Dr. Verde is

a member of several international statistical associations and is on the scientific board of SIS Group on Statistics and Data Science. She is associate editor of the journal Statistical Methods and Applications. Dr. Verde was an expert evaluator for the European Programme IST. She has coordinated various national and regional projects and has also participated in European projects. She has been a guest researcher and visiting professor at several foreign universities and research centres. Her main fields of research are classification, symbolic data analysis, data stream analysis, and functional data analysis. She is a co-author of more than 100 papers in journals and in proceedings of international conferences.

Complex Data Analytics

Monitoring the Spatial Correlation Among Functional Data Streams Through Moran's Index

Antonio Balzanella, Elvira Romano, Rosanna Verde, Francesca Fortuna, Fabrizio Maturo, Stefano Antonio Gattone and Tonio Di Battista

Abstract This paper focuses on measuring the spatial correlation among functional data streams recorded by sensor networks. In many real world applications, spatially located sensors are used for performing at a very high frequency, repeated measurements of some variable. Due to the spatial correlation, sensed data are more likely to be similar when measured at nearby locations rather than in distant places. In order to monitor such correlation over time and to deal with huge amount of data, we propose a strategy based on computing the well known Moran's index and Geary's index on summaries of the data.

Keywords Data stream mining · Functional data analysis · Spatial correlation

1 Introduction

Functional Data Analysis (FDA) has become a topic of interest in Statistics due to the increasing ability to measure and record over a continuous domain results of natural phenomena [9]. In environmental sciences, monitoring a physical phenomenon in different places of a geographic area is becoming very common due to the availability of sensor networks which can perform, at a very high frequency,

This research was funded by PRIN (year 2015, ERC Sector PE1, Prot. $20157PRZC4-004$).

A. Balzanella (✉) · E. Romano · R. Verde
University of Campania Luigi Vanvitelli, Caserta, Italy
e-mail: antonio.balzanella@unicampania.it

E. Romano
e-mail: elvira.romano@unicampania.it

R. Verde
e-mail: rosanna.verde@unicampania.it

F. Fortuna · F. Maturo · S. A. Gattone · T. Di Battista
University G. d'Annunzio of Chieti-Pescara, Chieti, Italy
e-mail: gattone@unich.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_1

repeated measurements of some variable. We can think, for instance, at temperature monitoring, seismic activity monitoring, pollution monitoring, over the locations of a geographic space. In this context one works with data having complex characteristics including spatial dependence structures. Often, the data acquisition is performed by sensors having limited storage and processing resources. Moreover, the communication among sensors is constrained by their physical distribution or by limited bandwidths. Finally, the recorded data relate, often, to highly evolving phenomena for which it is necessary to use algorithms that adapt the knowledge with the arrival of new observations. The data stream mining framework offers a wide range of specific tools for dealing with these potentially infinite and on-line arriving data. An overview of recent contributions is available in [2].

An emerging challenge, in this context, is the monitoring of the spatial dependence among sensor data. The First Law of Geography, also frequently known as Tobler Law [8], states that "everything is related to everything else, but near things are more related than distant things". This law finds its major developments in Geostatistics but is still valid in the framework of data stream mining, when the data is collected by spatially located sensors. For instance, surface air temperatures streams, are more likely to be similar when measured at nearby locations rather than in distant places.

Measuring the spatial dependence among fast and potentially infinite data streams is a very challenging task. This is due to a set of stringent constraints: (i) the available time for processing the incoming observations is small and constant; (ii) the allowed memory resources are orders of magnitude smaller than the total size of input data; (iii) only one scan of the data is feasible; (iv) the communication between the sensors should be very limited.

This paper introduces a new strategy for monitoring the spatial dependence over time which adapts the classic Moran's index [7] and Geary's index [4] to the challenge of functional data stream processing.

We assume that sensors do not communicate with each other but only with a central node. Thus, a first part of the processing is performed at the sensors while a second part is performed at the central computation node using the output of the sensors. In particular, each data stream recorded by a sensor is processed individually through two summarization steps. The first one, splits the incoming data stream into non overlapping windows and provides a compact representation of the observation in each window. The second step, performs on each data stream a CluStream [1] algorithm adapted for working on functional data subsequence. CluStream groups the incoming data into homogeneous micro-clusters and represent these through prototypes.

With the flowing of data, each sensor performs two kinds of data transmission to the central computation node. The first one is a snapshot of the micro-cluster centroid at predefined time stamps. The second one, which is performed at each windows, consists in sending the identifier of the micro-cluster to which the subsequences have been allocated. In this way, the communication between the sensors and the central node requires a low bandwidth as well as low memory resources. Only few micro-cluster prototypes are stored for each data stream at the central node and the sensor data are replaced by the micro-cluster centroid to which they have been allocated by the CluStream.

The central processing node is, thus, used for measuring the spatial dependence among the streams through the Moran's and the Geary's index on the micro-cluster centroids.

The next sections provide the details of the processing setup.

2 Sensor Data Summarization Through On-Line Clustering

Let $Y = \{Y_1(t), \ldots, Y_i(t), \ldots, Y_n(t)\}$ be a set of n functional data streams. $Y_i(t)$, $t \in T$, denotes a function defined on an interval with $T \subseteq \Re$. Each functional data stream $Y_i(t)$ is made by observations recorded by a sensor located at $s_i \in S$, with $S \subset \Re^2$ be the geographic space.

We assume that the potentially infinite data is recorded on-line so that we can keep into memory only subsets of the streams. Thus, the analysis is performed using the observations in the most recent batch and some synopsis of the old data, no longer available.

In reality, we observe the data at a grid of N points, $t_1, \ldots, t_N$. The functional data analysis viewpoint may be described by the following non-parametric model:

$$Y_{ij} = Y_i(t_j) + \varepsilon_{ij} \tag{1}$$

where $Y_i(t)$ is the underlying signal curve, ε_{ij} is an observation noise with mean zero and null covariance and Y_{ij} denote the observed noisy data, $i = 1, \ldots n$ and $j = 1, \ldots, N$.

We split the incoming data streams into non overlapping windows so that each window is an ordered subset of T, having size b which frames, for each $Y_i(t)$, a data batch $Y_i^w(t) = \{Y_i(t)\}_{t=0}^b$ (where $w = 1, \ldots, \infty$ allow to index each window).

The CluStream [1] algorithm, suitably adapted for working with the functional subsequences $Y_i^w(t)$ of the data stream $Y_i(t)$ is used for providing a fast to compute summarization of the stream. This allows to collect information about past data which are discarded after the processing.

The intuition that underlies the method, is to represent the incoming data through the center of low variability (micro-clusters). In order to have a high representativity of the input data, the number of clusters to keep updated is not specified *apriori* but only a threshold on their maximum number is fixed, to manage the memory resources.

Once the sensors have reached their storage limit, each stream of data is summarized by a set of microclusters $\left\{\mu C^1, \ldots, \mu C^K\right\}$. The generic micro-cluster μC^k records the following information:

- $\overline{Y}^k(t), t \in w$: the cluster centroid;
- n^k: number of allocated functions;
- $\sigma^k(t), t \in w$: the within micro-cluster standard deviation;

- Sw^k: Sum of window indexes;
- SSw^k: Sum of squared window indexes.

Whenever a new window w of data is available, CluStream allocates the subsequence $Y^w(t)$ to an existing micro-cluster or generates a new one. The first preference is to assign the data stream to a currently existing micro-cluster.

With the squared L^2 distance, the dissimilarity metric between two functions is defined as

$$d^2(Y_i, Y_r) = \int_0^T [Y_i(t) - Y_r(t)]^2 dt. \tag{2}$$

The algorithm is based on the following rule:

$Y_i^w(t)$ is allocated to the micro-cluster μC^k if

$$d^2\left[Y_i^w(t), \overline{Y}^k(t)\right] < d^2\left[Y_i^w(t), \overline{Y}^{k'}(t)\right] \tag{3}$$

and

$$d^2\left[Y_i^w(t), \overline{Y}^k(t)\right] < u \frac{\sum_{j \in \mu C^k} \int_{t \in w} \left[Y_j^w(t) - \overline{Y}^k(t)\right]^2 dt}{n^k} \tag{4}$$

with $k \neq k'$ and $k = 1, \ldots, K$.

The threshold value u allows to control if $Y_i^w(t)$ falls within the maximum boundary of the micro-cluster, which is defined as a factor of within cluster variance of μC_i^k. In order to take into account the functional nature of the data, a pre-smoothing step may be applied before clustering [3, 6].

The allocation of a subsequence to a micro-cluster involves the update of all its information: the micro-cluster size, centroid and standard deviation. Furthermore, it is necessary to update the sum and the sum of squares of the time window w.

If $Y_i^w(t)$ is outside the maximum boundary of any micro-cluster because of the evolution of the data stream, the a new micro-cluster, say μC^l is initialized by setting its centroid equal to $Y_i^w(t)$ and the micro-cluster size to $n_i^l = 1$. The functional standard deviation $\sigma_i^l(t)$ is defined in a heuristic way by setting it to the pointwise squared Euclidean distance to the closest cluster.

With the creation of a new micro-cluster, it is necessary to evaluate if the number of micro-clusters for the stream $Y_i(t)$ is higher than the available memory resources. In such a case, one of the old micro-clusters has to be removed in order to release memory space. This can be achieved by either deleting an old micro-cluster or joining two of the old clusters. The choice between these two alternatives, involves to evaluate if some current micro-cluster collects information about data behaviours no longer active in the recent history of the stream. Only in this case, there is a deletion. To verify this, it is possible to look at the value stored in the fields Sw^k and SSw^k of each micro-cluster and compute the average $\overline{w}^k$ and the variance $\sigma^2_{w^k}$ of the allocation

times. Behaviours no longer active will be summarized by micro-clusters having low values of $\overline{w}^k$ and $\sigma^2_{w^k}$, since they have not been updated recently.

If there is no old micro-cluster to delete, there is the merging of two nearest micro-clusters into one.

The proposed procedure, performed in a parallel way on all the streams, permits to keep, at each time instant, a snapshot of the data behavior. This is due to the availability of the set of subsequences used as representatives.

3 Monitoring Spatial Dependence on Data Stream Summaries

In this section we introduce the approach we propose for monitoring the spatial dependence among data streams summarized by functional micro-clusters.

Our idea is to measure the spatial dependence on the data of temporally consecutive time windows rather than providing a measure at each time instant. For instance, we aim at providing a measure of spatial dependence at each hour of data rather than at each second.

Since the measurement of the spatial dependence is performed at the central computation node where only the micro-clusters summarizing each data stream are available, we propose to adapt the classic Moran's I [7] and Geary's C [4] spatial autocorrelation measures, to functional data summarized by micro-cluster centroids.

We recall that the Moran's index is a widely used measure for testing the global spatial autocorrelation in spatial data. It is based on cross-products of the deviations from the mean and is calculated for the n observations of a variable Y at locations i, j, as:

$$I = \frac{n}{\sum_i \sum_j a_{i,j}} \frac{\sum_i \sum_j a_{i,j}(Y_i - \overline{Y})(Y_j - \overline{Y})}{\sum_i (Y_i - \overline{Y})^2} \tag{5}$$

where the weights $a_{i,j}$ define the relationships between locations in the geographic area.

Morans index is similar, but not equivalent, to a correlation coefficient. It varies from -1 to $+1$. In the absence of autocorrelation and regardless of the specified weight matrix, the expectation of Morans I statistic is $-1/(n-1)$, which tends to zero as the sample size increases.

According to the processing setup introduced above, at the central computation node it is kept a snapshot of micro-cluster centroids of each stream. Every time a new window becomes available, it is possible to measure the spatial autocorrelation by receiving at the central node, from each data stream, the identifier of the micro-cluster to which the subsequence of the window has been allocated. This approach, allows to measure the spatial dependence of the data in a window by using the micro-cluster centroids rather than the raw sensor data.

The functional Moran's index in each window w can be computed as follows:

$$I_{\mu C}^{w} = \frac{n}{\sum_i \sum_j a_{i,j}} \frac{\sum_i \sum_j a_{i,j} \int_{t \in w} [\overline{Y}^{k_i}(t) - \overline{Y}(t)][\overline{Y}^{k_j}(t) - \overline{Y}(t)]dt}{\sum_i \int_{t \in w} [\overline{Y}^{k_i}(t) - \overline{Y}(t)]^2 dt} \tag{6}$$

where $\overline{Y}^{k_i}(t)$ and $\overline{Y}^{k_j}(t)$ are the micro-cluster centroids to which, respectively, the subsequences $Y_i^w(t), Y_j^w(t)$, have been allocated and $\overline{Y}(t)$ is the average subsequence.

The proposed Moran's index can be used for obtaining a different measure of spatial dependence at every time window w, starting from the micro-cluster identifiers sent by the sensors to the central communication node.

Similarly, we can compute the Geary's C autocorrelation measure at each window.

We recall that the Geary's C measure is based on comparing measurements on spatial adjacent units and is calculated, for the n observations of a variable Y at the locations i, j, as:

$$C = \frac{n}{2\sum_i \sum_j a_{i,j}} \frac{\sum_i \sum_j a_{i,j}(Y_i - Y_j)^2}{\sum_i (Y_i - \overline{Y})^2} \tag{7}$$

where the weights $a_{i,j}$ define the relationships between locations in the geographic area as before.

It ranges between 1 and 2 however, values can be greater than 2 on occasion [5]. Positive spatial autocorrelation is found with values ranging from 0 to 1 and negative spatial autocorrelation is found between 1 and 2. Geary's C is inversely related to Moran's I, but it is not identical. Moran's I is a measure of global spatial autocorrelation, while Geary's C is more sensitive to local spatial autocorrelation.

As before for the Moran's index, we can measure the spatial dependence among data streams through the Geary's C measure on the micro-cluster centroids by:

$$C_{\mu C}^{w} = \frac{n}{2\sum_i \sum_j a_{i,j}} \frac{\sum_i \sum_j a_{i,j} \int_{t \in w} [\overline{Y}^{k_i}(t) - \overline{Y}^{k_j}(t)]^2 dt}{\sum_i \int_{t \in w} [\overline{Y}^{k_i}(t) - \overline{Y}(t)]^2 dt}. \tag{8}$$

4 An Application

The test dataset collects the records of 54 sensors placed at the Intel Berkeley Research lab between 28th and April 5th, 2004. Mica2Dot sensors with weather boards collected time stamped topology information, along with humidity, temperature, light and voltage values once every 31 s. Data was collected using TinyDB in-network query processing system, built on the TinyOS platform. The dataset includes the x and y coordinates of sensors expressed in meters relative to the upper right corner of the lab. We have analyzed the light records of each sensor so that we have a set of 54 time series each one made by 65000 observations.

The main focus of the experiment is to evaluate the performance of the proposed method in recovering the functional spatial autocorrelation of the data. In particular,

we will compare the estimated functional Moran's index $I^w_{\mu C}$ of Eq. (6) with the Moran's index I computed on the real data as in Eq. (5). Similarly, we compare the estimated functional Geary's measure C on real data, as in Eq. (8), and that on real data, as in Eq. (7). The goodness of fit of both the indexes is measured by the mean absolute relative error. For the Moran's index, it is given by:

$$RMAE_M = \frac{1}{W}\sum_{w=1}^{W}\left|\frac{I^w - I^w_{\mu C}}{I^w}\right|. \tag{9}$$

where I^w is the Moran's index computed on the real data of the window w.

For the Geary's measure, it is given by:

$$RMAE_G = \frac{1}{W}\sum_{w=1}^{W}\left|\frac{C^w - C^w_{\mu C}}{C^w}\right|. \tag{10}$$

where C^w is the Geary's measure computed on the real data of the window w.

The temporal window w has size $b = 464$ such to cover approximately 4 h. An example of the dataset is provided in Fig. 1, where the light values (Lux) recorded in one temporal window in each of the 52 sensors are displayed together with their smoothed version. The smoothing has been obtained by using natural cubic splines with a number of knots equal to $\frac{b}{20}$. In order to initialize the micro-clusters of the Clustream algorithm, an initial clustering has been performed on the first 25 temporal windows of each sensor. The results we show have been obtained setting, for each sensor, the number of maximum clusters equal to 20 and the threshold value equal to

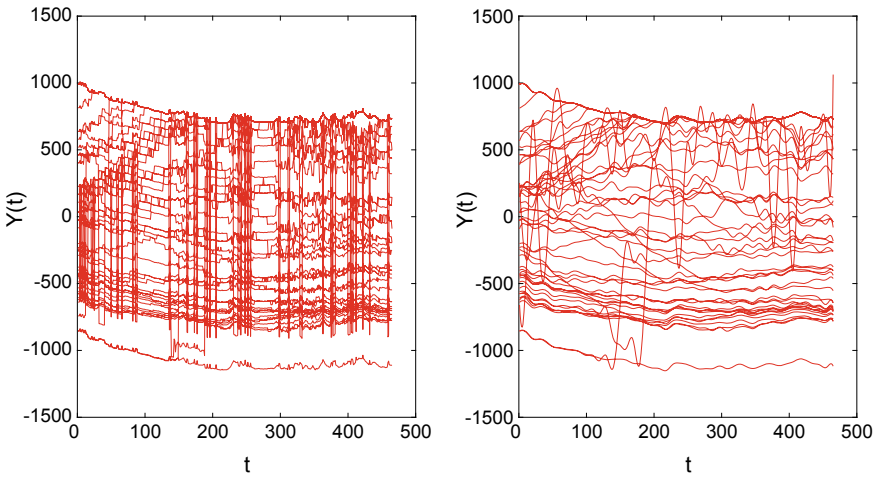


Fig. 1 Light data: raw data (left panel) and natural cubic splines smoothing (right panel)

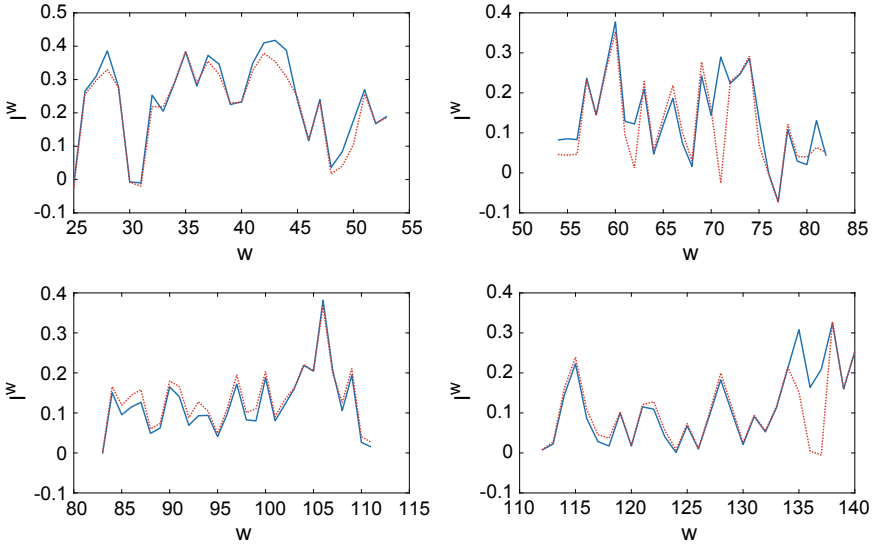


Fig. 2 Moran's index computed on real data (solid line) and on cluster prototypes (dotted line)

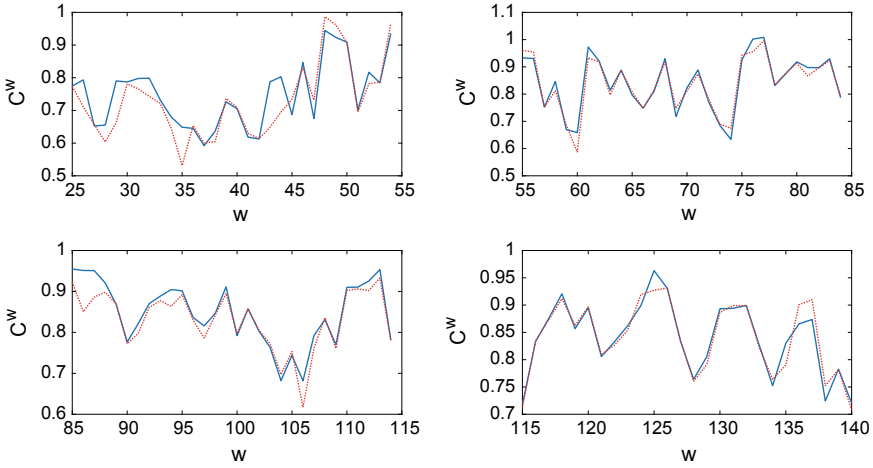


Fig. 3 Geary's C measure computed on real data (solid line) and on cluster prototypes (dotted line)

$u = 2$. We have made tests using different values of u in the range $0.5 - 10$ which show that with the growing of u there is a reduction of the the number of generated micro-clusters. The choice $u = 2$ has been made evaluating the compromise between accuracy and storage requirements. In Fig. 2, we display the functional Moran's index computed by using only the micro-cluster prototypes and the index computed using all the data available. Figure 3, replicates the plot of Fig. 2, using the Geary's C measure.

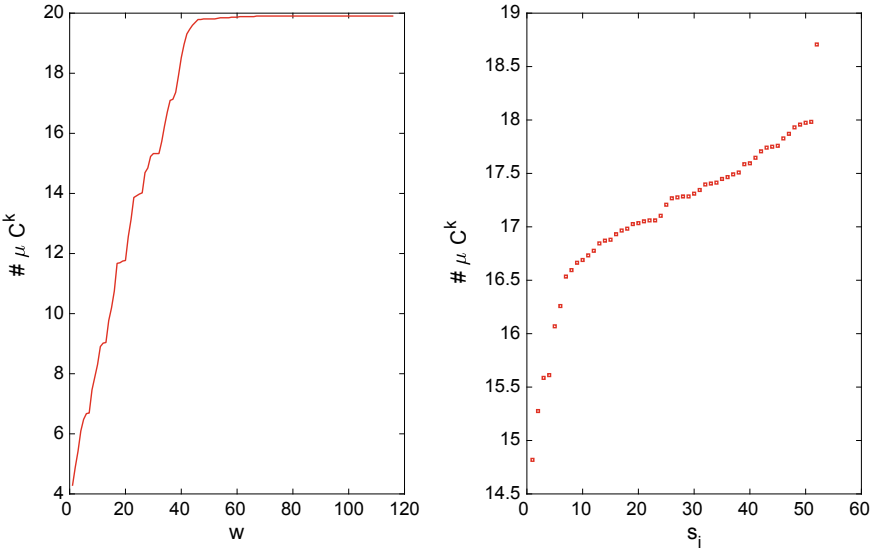


Fig. 4 Average cardinality of the micro-clusters sets across the temporal windows w (left panel) and the sensors s_i (right panel)

The relative absolute error is equal to $RMAE_I = 0.1571$ for the Moran's index, and $RMAE_G = 0.1722$, for the Geary's measure. Results show how the proposed methodology provide a good recovery of the spatial autocorrelation between the sensors. In Fig. 4, the average cardinality of the micro-cluster sets μC is displayed both across the temporal windows and the sensors.

5 Conclusions and Perspectives

In this paper we have introduced an approach for measuring the spatial autocorrelation among functional data streams recorded by sensors. Since the main spatial dependence measures require a high computational effort, we have proposed to perform a data summarization and to compute the spatial autocorrelation on the summaries rather than on the original data. Unlike to original data streams, summaries can be easily stored, thus, our method supports the possibility to recover information about the spatial correlation on past time periods for which sensor records have been deleted. Preliminary tests on a real data set confirm the effectiveness of the proposed summarization strategy in keeping track of the spatial correlation structure. Future work will focus on performing further tests on the sensitivity of the method to input parameters.

References

1. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for clustering evolving data streams. In: VLDB 2003: Proceedings of the 29th International Conference on Very Large Data Bases, p. 812. VLDB Endowment (2003)
2. Garofalakis, M., Gehrke, J., Rastogi, R: Data Stream Management: Processing High-Speed Data Streams. Springer, New York (2016)
3. Gattone, S.A., Rocci, R: Clustering curves on a reduced subspace. J. Comput. Graph. Stat. **21**(2), 361–379 (2012)
4. Geary, R.C.: The contiguity ratio and statistical mapping. Inc. Stat. **5**(3), 115145 (1954). https://doi.org/10.2307/2986645
5. Griffith, D.A.: Spatial autocorrelation: a primer. Resource publications in geography. Association of American Geographers (1987)
6. Hitchcock, D.B., Casella, G., Booth, J.G.: Improved estimation of dissimilarities by presmoothing functional data. J. Am. Stat. Assoc. **101**(473), 211–222 (2006)
7. Moran, P.A.P.: Notes on continuous stochastic phenomena. Biometrika **37**, 17–23 (1950)
8. Tobler, W.: A computer movie simulating urban growth in the Detroit region. Econ. Geogr. **46**(2), 234–240 (1970)
9. Wang, J.L., Chiou, J.M., Muller, H.G.: Functional data analysis. Annu. Rev. Stat. Its Appl. **3**, 257–295 (2016)

User Profile Construction Method for Personalized Access to Data Sources Using Multivariate Conjoint Analysis and Collaborating Filtering

Oumayma Banouar and Said Raghay

Abstract Current information systems provide access to multiple, distributed, autonomous and potentially redundant data sources. Their users may not know the sources they questioned, nor their description and content. Consequently, their queries reflect no more a need that must be satisfied but an intention that must be refined. The purpose of personalization is to facilitate the expression of users' needs. It allows them to obtain relevant information by maximizing the exploitation of their preferences grouped in their respective profile. In this work, we present a collaborative filtering method based on a Multivariate Conjoint Analysis approach to get these profiles. The proposed strategy provides a representation of the users and of the items, according to their characteristics, on factorial plans; whereas, the collaborative approach predicts the missing preferences.

Keywords Personalization · User profile · Preferences' predicates

1 Introduction

A user accessing an information system, for satisfying an information need, has to reformulate the query issued several times and sift many results until to obtain a suitable answer. This is a very common experience. The multiplicity of data sources, their scalability and the increasing difficulty to control their descriptions and their contents are the reasons behind the emergence of the need of users' requests personalization. A major limitation of information systems is their inability to classify and discriminate users based on their interests, their preferences and their query context. They cannot deliver relevant results according to their respective profiles [1]. Consequently, the execution of the same request, expressed by different users, over a data source will necessarily not provide the same results. We talk here about a per-

O. Banouar (✉) · S. Raghay
Laboratory of Applied Mathematics and Computer Science, Faculty of Science and Technics, Cadi Ayyad University, Marrakesh, Morocco
e-mail: o.banouar@edu.uca.ma

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_2

sonalized access to data sources. A critical observation is that: different users may find different things relevant when searching an information, because of different preferences, goals etc. Thus, they may expect different answers to the same query [2].

A special instance of this problem is the now famous problem known as Netflix problem. Users submit ratings on a subset of entries in a database, and the vendor provides recommendations based on users' preferences. Consider a simple case of two users, Al and Julie, access a web-based movies database both searching for comedies. Al is a fan of director W. Allen, while Julie is not. Most systems would consider only the request issued and return to both users the same, exhaustive list of comedies. However, storing user's preferences in his profile gives a system the opportunity to return more focused personalized (and hopefully smaller) answers.

The objective of the query personalization process is to enhance the user query with his related preferences stored in his profile. This process focuses on the system user, enables the exploitation of what is called personal relevancy [3] instead of consensus relevancy.

To construct the users' profiles, observations are stored in a matrix. It contains the scores that reflect the interest of users to items. The main characteristic of this matrix is its sparsity.

This work presents a collaborative filtering-based approach to predict the preferences of users to items. It is based on a Multivariate Conjoint Analysis approach [27] to represent users and items according to their characteristics, on factorial plans.

This approach focuses on the correspondence between the users preferences and the characteristics of the items.

On the factorial coordinates of the users and of the items is performed a biclustering by a double K-means, to identify clusters of users and items. Then, the prediction process exploits these clusters to predict the missing scores. To compute the score of a given user u to an item i, the process minimizes the nuclear norm of users-preferences matrix that contains the scores provided by the users that belong to the cluster of the selected user to the items of the cluster that contains the selected item.

The proposed algorithm is compared to existing matrix completion methods and multiplex network-based methods according to the following standard measures: Mean Absolute Error (MAE), Root Mean Square Error (MSE).

These methods are compared on the MovieLens dataset (https://grouplens.org/datasets/movielens/), that is a standard dataset used in collaborating filtering frameworks.

The remaining of this article is organized as follows: Sect. 2 refers to related works; Sect. 4 presents the problem statement and the nuclear norm minimization in the optimization of sparse and low-rank matrices; it also introduces the multivariate conjoint analysis and the proposed approach; Sect. 5 discusses application results on real data; the Conclusion section close the paper.

2 Related Work

Different works exist in literature that have proposed methods for user queries enrichment in documents retrieval context and especially in the educational domain. Each research work adopted a user profile model to personalize his query. The AHAM (Adaptive Hypermedia Application Model) is based on generic Dexter Model [4]. It splits up the storage layer of the Dexter Model into an adaptation model, a domain model, and a user model. The user model is an overlay of the domain model. Like in AHAM, authors in [5] have described the Munich reference model for adaptive hypermedia applications, where the adaptation engine can implement not only the educational-oriented rules but also other rules. An expert can define these rules for a particular domain. With the same concept, AHA! System [6] is defined for adaptive Web application. AHAM, Munich reference model and AHA! perform rule-based adaptation. Consequently, the adaptation engine implements only domain-based rules. Some works tried to reduce this restriction by defining non-persistent properties and post and pre-concepts access rules execution. Authors in [7] defined GAM that is a generic theoretical model for describing the user adaptive behaviour in a system in order to adapt interactive systems. This problem has been dealt with authors in [8] when blind users interact with information systems. These different architectures for adaptive hypermedia system are oriented towards particular domains or interactive systems [9–11]. Therefore, it is essential to change the domain model by a model relative to any domain. In this case, the user profile can include different interests. It is built from the analysis of user's queries. Another approach consists in the exploitation of observations about users.

The authors in [12] presented a solution that predict the missing ratings in users-preferences matrix. It starts by creating the multiplex network related to users and the one related to items. A community detection algorithm is then applied on the multiplex networks in order to find items partitions and users partitions. Once clusters are found for users and items, the predicting system finds, for a user u and for an item i, the predicted rate. It is the result of aggregating rates found by intersection between the clusters to which the user belongs and the cluster to which the item belongs to.

Clustering approaches for community detection in multiplex networks are the following:

- Methods based on monoplex approaches: [13–18].
  It consists on transforming the problem of clustering in multiplex network to a problem of clustering in simple graphs.
- Extending existing algorithms to deal directly with multiplex networks [19–24].

These solutions correspond to collaborating filtering through multiplex network clustering. They overcome the drawback of the presented solutions at first. Indeed, these solutions are independent of the application domain. However, they are applicable on data that can be represented as a graph.

Authors in [25] present an adaptation process of users' queries based on a profile construction method that is independent to the domain of application and to data representation. They exploited a matrix completion method by minimizing the nuclear

norm of users-preferences matrix. Matrix completion methods are very popular in image processing. They are mostly used for image recovery. Authors in this work, proposed two steps solution. It starts by applying a bi-clustering step to identify users, respectively, items' clusters. Then, it predicts the missing rating by computing the Singular Value Thresholding algorithm (SVT) on sub matrices of user-preferences matrix. For a user u and an item i, when a rate is unknown, a sub matrix is created. It contains the ratings that users belonging to the u cluster gave to items belonging to i cluster. It minimizes the nuclear norm of these sub matrices until the rates are totally predicted. A main drawback of this solution is it runs on the strong assumption that the users-preferences matrix contains at least on observation in each column. Therefore, before starting the computation of the proposed steps, a data-filtering step is mandatory. It eliminates the items that are not rated by any user. Consequently, it is impossible to predict the ratings corresponding to these items. They are already inexistent during the computation of the proposed solution. In this paper, we aim to propose a solution that is more complete.

3 Methods of User Profile Construction

3.1 Problem Statement

In many practical problems of interest, one would like to guess the missing entries of an $n_1 \times n_2$ matrix from a sampling Ω of its entries. This problem is known as the matrix completion problem. It comes up in a great number of applications including those of collaborating filtering. The collaborating filtering is the task of making automatic predictions about the interests of a user by collecting taste information from many users. In each instance, the objective is to predict the preferences of a user for all items from a partial list of his preferences for a few rated items or information gleaned from other users.

In mathematical terms, this problem is posed as follows:

A data matrix $\mathbf{M} \in \mathbb{R}^{n1 \times n2}$ is the matrix to be known as much as possible. The only information available about it is a sampling set of entries M_{ij}, $(i, j) \in \Omega$, where Ω is a subset of the complete set of entries $\{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$.

Very few factors contribute to an individual's taste. Then, the problem of matrix completion is a problem of a low-rank r matrix from a sample of its entries. The matrix rank satisfies $r \leq \min(n_1, n_2)$.

Such a matrix is represented by counting $n_1 \times n_2$ numbers but has only $r \times (n_1 + n_2 - r)$ degrees of freedom. When the matrix rank is small and its dimension is large, then the data matrix carries much less information than its dimension suggests.

At users, along of the rows of the matrix, are given the opportunity to rate items, columns of the data matrix. However, they usually rate very few ones so there are very few scattered observed entries of this data matrix. In this case, the users-preferences matrix is approximately low rank because as mentioned, it is commonly believed

Fig. 1 Users-preferences matrix

	f_1	f_2	f_3	f_4	f_5	f_6	f_7
u_1	2	-	-	-	4	-	-
u_2	-	-	3	-	5	-	1
u_3	-	4	-	2	-	1	-
u_4	-	-	5	3	-	2	-
u_5	5	-	-	1	-	2	5

that only very few factors contribute to an individual's tastes or preferences. These preferences are stored in a user profile.

The following figure illustrates an example of users-preferences matrix concerning five users denoted as u_1 to u_5 and seven films denoted as f_1 to f_7. Each user rates some preferences as to express the interest in each one. The ratings are usually numerical five-star scale. One and two stars represent negative ratings, three stars represent ambivalence while four and five stars represent positive ratings. The objective of our work is to predict the missing ratings in the matrix and then construct a complete user profile (Fig. 1).

3.2 Preliminaries

– *Matrix completion using nuclear norm minimisation*

Let $P_\Omega : \mathbb{R}^{n1\times n2} \rightarrow \mathbb{R}^{n1\times n2}$ be the orthogonal projection onto the subspace of matrices that vanish outside of Ω. We note that $(i, j) \in \Omega$ if and only if M_{ij} is observed.

Let $\mathbf{Y} = P_\Omega(X)$ be defined as follows:

$$Y_{ij} = \begin{cases} X_{ij}, & (i, j) \in \Omega \\ 0, & \text{otherwise,} \end{cases}$$

The data known in is given by $P_\Omega(\mathbf{M})$. The matrix is recovered, then from $P_\Omega(\mathbf{X})$ if it is the unique matrix of rank less or equal to r and consistent with the data, which means that, is the unique solution to:

$$\begin{aligned} &\text{minimize rank}(\mathbf{X}) \\ &\text{subject to } P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}) \end{aligned}$$

In practical point of view, the rank minimization problem is an NP-hard problem. Algorithms are not capable to resolve it in time once the matrices have an important dimension. They require time doubly exponential in the dimension of the matrix to find the exact solution.

Authors in [26] proposed the resolution of matrix completion problem by solving the nuclear norm minimization problem:

$$\begin{aligned} &\text{minimize} \|\mathbf{X}\|_* \\ &\text{subject to } \mathrm{P}_\Omega(\mathbf{X}) = \mathrm{P}_\Omega(\mathbf{M}) \end{aligned}$$

where the nuclear norm $\|\mathbf{X}\|_*$ is defined as the sum of its singular values: $\|\mathbf{X}\|_* : \sum_i \sigma_i(\mathrm{X})$.

Matrix completion problem is not as ill posed as thought. It is possible to resolve it by convex programming. The rank function counts the number of nonvanishing singular values when the nuclear norm sums their amplitude. The nuclear norm is a convex function. It can be optimized efficiently via semidefinite programming. Therefore, the first-order methods are used to complete large low rank matrices by solving the convex problem.

In the special matrix completion setting presented, $\mathrm{P}_\Omega(\mathbf{X})$ is the orthogonal projector onto the span of matrices vanishing outside of Ω. Therefore the (i, j) component of $\mathrm{P}_\Omega(\mathbf{X})$ is equal to $\mathbf{X}_{ij}$ if $(i, j) \in \Omega$ and 0 otherwise. $\mathbf{X} \in \mathbb{R}^{n1 \times n2}$ is then the optimization variable. Fix $\tau > 0$ and a sequence $\{\delta_k\}_{k \geq 1}$ of scalar step sizes.

Denoted shrink(Y_{k-1}, τ) a nonlinear function that applies a soft-thresholding rule at level τ to the singular values of the input matrix, and initialising $\mathrm{Y}_0 = 0 \in \mathbb{R}^{n1 \times n2}$, the algorithm computes:

$$\begin{cases} \mathrm{X}_k = \text{shrink}(\mathrm{Y}_{k-1}, \tau) \\ \mathrm{Y}_k = \mathrm{Y}_{k-1} + \delta_k \mathrm{P}_\Omega(\mathrm{M} - \mathrm{X}_k), \end{cases}$$

until a stopping criterion is reached.

The key property here is that for large values of τ, the sequence $\{\mathrm{X}_k\}$ converges to a solution which very nearly minimizes the nuclear norm. Hence, at each step, one only needs to compute at most one singular value decomposition and perform a few elementary matrix additions. The singular value shrinkage operator is the key building block of the matrix completion matrix SVT (Singular Value Thresholding) algorithm.

Consider the singular value decomposition SVD of a matrix $\in \mathbb{R}^{n1 \times n2}$ of rank r, $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^*$, $\Sigma = \text{diag}(\{\sigma_i\}_1 \leq i \leq r)$, where $\mathbf{U}$ and $\mathbf{V}$ are respectively $n_1 \times r$ and $n_2 \times r$ matrices with orthonormal columns, and the singular values σ_i are positive.

For each $\tau \geq 0$, the soft-thresholding operator D_τ is defined as follows:

$$\mathrm{D}_\tau(\mathbf{X}) := \mathbf{U}\mathrm{D}_\tau(\Sigma)\mathbf{V}^*, \mathrm{D}_\tau(\Sigma) = \text{diag}\big(\{\sigma_i - \tau\}_+\big),$$

where t_+ is the positive part of t, namely, $t_+ = \max(0, t)$.

This operator applies a shrinking operation to the singular values of $\mathbf{X}$. Effectively, it shrinks them towards 0. Even though the SVD may not be unique, it is easy to see that the singular value shrinkage operators are well defined. In some sense, this shrinkage operator is a straightforward extension of the soft-thresholding rule

for scalars and vectors. In particular, note that if many of the singular values of $\mathbf{X}$ are below the threshold τ, the rank of $D_\tau(\mathbf{X})$ may be considerably lower than that of $\mathbf{X}$, just like the soft-thresholding rule applied to vectors leads to sparser outputs whenever some entries of the input are below threshold. The singular value thresholding operator is the proximity operator associated with the nuclear norm.

The singular value thresholding SVT algorithm approximates the minimization by:

$$\begin{aligned} &\min \tau\|\mathbf{X}\|_* + \frac{1}{2}\|\mathbf{X}\|_F^2 \\ &\mathrm{X} \\ &\text{subject to } \mathrm{X_{ij}} = \mathrm{M_{ij}}, (\mathrm{i}, \mathrm{j}) \in \Omega \end{aligned}$$

with a large parameter τ. The matrix Frobenius norm or the square root of the summation of squares of all entries is denoted as $\| \cdot \|_F$ denotes.

Then, it applies a gradient ascent algorithm to its dual problem. The iteration is:

$$\begin{cases} \mathrm{X_k} = \mathcal{D}_\tau(\mathrm{Y_{k-1}}), \\ \mathrm{Y_k} = \mathrm{Y_{k-1}} + \delta_\mathrm{k} \mathrm{P}_\Omega(\mathrm{M} - \mathrm{X_k}), \end{cases}$$

where $\mathcal{D}_\tau$ is the SVT operator defined as:

$$\mathcal{D}_\tau(\mathrm{Y}) := \underset{\mathbf{X} \in R^{n1 \times n_2}}{\arg\min} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\|_F + \tau\|\mathbf{X}\|_*$$

– *Multivariate Conjoint Analysis using Principal Component Analysis onto a Reference Subspace PCAR*

Conjoint Analysis (CA) [2] deals with preference judgments expressed by users about a set of items, described by several attributes. Levels are the values assumed by each attribute. An experimental design regroups the level combinations for every item. CA aims to evaluate the importance of the attribute-levels in the determination of global preference for an item.

CA provides individual estimates according to each user.

The Metric Conjoint Analysis approach uses the multiple regression model in order to estimate the path-worth coefficient of each level. We talk here about items enrichment by their characteristics.

A factorial approach of Conjoint Analysis, hereafter denoted FCA, was proposed by [27]. Defined as:

– The design matrix $\mathbf{X}$ of dimension $\mathcal{Q} \times \mathrm{K}$, where the rows refer to $\mathcal{Q}$ items and the columns to the levels of p attributes.
– The preference matrix $\mathbf{Y}$ of dimension $\mathcal{Q} \times G$, where the rows refer to $\mathcal{Q}$ items and the columns represent the score given by G users.

The Metric Conjoint Analysis approach estimates the part-worth or utility coefficients by minimizing the following expression: $\min_{\mathbf{B}} \left\| \mathbf{Y} - \mathbf{XB}^2 \right\|$ where $\mathbf{B}$, of dimension $K \times G$, is the matrix of the individual part-worth coefficients associated to the attribute-levels in $\mathbf{X}$. It leads to the classical Ordinary Least Squares OLS results: $\hat{\mathbf{B}} = \Delta_X^{-1}\mathbf{X}'\mathbf{Y}$ where $\Delta_X = Diag(\mathbf{X}'\mathbf{X})$.

The decomposition of the part-worth coefficients on a factorial plan is carried out by a SVD of the matrix $\mathbf{A} = \hat{\mathbf{B}}'\Delta_X\hat{\mathbf{B}}$, assuming a system of weights elements of the diagonal matrix Δ_X to take into account the different number of levels of the items, given by:

$$\text{SVD}(\mathbf{A}) = \mathbf{U\Sigma V}' \text{under the constraints: } \mathbf{U}'\mathbf{U} = \mathbf{V}\Delta_X\mathbf{V}' = \mathbf{I} \text{ and } \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_Q).$$

The projection of the users and of the item-levels on the factorial plans, allows to analyse in a reduced space the relationship among users and attribute-levels with respect to the characteristics of the items defined by the design matrix [27].

- The coordinates of the users on the first m principal axes are computed according to the following expression: $\varphi_\alpha = V_\alpha\sqrt{\lambda_\alpha}, \alpha = 1, \ldots\ldots, m.$
- The coordinates of the levels are obtained by the following expressions: $\psi_\alpha = \Delta_X^{-1/2}\widehat{\boldsymbol{B}}V_\alpha\Delta_X^{-1}\mathbf{X}'\mathbf{Y}V_\alpha, \alpha = 1, \ldots\ldots, m.$
- The projections of items on the factorial plan are computed by means of the following formula: $\sigma_\alpha = \mathbf{X}\Delta_X^{-1}\mathbf{X}'\mathbf{Y}V_\alpha\sqrt{\lambda_\alpha}, \alpha = 1, \ldots\ldots, m.$

3.3 Proposed Approach for User Profile Construction

The proposed process of user profile construction relies on the following steps:

1. Enriching the users-preferences matrix by users and items characteristics by using FCA.
2. Identifying users and items' clusters by applying a learning process using a bi-clustering based K-means.
3. Deleting items that are not rates by any user through a data.
4. Imputing ratings by predictive method based on SVT algorithm according to the clusters achieved in step 2.
5. Estimating the ratings of eliminated items in step 3 through an assignment function.

- Multivariate Conjoint Analysis for users and preferences clusters identification

We adopt the FCA to obtain the correspondences between the users and items characteristics, by considering the design matrix **XU**. It contains in its columns the characteristics of users and the preference matrix **YI** has in its rows the items and its columns the users.

Consequently, the users and items' partitions in our approach are obtained by applying the K-means based bi-clustering step on respectively users and item-levels coordinates matrix $\boldsymbol{\sigma}\mathbf{V}_{\boldsymbol{\alpha}}$ and $\Delta_X^{-1/2}\mathbf{B}\mathbf{U}_{\boldsymbol{\alpha}}$.

The number of classes K are obtained by applying hierarchical clustering according to classical criterion of cutting of the dendrogram.

$\mathbf{YU}\sigma u_\alpha$ and $\mathbf{YI}\boldsymbol{\sigma}\mathbf{i}_{\boldsymbol{\alpha}}$. The users' coordinates matrix $\boldsymbol{\sigma}\mathbf{u}_{\boldsymbol{\alpha}}$ is obtained by applying the FCA by considering the design matrix **XU**. It contains in its columns the characteristics of users. The matrix of items coordinates $\boldsymbol{\sigma}\mathbf{i}_{\boldsymbol{\alpha}}$ is obtained by performing FCA on the design matrix **XI**. It contains in its columns the items' characteristics. The preference matrix **YI** has in its rows the items and its columns the users. The matrix **YU** is defined as the following $\mathbf{YU} = \mathbf{YI}'$. The number of classes k corresponds to the one obtained by applying hierarchical clustering.

– Matrix completion SVT algorithm for ratings prediction

The SVT algorithm works under the strong assumption that the preferences matrix contains at least one observation in each column. Therefore, before starting the computation of the proposed steps, a data-filtering step is mandatory. It eliminates the items that are not rated from the preferences matrix **YI**. The SVT algorithm then is applied on the sub matrices extracted from **YI**.

For a given user, respectively an item, we identify clusters in which the selected user, respectively the item, belongs. The predicted rate is the result of SVT algorithm applied on the matrix containing rates that users in the selected user cluster gave to items in the selected item cluster [25].

In some cases, the application of the SVT algorithm in blocks provides certain results that are out of range (The rates to be predicted has in most application a determined scale). In this case, we use an aggregation process to predict the following rates. It is equal to the mode of all rates found by intersection between the cluster to which the user belongs and the cluster that contains the selected item.

– Assignment function to estimate the ratings of eliminated

To provide a relevant solution, the proposed process uses an assignment function. This function has as an objective to find the users class that are interested by the selected item. It exploits the characteristics of the items to enrich the data matrix. It provides as a result the matrix of items weighted according to users' classes. Then for a certain item, it is possible to know the class of users that will be the most interested by it. The exact rate will be then equal to the aggregation of rates provided by the users of this class.

4 Experimental Results

To evaluate the proposed approach for users' profiles construction, we applied it on the MovieLens dataset. It is the standard dataset used for collaborative filtering testing. It consists of:

– 100 000 ratings from 943 users on 1682 films from 1 to 5.
– Each user has rated at least 20 movies.
– The dataset is 80, 20% splits into training and test data. It procures data through 5 bases (u1.base, u2.base…) with their test files (u1.test, u2.test …).
– It characterizes users by their: age, gender, occupation and zip code, where it characterizes the movies by 19 genres.

In the objective to demonstrate the efficiency of combining the aggregation method and the SVT algorithm per blocks, we applied several methods of Low-Rank Matrix Recovery and Completion over the same experimental data. These methods minimize also the nuclear norm of their users-preferences matrix in the aim to recover the missing data with precise rank. We cited Augmented Lagrange multiplier method ALM [28], Accelerated Proximal Gradient method APG [29], Dual Method DM [29] and Fixed-Point Continuation method FPC [30]. Only SVT, FPC and ALM algorithms recovered the matrix with the desired rank 943.

The proposed approach used the predictive process using the clusters obtained by the bi-clustering over the users and items coordinates in the factorial plan using PCAR. We compared the results of the same predictive process using the clusters obtained by performing the bi-clustering on the principle component scores and the correlation matrix to identify respectively users and items partitions [25] and applying the bi-clustering on the result of the SVD.

In addition, our approach is compared to the collaborating filtering approach using multiplex network. This approach used different community detection methods: Muxlicod algorithm [24], Layer aggregation LA [13] using Louvain [16] and Walktrap [17], Partition aggregation PA [18] using Louvain and Walktrap and Generative Topographic Mapping GTM [31].

As mentioned, The MoviesLens 100 K dataset is composed of 5 bases. The results then are the average of ones obtained by applying each approach on these bases. The following table presents the results obtained by the cited works and the proposed approach PA (Table 1).

The fact that our approach enriches the data matrix with users and items characteristics using PCAR augmented the precision of our clustering step. This also had a direct impact on reducing the MAE and RMSE. Indeed, finding precise clusters augment the precision of our predictive process. Our approach is based on the strong assumption that if users rate certain items similarly, or have similar behaviours or share similar characteristics, then they will rate or act on other items similarly.

Table 1 Comparaison results

Method	MAE	RMSE
PA	0.8044	1.1088
SVT in blocks + PCA	0.8062	1.1089
SVT in blocks + SVD	0.8204	1.1869
SVT	0.8956	1.3003
FPC	0.9759	1.3108
ALM	0.9781	1.3194
GTM	0.9441	1.2549
Muxlicod	0.9635	1.2773
LALouvain	0.8352	1.1509
LAwalktrap	0.8216	1.1155
PALouvain	0.8713	1.1917
PAwalktrap	0.8801	1.2023

5 Conclusion

The used prediction process in [25] runs on the assumption that the initial matrix contains at least one observation per row and one observation per column. That is why during its application on the MovieLens dataset, it provided rates out of scale. Therefore, to provide a relevant solution, we proposed to eliminate the items unrated by any user then applying the prediction process. In addition, we used an assignment function to estimate the rates of items eliminated. We also demonstrated that the enrichment of users-preferences matrix with users and items characteristics plays an important role in the prediction process. The exploitation of these characteristics allowed us to decrease the MAE and RMSE errors. We adopted a multivariate conjoint analytics based on the PCAR method to get the projections of users and items over a factorial plan. The bi-clustering step was then applied on these projections values.

References

1. Banouar, O., Raghay, S.: User profile construction for personalized access to multiple data sources through matrix completion method. Int. J. Comput. Sci. Net. Secur. **16**(10), 51–57 (2016)
2. Bozdag, E.: Bias in algorithmic filtering and personalization. Ethics Inf. Technol. **15**(3), 209–227 (2016)
3. Koutrika, G., Ioannidis, Y.: Personalizing queries based on networks of composite preferences. ACM Trans. Database Syst. **35**, 1–50 (2010)
4. D. Lewandowski, Evaluating the retrieval effectiveness of web search engines using a representative query sample, Journ. Of the Association for Info. Science and Technology, vol. 66, 2015, pp. 1763–1775

5. Gernanakos, P., Belk, M.: Human-Centered Web Adaptation and Personalization: From Theory to Practice. Human-Computer Interaction Series. Springer, Berlin (2016)
6. Kobsa, A., Koenemann, J., Pohl, W.: Personalised hypermedia presentation techniques for improving online customer relationships. Knowl. Eng. Rev. **16**(11), 111–155 (2001)
7. Bra, P.D., Aerts, A., Berden, B., de Lange, B., Rousseau, B., Santic, T., Smits, D., Stash, N.: Aha! the adaptive hypermedia architecture. In: Proceedings of 14th Conference on Hypertext and Hypermedia (HYPERTEXT'03), Nottingham, UK, 2003, pp. 81–84 (2003)
8. van der Weide, T., Bommel, P.v.: GAM: A Generic Model for Adaptive Personalisation, Technical Report ICIS–R No. 06022, Radboud University Nijmegen, Nijmegen, The Netherlands, EU, June 2006
9. Yakoubi, Z., Kanawati, R.: Licod: leader-driven approaches for community detection. Vietnam J. Comput. Sci. **14**, 241–256 (2014)
10. Turrin, R.: Personalization challenges in e-learning. In: Proceedings of 11th Conference on Recommender Systems (RecSys'17), Como, Italy, 2017, pp. 345–345 (2017)
11. Klašnja-Milićević, A., Vesin, B., Ivanović, M., Budimac, Z., Jain, L.C.: Personalization and adaptation in e-learning systems. E-Learn. Syst. **112** (2016)
12. Encelle, B., Jessel, N.: Adapting presentation and interaction with XML documents to user preferences. In: Proceeding of 9th Conference Computers Helpring People (ICCHP'04),Paris, France, 2004, pp. 143–150 (2004)
13. Falih, I., Grozavu, N., Kanawati, R., Bennani, Y.: A recommendation system based on unsupervised topological learning. In: Proceedings of 22nd Conference of Neural information Processing (ICONIP'15), Istumbul, Turkey, 2015, pp. 224–232 (2015)
14. Suthers, D.D., Fusco, J., Schank, P.K., Chu, K.H., Schlager, M.S.: Discovery of community structures in a heterogeneous professional online network. In: Proceedings of 46th Hawaii International Conference on System Sciences (HICSS'13), Maui, USA, 2013, pp. 3262–3271 (2013)
15. Berlingerio, M., Pinelli, F., Calabrese, F.: Abacus: frequent pattern mining-based community discovery in multidimensional networks. Data Min. Knowl. Discov **27**(3), 294–320 (2013)
16. Ahmed, R.K.A.: Applications of artificial neural networks in e-learning personalization. Int. J. Comput. Appl. **158**, 37–39 (2017)
17. Blondel, V.D., Guillaume, J.L, Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. (2008)
18. Pons, P., Latapy, M.: Computing communities in large networks using random walks. J. Graph Algo. Appl. **10**(2), 191–218 (2006)
19. Strehl, A., Ghosh, J., Cardie, C.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2002)
20. Lambiotte, R.: Multi-scale modularity in complex networks. In: Proceedings of 8th Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt'10), Avignon, France, 2010, pp. 546–553 (2010)
21. Amelio, A., Pizzuti, C.: A cooperative evolutionary approach to learn communities in multilayer networks. Parallel Problem Solving from Nature–PPSN XIII, 2014, pp. 222–232 (2014)
22. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community structure in timedependent, multiscale, and multiplex networks. Science **328**, 876–878 (2010)
23. Good, B.H., de Montjoye, Y.A., Clauset, A.: The performance of modularity maximization in practical contexts. Phys. Rev. (2010)
24. Kanawati, R.: Seed-centric approaches for community detection in complex networks. In: Proceedings 6th International Conference on Social Computing and Social Media (SCSM'14), Crete, Greece, 2014, pp. 197–208 (2014)
25. Banouar, O., Raghay, S.: Enriching SPARQL queries by user preferences for results adaptation. Int. J. Soft. Eng. Know. Eng. **28**, 1195–1221 (2018)
26. Cai, J., Candès, J.E., Zuowei, C.: A singular value thresholding algorithm for matrix completion. SIAM J. Opt. **20**(4), 1956–1982 (2010)
27. Lauro, C., Giordano, G., Verde, R.: A multidimensional approach to conjoint analysis. Appl. Stoch. Model. Data Anal. J. **14**, 265–274 (1998)

28. Lin, Z., Chen, M., Wu, L., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. UIUC Technical Report UILU-ENG No 09-2215 (2009)
29. Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M., Ma, Y.: Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. UIUC Technical Report UILU-ENG No 09-2214, (2009)
30. Hale, E.T., Yin, W., Zhang, Y.: Fixed-point continuation for l1-minimization: methodology and convergence. SIAM J. Opt. **19**(9), 1107–1130 (2008)
31. Bishop, C.M., Svensén, M., Williams, C.K.I.: Gtm: the generative topographic mapping. Neural Comput. **10**(1), 215–234 (1998)

Clustering Communities Using Interval K-Means

Carlo Drago

Abstract With regard to large networks there is a specific need to consider particular patterns relatable to structured groups of nodes which could be also defined as communities. In this work we will propose an approach to cluster the different communities using interval data. This approach is relevant in the context of the analysis of large networks and, in particular, in order to discover the different functionalities of the communities inside a network. The approach is shown in this paper by considering different examples of networks by means of synthetic data. The application is specifically related to a large network, that of the co-authorship network in Astrophysics.

Keywords Network analysis · Community detection · Interval data · Symbolic data analysis · Clustering · Symbolic clustering

1 Introduction

In recent years,[1] the amount of collected data possible to observe has increased exponentially (Manyika et al. [25]). At the same time there has been a data collection growth on an exceptional scale; this has been made possible by the great advances in databases and computer technology. In particular the lower costs of digital storage was a relevant determinant of this growth (Vjgen [34]). Due to this it is possible to consider the relevant intersections and of that on big data analysis and the research on the network analysis (Sellis and Horadam [33]). Networks are nowadays ubiquitous and they typically produce many data which can be explored and analyzed. These analyses can represent a clear added value (see Atzmueller et al. [2]). Network analyses can be conducted in a variety of ways: analyze the different network features present on data, but at the same time it is possible to mine into the different operations and events occurring on the network (Aggarwal [1]). A community is a set of

[1] I thank the referees for their helpful suggestions

C. Drago (✉)
Universitá degli Studi di Roma "Niccoló Cusano", Via Don Carlo Gnocchi 3, Rome, Italy
e-mail: carlo.drago@unicusano.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_3

nodes densely connected to each other but weakly connected when considering each different group of nodes (Fortunato [15]). There are important cases in which it could be very important to cluster the communities of a network. For example an important case is explained by the same author [15]: different communities are associated to different behaviors on the network. The different nodes on the community can be associated with a specific function or behavior inside the network. In order to predict the future behavior of the different nodes it could be crucial to determine the different communities and to understand the different patterns of observed similarity. In this sense, understanding the concept of a community on a network leads to a better understanding of the network behavior as a whole (see Coscia et al. [9]). For this, it is necessary to understand the entire network. The problem of the adequate representation is particularly relevant on big data, in this sense we have to decide the best representation to use community clustering means so as to take into account on the clustering process all the structural features of the communities and the attributes of the different nodes considered as a whole. In this sense, we can consider the structural characteristics and the attributes of the nodes (which characterize the communities). In order to cluster communities it is necessary to adequately represent the problem of representing the community structure of a network (Drago [12]). Thus it is important to consider where it is possible to find a relevant property on networks in which the nodes on the community show characteristics which are loosely connected to each other as well as with other nodes belonging to other communities (see in particular Girvan and Newman [20]). At the same time it is particularly important to propose an approach which could be based on interval data because we wish to take into consideration the entire community. In fact interval data can be relevant in representing adequately the different communities considered: this allows the retention of the relevant information of the communities considered. It is for this reason that communities are a very relevant object to consider. In fact, on a specific network the different vertices tend to react as a whole and so it could be relevant to cluster them as a whole. In this sense symbolic data (Billard and Diday [4]) are a natural way to represent the data we are considering. Networks can be considered and represented as symbolic data and this specific representation allows the handling of huge quantities of data (see in this sense the works of Giordano and Brito [18], Giordano et al. [19]). In this context following Chavent et al. [7] the observations can be collected on special data matrices. These data are characterized by multivalued descriptors, by which each specific community can be represented on the rows of the data matrix. These special data tables are called symbolic data tables. It is possible to observe that each value on the data matrix is characterized by an interval data. See in this sense Chavent et al. [7]. In our case the data matrix is represented by different intervals, for each community we have an interval data for each different variable considered (structural characteristic or attribute). For instance considering a structural characteristic like betweenness we can compute the interval of betweenness for each considered node of the community. At the same time we can build an interval for the measurement of attributes like the age (the interval of the age of the different members of the community considered). Here, the interval represents the range of variation between the betweenness or an attribute (for instance the age) considered inside the community.

In Sect. 2 we define the problem of community identification in the network whereas in Sect. 3 we present the interval representation of the network communities and the K-means clustering. In the Sects. 4 and 5 we consider some relevant examples from synthetic data (Sect. 4) and on real data (Sect. 5).

2 Community Identification

The first step of the analysis is based on the need to determine the different communities inside the network. In order to do this is needed an appropriate community detection algorithm (Zhao et al. [35] and Blondel et al. [5]). We start from a network G defined as:

$$G = (V, E) \tag{1}$$

where E are the edges and V the vertices of the network. First of all it is possible to characterize the single nodes or vertices V by their structural characteristics like the centrality which can be measured as the Freeman degree for the generic node w:

$$FDeg_{centrality}(w) = deg(w) \tag{2}$$

The degree is the number of connections for each considered node (see Wassermann and Faust [36]). So it could be defined as a local measure of centrality. A global measure of centrality can be considered the betweenness:

$$Betw_{centrality}(w) = \sum_{s \neq w \neq k \in V} \frac{\sigma_{s,z}(w)}{\sigma_{s,z}} \tag{3}$$

where s and z are two generic nodes distinct from w. The numerator of the expression is related to the paths which pass through the node w considering the path between s and z. The denominator is related to the total number of paths between s and z. See Wassermann and Faust [36]. Finally another relevant local centrality measure is the eigenvector centrality. This measure computes the centrality of a node, considering the centrality of the neighbors. So we can have the $Eigv_{centrality}$:

$$Av = \lambda v \tag{4}$$

where A is the adjacency matrix, and an eigenvalue λ. The adjacency matrix A is a square matrix representing the adjacency of two nodes on a graph (Wasserman and Faust [36]). The greatest eigenvalues it is possible to obtain on the expression are the considered centrality (see Newman [28]). There are many ways to identify the communities on a network; for a review and a comparison of the different methodologies which can be used for community detection on large networks see Harenberg [21]. It is important to note that there does not exist a single specific definition of

community (Fortunato [15]). The lack of a unique definition is also due to the fact that there can be many different configurations of nodes (different partitions) (see Reichardt and Bornholdt [31]). The author [15] in particular defines a very simple way to identify a community: when there are more vertices inside the relevant set of node than there outside of it. In this sense it is possible to say that the part of the network belonging to the different communities is more strongly connected than the zones relating to different connections, which are weakly connected. Fortunato [15] defines some indices which can be useful in identifying communities: giving the definition of density as the number of the connections in each node divided by the theoretical connections for each node (see Wassermann and Faust [36]). We can define as intra-cluster density:

$$d_{internal}(C) \tag{5}$$

where d is the density for each community C and at the same the inter-cluster density:

$$d_{external}(C) \tag{6}$$

In this sense we can have:

$$D = d_{internal}(C) - d_{external}(C) \tag{7}$$

and the value obtained as D is higher when it is possible to identify a community on the network (see Mancoridis et al. [24]). Typically the community detection methods tend to focus on the connections between the different nodes which are part of the same community. There are cases in which the different nodes tend not to be a unique part of an identified single community. The general assumption of these methodologies is that there is a direct emphasis on considering the connections inside the communities more than the connections between members of different communities (Zhao et al. [35]). The relevant requirement for detecting a community is connectedness. In particular we can expect a strong connection between the nodes which are part of the community. Each community can be also seen as a module of the entire network; in order to detect communities we need to take into account the modularity which can measure explicitly the capacity of a network to be divided into different modules (Blondel et al. [5]). The higher the modularity means that it is possible to divide a network into different communities. Where the modularity is lower it is not possible to detect relevant communities. So in this way the modularity allows to identify the different communities. The modularity (see Newman [26]) needs to be computed by considering a null model. In this sense the structure seems to be non existent (i.e. a random graph). So following Fortunato [15] we can define the modularity in this way:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{i,j} - K_{i,j}) \gamma(C_i, C_j) \tag{8}$$

where m is the number of the edges on the network, A is the adjacency matrix considered and $K_{i,j}$ is the number of edges which can be considered between the vertices i and j on the null model. Finally it is possible to consider the γ function which returns two possible values: 1 where the two vertices i and j belong to the same community and 0 if they are part of different communities. However in order to take into account also the degree of the vertices i and j (it is important to consider the degree distributions) we can write the modularity as follows (Fortunato [15] and Newman [26]):

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) \gamma(C_i, C_j) \tag{9}$$

where k_i and k_j are degree values for different vertices. In this way we obtain the communities inside the network. An alternative approach is that followed by Reichardt and Bornholdt [31] which introduces a different methodology based on the comparison between null models and a general one (see Newman [27]). These communities are important because they are a stylized way to represent the different structure of the network. In this sense we need to take into account the entire groups of nodes as a whole in order to cluster the communities considered entirely. Thus we consider all the nodes singularly by considering their statistical characteristics (also the structural characteristics related to the network structure see Wasserman and Faust [36]). We can start to measure the different communities as an interval data based on the different structural characteristics (for instance degree, betweenness ecc.). Then we can obtain a data matrix based on the different interval data. In particular we must consider for each community the measurements of the structural indicators (for instance betweenness) by taking into account the different intervals of each indicator by considering all the nodes belonging to the community. Finally we can proceed through the cluster analysis of the different intervals representing the communities.

3 K-Means Clustering of the Communities

The different communities (denoted as X) are characterized by a vector related to the different n observations considering the single node for the same variable b. So we can have:

$$X^b = (x_1, x_2, \ldots, x_n) \tag{10}$$

We can write the interval data (the measurement for the network community) in this way:

$$X^{I,b} = [\overline{x}, \underline{x}] \tag{11}$$

Table 1 The data matrix considered (the interval data relating to each community representation)

Node	degr	betw	att1	…
Community 1	[1:2]	[5:7]	[3:5]	…
Community 2	[2:3]	[1:4]	[1:8]	…
…	…	…	…	…

Each interval of the considered variable represents a measurement for the single community. By doing this we obtain for each different community X an interval I. Each interval is characterized by their upper bound $\overline{x}$ and lower bound $\underline{x}$. It is important to note that the intervals can at the same time be characterized by their radii and the midpoints. In this case each community can be also represented by a midpoint value. In particular it is possible to obtain a value of the interval midpoint for the generic variable b:

$$X_{center}^{I,b} = \frac{1}{2}(\underline{x} + \overline{x}) \tag{12}$$

and the radius of the interval:

$$X_{radius}^{I,b} = \frac{1}{2}(\overline{x} - \underline{x}) \tag{13}$$

The different intervals can be compared and can be considered in the descriptive analysis of a network. It is possible, for example, to consider the mean of the intervals when we want to obtain a mean for the different l communities which are part of a network; for the basic statistical methods for interval data see Gioia and Lauro [17]:

$$M^{I,b} = \left[\frac{1}{N}\sum_{l=1}^{N}\underline{x_l}, \frac{1}{N}\sum_{l=1}^{N}\overline{x_l}\right] \tag{14}$$

In order to cluster the different communities we depart from the measurement of the community characteristics by using interval data (so we consider the data matrix in Table 1). Many different clustering algorithms have been proposed in order to cluster interval data. In particular, interval clustering was considered first of all. One of the authors proposing clustering algorithms is Bock [6]. Our starting point is the data matrix shown by Chavent et al. [7]. In this sense following the authors we have:

At this point we need to consider the K-Means clustering algorithm in order to classify adequately the different communities and obtain the meta-communities. Following De Carvalho et al. [11] and Pen and Li [30] from the initial data matrix representing the different communities we can perform the K-means interval clustering by adequately considering the number of the clusters. The method considered generalizes the classical k-means to interval data. In particular we start from the data representation given by considering the intervals of the different communities, then we consider the method of Hartigan Wong in order to allocate the different intervals

on the k considered clusters (see Pen and Li [30]). In particular the number of the initial clusters can be derived from the a-priori information of the clustering process and then it is possible to repeat the procedure in order to measure the robustness of the clusters obtained (see Lauro and Gherghi [16]). In this sense we consider a sensitivity analysis of the results obtained by the first cluster analysis. At the same time in order to explore the different partitions which can be obtained we consider different matrices simultaneously with different variables to determine if the partition could be specifically related to a different matrix configuration.

4 Simulation Study

We can start by considering different networks simulating variant characteristics. In this study we consider different networks obtained by applying the R package igraph (see Csardi and Nepusz [10]). In order to perform the data analysis at community level we have also used the package RSDA (Rodriguez [32]). Thus we consider these different groups of networks:

- Barabasi Albert graph models (Barabasi and Albert [3])
- Erdos-Renyi graph models (Erdos and Renyi [13])
- Random Dot Product graph models (Nickel [29])
- Forest Fire network models (Leskovec et al. [22])

and other types of networks. We have considered many different networks in order to test the approach on different structures. Here we present some examples we have obtained from the simulation study. The results obtained from examples are important in order to derive some interpretation rules which can be considered on the results of the proposed approach. In the first case we consider an example from the network based on the Barabasi game. In Fig. 1 a network with 100 nodes based on the Barabasi model is shown. We obtain 9 communities as part of the community structure obtained by using the greedy optimization of the modularity (Clauset et al. [8]). In particular in Fig. 1 we are able to identify the different communities and the nodes to which they belong.

From the clustering analysis of the interval data relating to the communities we observe that there is a group or cluster of communities on the center with similar structural characteristics. This could be also observed by considering the different interval scatterplot diagrams (see Fig. 1). That means we are able to identify some groups of communities which tend to show similar characteristics (Freeman degree and eigenvector centrality, see Fig. 1).

Then we consider the 9 different communities obtained on the clustering process using the K-Means. In this case we look at the different data matrices using different specifications and structural measures in order to evaluate the different results. Two interpretative results need to be noted on the simulations: it is possible to observe that the lower bound appears not to be relevant; in particular the upper bound is relevant in discriminating the different communities. On the other hand, the differences which

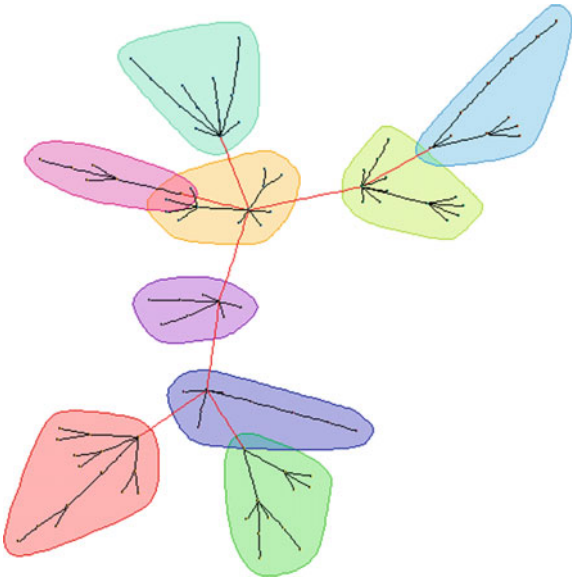


Fig. 1 Barabasi model simulation using 100 nodes: community structure

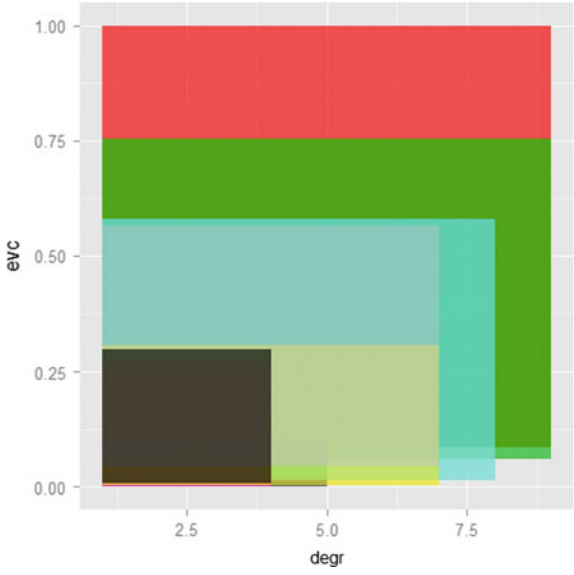


Fig. 2 Barabasi model simulation using 100 nodes: community structure visualized. On the x-axis the degree, on the y-axis the eigenvector centrality scores

can be observed by the different intervals can be determined specifically by the differences on the upper bounds. In fact it is possible to observe a higher heterogeneity between upper bounds rather than lower bounds. The visualization in Fig. 2 shows an overlapped structure because there is a centralized structure of the network. This structure tends to cluster specifically the communities in a central position. In this case the betweenness is related to the higher degree. It is also possible to note that

the different nodes can be characterized by different groups of similar nodes when they are considered specifically on the communities. We consider here as another example the case of the Erdos Renyi Model, using at the same time 100 nodes (see the interval scatterplot diagram in Fig. 2). We observe that the communities tend to be part of densely connected networks.

In these cases we can see that the different meta-communities can be influenced by the characteristics of the communities as well. In the next section we consider an experiment which has been carried out on real data.

5 Application on Real Data

The data used on the application are specifically related to the network of the researchers in t Theoretical Physics. The dataset is also present on the SNAP web-page (Leskovec and Krevl [23]). In particular we have used the dataset related to the 'General Relativity and Quantum Cosmology collaboration network' (Arxiv Gr-Qc). Here we have used the approach seen above and thus community detection based on the greedy optimization of the modularity (see Clauset et al. [8]) and the cluster analysis of the considered representations as intervals (see the interval scatterplot diagram obtained in Fig. 3). The network observed reveals a significant community structure which can be detected using the algorithms of community detection (the greedy optimization of the modularity). At the same time it is possible to visualize the community structure in such a way as to detect the general structure of the data. The results obtained from the analysis are coherent with Fay and Gautias [14]. The

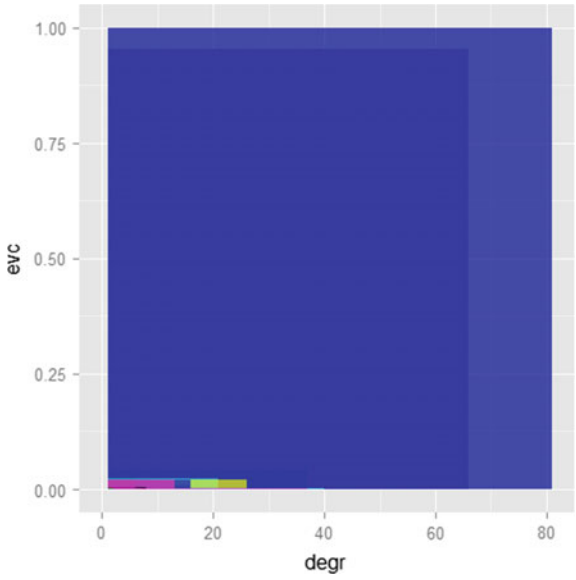


Fig. 3 Erdos Renyi model simulation using 100 nodes: community structure visualized. On the x-axis the degree, on the y-axis the eigenvector centrality scores

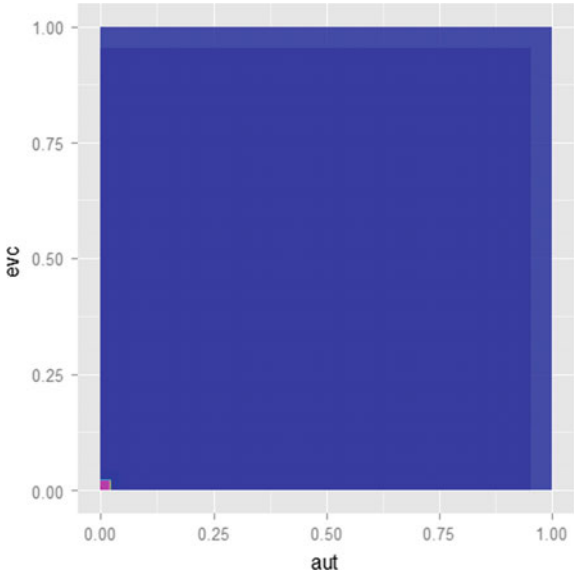


Fig. 4 Gr-Qc network. On the x-axis the degree, on the y-axis the eigenvector centrality scores

advantage of using this approach is the possibility to detect the different communities and the different meta-communities (the clusters of the different communities) which tend to show similar characteristics. In particular the approach based on the interval K-means offers the possibility to understand the different prototypes of the different communities it is possible to find. The final conclusion indicates a centralized structure of the network. By observing the meta-communities we can observe that there is an interesting difference on the prototypes of the most central communities with similar different levels of betweenness and degree centrality. This can be interpreted as indicating cooperative behavior between the vertices on the network (Fig. 4).

6 Conclusions

The results we have obtained confirm the usefulness of the approach considered in this work on large networks. The result is particularly useful to determine specifically the community structure and some different meta-communities which can be identified on a specific network. By starting from the meta communities it is possible to obtain the different prototypes. In particular a relevant observation relating to the results is that in the case of the clustering communities as groups of nodes we can obtain different results from those when considering only the clustering of the single node. In this sense the analysis can be enriched by the fact that in some cases the nodes have on their communities relevant dissimilarities which need not be taken into account when the analysis is performed by considering the communities as a whole. At the same time clusters of communities (or meta-communities) obtained by the clustering

of the interval data can be characterized as behavior by nodes which participate in the community on different levels. As in the case of scientific cooperation networks in Astrophysics it could indicate complex behaviors inside the same communities. The approach considered in this work allows the exploration of these levels of interaction between the different nodes.

References

1. Aggarwal, C.C.: Network Analysis in the Big Data Age: Mining Graphs and Social Streams. Keynote Talk, ECML/PKDD, 2014 (2014)
2. Atzmueller, M., Hotho, A., Strohmaier, M., Chin, A. (Eds.): Analysis of Social Media and Ubiquitous Data: International Workshops MSM 2010, Toronto, Canada, June 13, 2010, and MUSE 2010, Barcelona, Spain, September 20, 2010, Revised Selected Papers, vol. 6904. Springer (2011)
3. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
4. Billard, L., Diday, E.: From the statistics of data to the statistics of knowledge: symbolic data analysis. J. Am. Stat. Assoc. **98**(462), 470–487 (2003)
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. **2008**(10), P10008 (2008)
6. Bock, H.-H.: Clustering algorithms and kohonen maps for symbolic data. In: 'ICNCB Proceedings', Osaka, pp. 203–215 (2001)
7. Chavent, M, Francisco de A.T. De Carvalho, Yves Lechevallier, Rosanna Verde. New Clustering methods for interval data. Computational Statistics, vol. 21, pp. 211–229. Springer, Berlin (2006)
8. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**(6), 066111 (2004)
9. Coscia, M., Giannotti, F., Pedreschi, D.: A classification for community discovery methods in complex networks. Stat. Anal. Data Min.: ASA Data Sci. J. **4**(5), 512–546 (2011)
10. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJ. Complex Syst. **11**, 1695 (2006). http://igraph.org
11. De Carvalho, F., Souza, R., Chavent, M., Lechevallier, Y.: Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognit. Lett. **27**(3), 167–179 (2006)
12. Drago, C.: Exploring the community structure of complex networks. Annali del MEMOTEF - Note e Discussioni 10/2015; 2(forthcoming) (2015)
13. Erdos, P., Renyi, A.: On random graphs. Publ. Math. **6**(195), 290–297 (1959)
14. Fay, S., Gautrias, S.: A scientometric study of general relativity and quantum cosmology from 2000 to 2012. arXiv:1502.03471 (2015)
15. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3), 75–174 (2010)
16. Gherghi, M., Lauro, C.: Appunti di analisi dei dati multidimensionali: metodologia ed esempi. RCE edizioni (2004)
17. Gioia, F., Lauro, C.N.: Basic statistical methods for interval data. Stat. Appl. **17**(1), 75–104 (2005)
18. Giordano, G., Brito, P.: Social networks as symbolic data. In: Vicari, D., Okada, A., Ragozini, G., Weihs, C. (eds.) Analysis and Modeling of Complex Data in Behavioral and Social Science, pp. 133–142. Springer, Heidelberg (2014)
19. Giordano, G., Signoriello, S., Vitale, M.P.: Comparing social networks in the framework of complex data analysis. CLEUP Editore, Padova: pp. 1–2, In: XLIV Riunione Scientifica Societ Italiana di Statistica (2008)
20. Girvan, M., Newman, M.E.: Community Structure in Social and Biological Networks (2002)

21. Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Samatova, N.: Community detection in large scale networks: a survey and empirical evaluation. Wiley Interdiscip. Rev.: Comput. Stat. **6**(6), 426–439 (2014)
22. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: densification and shrinking diameters. ACM Trans. Knowl. Discov. Data (TKDD) **1**(1), 2 (2007)
23. Leskovec, J., Krevl, A.: SNAP datasets: Stanford large network dataset collection. http://snap.stanford.edu/data (2014)
24. Mancoridis, S., Mitchell, B.S., Rorres, C., Chen, Y., Gansner, E.R.: in IWPC '98: Proceedings of the 6th International Workshop on Program Comprehension. IEEE Computer Society, Washington, DC, USA (1998)
25. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute (2011)
26. Newman, M.E.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
27. Newman, M.E.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**(3), 036104 (2006)
28. Newman, M.E.: The mathematics of networks. New Palgrave Encycl. Econ. **2**(2008), 1–12 (2008)
29. Nickel, C.L.M.: Random dot product graphs: A model for social networks, Vol. 68, no. 04. (2007)
30. Peng, W., Li, T.: Interval data clustering with applications. In: 2006. ICTAI'06. 18th IEEE International Conference on Tools with Artificial Intelligence, pp. 355–362. IEEE (2006)
31. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E **74**(1), 016110 (2006)
32. Rodriguez, O.R. with contributions from Calderon, O., Zuniga, R.: RSDA: RSDA- R to symbolic data analysis. R package version 1.2. http://CRAN.R-project.org/package=RSDA (2014)
33. Sellis, T., Horadam, K.: Big data and complex networks analytics. IEEE Access **4**, 1958–1996 (2015)
34. Vijgen, R.: Big data, big stories. New Challenges for Data Design, pp. 221–234. Springer, London (2015)
35. Zhao, Y., Levina, E., Zhu, J.: Community extraction for social networks. Proc. Natl. Acad. Sci. **108**(18), 7321–7326 (2011)
36. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)

Text Mining and Big Textual Data: Relevant Statistical Models

Fionn Murtagh

Abstract A general overview is provided through examples and case studies, retrieved from research experiences, to foster description and debate on effectiveness in Big Data environments. At issue are early stage case studies relating to: research publishing and research impact; literature, narrative and foundational emotional tracking; and social media, here Twitter, with a social science orientation. Central relevance and importance will be associated with the following aspects of analytical methodology: context, leading to availing of semantics; focus, motivating homology between fields of analytical orientation; resolution scale, which can incorporate a concept hierarchy and aggregation in general; and acknowledging all that is implied by this expression: correlation is not causation. Application areas are: quantitative and also qualitative assessment, narrative analysis and assessing impact, and baselining and contextualizing, statistically and in related aspects such as visualization.

Keywords Correspondence analysis · Chronological hierarchical clustering · Mapping narrative · Emotion tracking · Significance of style

1 Introduction

1.1 Statistical Analytical Challenges in Data Science

A seminal paper in statistics is [6], describing current challenges and new developments in statistics. It can be emphasized that Hand [6] is addressing the increasingly major and contemporary domains of Data Science and Big Data analytics. The title refers to "administrative and transaction data", coming from the statistical work of companies and of government and authority agencies. The section headings in [6] are mostly relating to data quality and implying also, data encoding, and data curation.

F. Murtagh (✉)
Centre of Mathematics and Data Science, University of Huddersfield, Huddersfield, UK
e-mail: fmurtagh@acm.org

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_4

For how comprehensive or insightful one's data sources are, there is the need (Sect. 3 title in [6], " 'Data=all'?") to interact with our data sources, to have visualization and verbalization of our data.

Following this is our focus of analysis, in [6], Sect. 4 title, "Answering the right question". In the Sect. 5, "Causality and intervention", observational data is now our main engagement and, of course, there are limitations in how we can interact and engage with the domains that are at issue. For establishing causality, "The most common way to do this is via a properly controlled experiment involving randomization."

The Sect. 6 title in [6], is "Combining data from different sources". This can lead to the importance of triangulation, which is so very important in understanding the narrative of behavioural patterns, of analytical processes, and many other such themes and issues. The Sect. 7 title is "Confidentiality, privacy and anonymization". For security and also for the ethical issue of the individual not being thoroughly replaced by the cluster or group, there can and there should be full and complete account taken of both security and such ethical issues.

Having noted in [6] that "administrative data are ... typically not random samples", in this paper here, we just consider such viewpoints and perspectives for the text data source context. It is clearly described how an issue of importance is that population rather than sampling is at issue. (Hence, here, the corpus of retained terms from our text sources.) Other than summarization, in practice another consideration can be that the data is aggregated, leading to different resolution scales of the analysis.

There is a very interesting presentation of the term "survey" in [6], p. 3, and then how administrative data is purely factual as opposed to having what may be implicit explanatory or contextual or causal properties. For sampled data, very often the context of the population can have known or presupposed distributional properties, possibly from the law of large numbers, or long tailed (exponential) distributions.

When starting with the distinction between data that are collected for statistical and administrative purposes directly and very clearly leads for any reader to the current epoch of Big Data, which "we might define as the result of some automatic data collection system" [6], and that so much and such data are "nowadays largely collected automatically".

It is just interesting to note how data collected for statistical purposes can be subject to bias but, if so, the Big Data context can be useful for calibration purposes: [8]. At issue is self-selection from social media, and hence the need to calibrate (i.e., benchmark) such data. Such calibration may also be considered as, or linked to, the context of the data sources, or the objectives of the analysis.

For the context [6], "data describing different kinds of entities might have different characteristics", so therefore data encoding can be important. The entire data quality discussion can be related to the context of the data sourcing. If "administrative data quality is a multidimensional issue, with a hierarchy of dimensions" [6], then it may follow that the real-number system is less relevant compared to other number systems (p-adic number systems; this of course includes binary numbers that are essential to computers, even if earlier and some current viewpoints are that computers could and should be based on ternary, 3-adic, numbers, [16]. Measurement, e.g. having

quantitative or qualitative (possibly also termed, categorical) variables, is central to such an issue. Data encoding is ultimately very crucial.

This may well lead, not just to nearest neighbour and clusterwise regression, but also to ultrametric regression. From the hierarchically structured semantics underpinning the past terrible social violence in Colombia, regression of that data with the dependent regression variable being from the drugs market in the United States is at issue in [20].

In a sense, this may be alternatively expressed as follows [6]: "quality is not a property of the data set itself, but of the interaction between the data set and the use to which it is put". So much current analytical work, and related research, suffers from the note made that machine learning "places more emphasis on the final modelling stage of data analysis. This can be unfortunate: feed data into an algorithm and a number will emerge, whether or not it makes sense." Perfect data is not to be assumed. Relevance is crucial. Data aggregation may also be relevant here. That can be expressed thus: resolution scale of our data is another contextual, including relevance, aspect. Continuing this summarization of [6], data storage and rebranding of data is, in effect, the issue of data curation. There is the pointing to how much of an issue, the dynamics, of what we are dealing with, can be in our analytics. Given how we are dealing with data population rather than sampling of data in the discussion here, this points to the need to have informative and revealing ways to associate our data point clouds in our multidimensional spaces. This can be data calibration, and certainly it should also be, for interpretative objectives, but also innovative pattern recognition and trend finding, in the discovery of, and determining of, homologies in our data clouds and subclouds ([16], Preface, and Sect. 2.6).

The Sect. 4 title in [6] is "Answering the right question", and here there is this necessary association of the analysis with the data: "it can be useful ... to have statisticians involved in the data collection process." So fundamental in Data Science is that there be integration of all the data and all of the analytics. This leads to the Data Scientist's perspective that analytics is all about the visualization and the verbalization of data, (an edited book title, [3]). Referring to how "Statistical analysis methods are often divided into *descriptive* and *inferential*." is nicely describing the analytical narrative.

Calibrating survey data, cf. [8], the way that Big Data now supports and backs up, and can even contextualize, survey data, and such classically and statistically oriented data sourcing. From [6] p. 16: "A particular merit of administrative data, and especially of transaction data, is that it is recorded as time progresses." It is therefore continuous, over time. "This means that administrative data can be very useful for early detection of changes in populations."

Reference [6], "As is well known, observational data present challenges in establishing causality." The issue here is that the data is all that we have, while we must have further, where necessary, interaction with the contextual framework, or basis, of our data. "To establish causality, we need to intervene to break all possible causal links except the link that we wish to test ... The most common way to do this is via a properly controlled experiment involving randomization." All in all here, we are involved with the visualization and verbalization of data. Section 6: "Combining data

from different sources". This involves: data integration, information fusion, and, in the analytics, complement, supplement, add to accuracy. This is so very relevant, for all aspects of validation and verification of the methodology at issue. This is so very important, among the statistical challenges: "Challenge 14. Develop improved methods for data triangulation, combining different sources and types of data to yield improved estimates." From our data analysis and mathematical perspective, we know so very clearly, that triangulation expresses anomaly and innovation, and the strong triangular inequality is a natural expression of what is inherently hierarchically structured, and that means, to be structured as an ultrametric topology.

In [16], it is noted how important hierarchical clustering is, and therefore ultrametric topology embedding, for continuity and anomaly detection. Such work, on the inherent ultrametricity, i.e., the inherent hierarchical structure in complex data has been carried out for medical and environmental data, for psychoanalysis, for literature and for other textual data.

1.2 Contexts and Frameworks for Statistical Analytics

Clearly, through integration of analytical methodology and domain of application, the choice of methodology or even its development is dependent on the specific requirements. However the following general aspects of contemporary analytics, including textual data analytics, are useful to be noted.

Ethical consequences of Big Data mining and analysis may be associated with the following, from [11]: "Rehabilitation of individuals. The context model is always formulated at the individual level, being opposed therefore to modelling at an aggregate level for which the individuals are only an 'error term' of the model."

In [8], "There is the potential for big data to evaluate or calibrate survey findings ... to help to validate cohort studies". Examples are discussed of "how data ... tracks well with the official", far larger, repository or holdings. It is well pointed out how one case study discussed "shows the value of using 'big data to conduct research on surveys (as distinct from survey research)". Limitations though are clear: "Although randomization in some form is very beneficial, it is by no means a panacea. Trial participants are commonly very different from the external pool, in part because of self-selection, ...". This is due to, "One type of selection bias is self-selection (which is our focus)". Important points towards addressing these contemporary issues include the following. "When informing policy, inference to identified reference populations is key": This is part of the bridge which is needed, between data analytics technology and deployment of outcomes.

Furthermore there is this: "In all situations, modelling is needed to accommodate non-response, dropouts and other forms of missing data. While "Representativity should be avoided", here is an essential way to address in a fundamental way, what we need to address: "Assessment of external validity, i.e. generalization to the population from which the study subjects originated or to other populations, will in principle proceed via formulation of abstract laws of nature similar to physical laws".

Hence our motivation for the following framework for analytical processes: Euclidean geometry for semantics of information; hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. A further useful case is when the hierarchy respects chronological or other sequence information.

2 Towards: Qualitative as well as Quantitative Research Effectiveness and Impact

In the previous section, at issue has been advanced challenges, relating to data encoding. Relating also to this is qualitative and quantitative data.

For analysis of research funding, of publishing, and of commercial outcomes, account needs to be taken of measures of esteem. Also account is taken of research impact, through impact of research products: (1) research results, (2) organisation of science (journal editing, running conferences), (3) knowledge transfer, supervision, (4) technology innovations.

Correspondence Analysis when based on part of an ontology or concept hierarchy (i.e., when having qualitative data from such sources) can be considered as "information focusing". Correspondence Analysis provides simultaneous representation of observations and attributes. We project other observations or attributes into the factor space: these are supplementary or contextual observations or attributes. A 2-dimensional or planar view is an approximation of the full cloud of observations or of attributes. Therefore there can be benefit in the following: define a small number of aggregates of either observations or attributes, and carry out the analysis on them. Then project the full set of observations and attributes into the factor space.

In support of "The Leiden Manifesto for research metrics", DORA (San Francisco Declaration on Research Assessment), Metrics Tide Report (HEFCE, Higher Education Funding Council England, 2015), qualitative judgement is primary. Research results may be assessed through first determining a taxonomic rank by mapping to a taxonomy of the domain (a manual action). There then will be unsupervised aggregation of criteria for stratification.

Research impact should be evaluated, first of all, based on qualitative considerations. Evaluation of research, especially at the level of teams or individuals can be organized by, firstly, developing and maintaining a taxonomy of the relevant subdomains and, secondly, a system for mapping research results to those subdomains that have been created or significantly transformed because of these research results. Of course, developing and/or incorporating systems for other elements of research impact, viz., knowledge transfer, industrial applications, social interactions, etc., are to be taken into account also.

In [22], there is such an implementation, by having features of published research work that are defined and valued "manually" from an ontology, i.e. a taxonomy, or a conceptual hierarchy, of the relevant subdomains, i.e. themes at issue for research.

From such qualitative characterizing of research work, in [22], there is the deriving of quantitative values and that is used for ranking and for stratification. The envisaged application domain there is for editorial work in research publishing, and also for managing research funding proposals.

This other work, [7], has a very interesting theme: by mapping out text content, and its vocabulary, as it changes or evolves over time, over the years, and associating this with influential published articles, this work was studying the historical evolution of statistics, from the content of publications.

3 Qualitative Style in Narrative for Analysis and Synthesis of Narrative

For [13], the composition of the movie, Casablanca, is "virtually perfect". Text is the "sensory surface" of the underlying semantics.

Here there is consideration as to how permutation testing and evaluation can be very relevant for qualitative appraisal. Considering the Casablanca movie, shot by Warner Brothers between May and August 1942, and also some early episodes of the CSI Las Vegas, Crime Scene Investigation, television drama series, from the year 2000, the attributes used were as follows, [19].

All is based on the following: Euclidean geometry for semantics of information; hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. The hierarchy respects chronological or other sequence information. Chronological hierarchical clustering, also termed contiguity constrained hierarchical clustering, is based on the complete link agglomerative clustering criterion [1, 9, 14].

1. Attributes 1 and 2: The relative movement, given by the mean squared distance from one scene to the next. We take the mean and the variance of these relative movements. Attributes 1 and 2 are based on the (full-dimensionality) factor space embedding of the scenes.
2. Attributes 3 and 4: The changes in direction, given by the squared difference in correlation from one scene to the next. We take the mean and variance of these changes in direction. Attributes 3 and 4 are based on the (full-dimensionality) correlations with factors.
3. Attribute 5 is mean absolute tempo. Tempo is given by difference in scene length from one scene to the next. Attribute 6 is the mean of the ups and downs of tempo.
4. Attributes 7 and 8 are, respectively, the mean and variance of rhythm given by the sums of squared deviations from one scene length to the next.
5. Finally, attribute 9 is the mean of the rhythm taking up or down into account.

For permutation testing, assessment was carried out relative to uniformly randomized sequences of scenes or sub-scenes.

This allows this concluding outcome: this is how we statistically assess how the movie, "Casablanca", approximates artistic perfection.

4 Statistical Significance of Impact

Underlying [21] is the testing of social media with the aim of designing interventions, associated with statistical assessment of impact. The application here is to environmental communication initiatives. Measuring impact of public engagement theory, in the sense of the eminent political scientist, Jürgen Habermas, involves public engagement centred on communicative theory; by implication therefore, discourse as a possible route to social learning and environmental citizenship.

The case study here, was directed towards:

1. Qualitative data analysis of Twitter.
2. Nearly 1000 tweets in October, November 2012.
3. Evaluation of tweet interventions.
4. Eight separate twitter campaigns carried out.

Mediated by the latent semantic mapping of the discourse, semantic distance measures were developed between deliberative actions and the aggregate social effect. We let the data speak in regard to influence, impact and reach.

Impact was algorithmically specified in this way: semantic distance between the initiating action, and the net aggregate outcome. This can be statistically tested through the modelling of semantic distances. It can be further visualized and evaluated.

A fundamental aspect of the Twitter analysis was how a tweet, considered as a "campaign initiating tweet", differed from an aggregate set of tweets. The latter was the mean tweet, where the tweets were first mapped into a semantic space. The semantic space is provided by the factor space, which is endowed with a Euclidean metric. For very high dimensions, we find "data piling" or concentration. See [16]. That is, the cloud of points becomes concentrated in a point. Now that could be of benefit to us, when we are seeking a mean (hence, aggregate) point in a very high dimensional space. A further aspect is when it is shown that the cloud piling or concentration is very much related to the marginal distributions.

Here we show how we can test the statistical significance of effectiveness.

The campaign 7 case, with the distance between the tweet initiating campaign 7, and the mean campaign 7 outcome, in the full, 338-dimensional factor space is equal to 3.670904.

Compare that to all pairwise distances of non-initiating tweets. We verified that these distances are normal distributed, with a small number of large distances. By the central limit theorem, for very large numbers of such distances, they will be normal distributed. Denote the mean by μ, and the standard deviation by σ. Mean and standard deviation are defined from distances between all non-initiating tweets, in the full dimensionality semantic (or factor) space. We find $\mu = 12.64907$, $\mu - \sigma = 8.508712$, and $\mu - 2\sigma = 4.368352$.

We find the distance between initiating tweet and mean outcome, for campaign 7, in terms of the mean and standard deviation of tweet distances to be: $\mu - 2.168451\sigma$. Therefore for $z = -2.16$, the campaign 7 effectiveness is significant at the 1.5% level

(i.e. $z = -2.16$, in the two-sided case, has 98.5% of the normal distribution greater than it in value).

In the case of campaigns 1, 4, 5, 6, their distances between initiating tweet and mean outcome are less than 90% of all tweet distances. Therefore the effectiveness of these campaigns is in the top 10% which is not greatly effective (compared to campaign 7).

In the case of campaigns 3 and 8, we find their distances to be less than 80% of all tweet distances. So their effectiveness is in the top 20%.

Finally, campaign 2 is the least good fit, relative to initiating tweet and outcome.

5 Baselining or Contextualizing Analysis

The following is in regard to baselining, i.e. contextualizing, against healthy reference subjects, from a case study, in chapter "Bayesian analysis of ERG models for multilevel, multiplex, and multilayered networks with sampled or missing data" of this book, [12], a most important book, the content of which and its title relate, in effect, all of the data mining analytical work at issue here, with statistical modelling and hypothesis testing. This repeats some of the description in [17], in regard to testing through statistically baselining or contextualizing in a multivariate manner.

In [2], there is an important methodological development, concerning statistical inference in Geometric Data Analysis, i.e. based on MCA, Multiple Correspondence Analysis. At issue is statistical "typicality of a subcloud with respect to the overall cloud of individuals". Following an excellent review of permutation tests, the data is introduced: 6 numerical variables relating to gait, body movement, related to the following; a reference group of 45 healthy subjects; and a group of 15 Parkinsons illness patients, each before and after drug treatment. Reference [10] (Sect. 11.1) relates to this analysis, of the, in total, $45 + 15 + 15$ observation vectors, of subjects between the ages of 60 and 92, of average age 74.

First there is correlation analysis carried out, so that when Principal Components Analysis (PCA) of standardized variables is carried out, it is the case that the first two axes explain 97% of the variance. Axis 1 is characterized as "performance", and axis 2 is characterized as "style". Then the two sets of, before treatment, and after treatment, 15 Parkinsons patients are input into the analysis as supplementary individuals. Reference [2] is directly addressing statistically the question of effect of treatment. Just as in [10], the healthy subjects are the main individuals, and the treated patients, before and after treatment, are the supplementary individuals. This allows to discuss the subclouds of the before, and of the after treatment individuals, relative to the first, performance, axis, and the second, style, axis. The test statistic, that assesses statistically the effect of medical treatment here, is a permutation-based distributional evaluation of the following statistic. The subcloud's deviations relative to samples of the reference cloud are at issue. The Mahalanobis distance based on covariance structure of the reference cloud is used. The test statistic is the Mahalanobis norm of deviations between subcloud points and the mean point of the reference cloud.

In summary, this exemplifies in a most important way, how supplementary elements and the principal elements are selected and used in practice. The medical treatment context is so very clear in regard to such baselining, i.e. contextualizing, against healthy reference subjects.

6 Tracking Emotion

This relates to determining and tracking emotion in an unsupervised way. This is as opposed to machine learning, like in sentiment analysis, which is supervised. Emotion is understood as a manifestation of the unconscious. Social activity causes emotion to be expressed or manifested. This can lead to later discussion of psychoanalyst, Matte Blanco. See [16].

The foundation of this tracking of emotion, and determining the depth of emotion, is using the methodology of metric space mapping and hierarchical topology. The former here maps the textual data into a Euclidean metric endowed factor space, and the latter may be chronologically constrained hierarchical clustering.

The examples to follow are based on: in the Casablanca movie, dialogue (and dialogue only) between main characters Ilsa and Rick, having selected this dialogue from the scenes with both of these protagonists (scenes 22, 26, 28, 30, 31, 43, 58, 59, 70, 75 and last scene, 77); and Chaps. 9, 10, 11, 12 of Gustave Flaubert's 19th century novel, Madame Bovary. This concerns the three-way relationship between Emma Bovary, her husband Charles, and her lover Rodolphe Boulanger.

Following [18], in Fig. 1 in the full dimensionality factor space, based on all interrelationships of scenes and words, the distance between the word "darling" in this space, was determined with each of the 11 scenes in this space. The same was done for the word "love". The semantic locations of these two words, relative to the semantic locations of scenes 30 and 70 are highlighted with boxes.

Then in Fig. 2, hierarchical clustering, that is sequence constrained, is carried out on the scenes used, i.e. scenes 22, 26, 28, 30, 31, 43, 58, 59 70, 75, 77 (using the dialogue, between Ilsa and Rick). See how the big changes, here, in regard to emotion, in scenes 30 and 70 are indicated in the previous figure.

Now there is consideration of the novel Madame Bovary, by Gustave Flaubert, with the 3-way interrelationships of Emma Bovary, her husband Charles, and her lover, Rodolphe. Figure 3 presents an interesting perspective that can be considered relative to the original text. Rodolphe is emotionally scoring over Charles in text segment 1, then again in 3, 4, 5, 6. In text segment 7, Emma is accosted by Captain Binet, giving her qualms of conscience. Charles regains emotional ground with Emma through Emma's father's letter in text segment 10, and Emma's attachment to her daughter, Berthe. Initially the surgery on Hippolyte in text segment 11 draws Emma close to Charles. By text segment 14 Emma is walking out on Charles following the botched surgery. Emma has total disdain for Charles in text segment 15. In text segment 16 Emma is buying gifts for Rodolphe in spite of potentially making Charles indebted. In text segments 17 and 18, Charles' mother is there, with a difficult

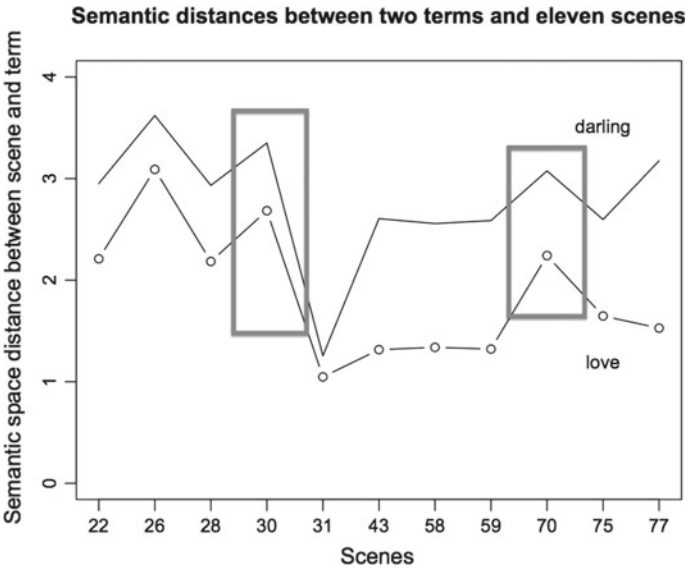


Fig. 1 In the full dimensionality factor space, based on all interrelationships of scenes and words, in the Casablanca movie filmscript. The distance was determined between the word "darling" in this space, with each of the 11 scenes in this space. This was also done for the word "love". The semantic locations of these two words, relative to the semantic locations of scenes 30 and 70 are highlighted with the boxes here

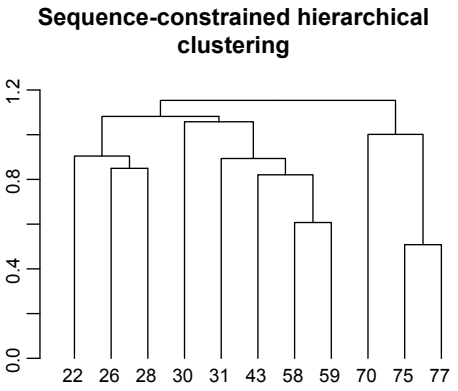


Fig. 2 Hierarchical clustering, that is sequence constrained, of the 11 scenes used, i.e. scenes 22, 26, 28, 30, 31, 43, 58, 59, 70, 75, 77 (all with dialogue, and only dialogue, between Ilsa and Rick). Rather than projections on factors, here the correlations (or cosines of angles with factors) are used to directly capture orientation

mother-in-law relationship for Emma. Plans for running away ensue, with pangs of conscience for Emma, and in the final text segment there is Rodolphe refusing to himself to leave with Emma.

There can be also display of the evolution of sentiment, pursued in [18], expressed by (or proxied by) the terms "kiss", "tenderness", and "happiness". That will permit to be seen that some text segments are more expressive of emotion than are other text segments.

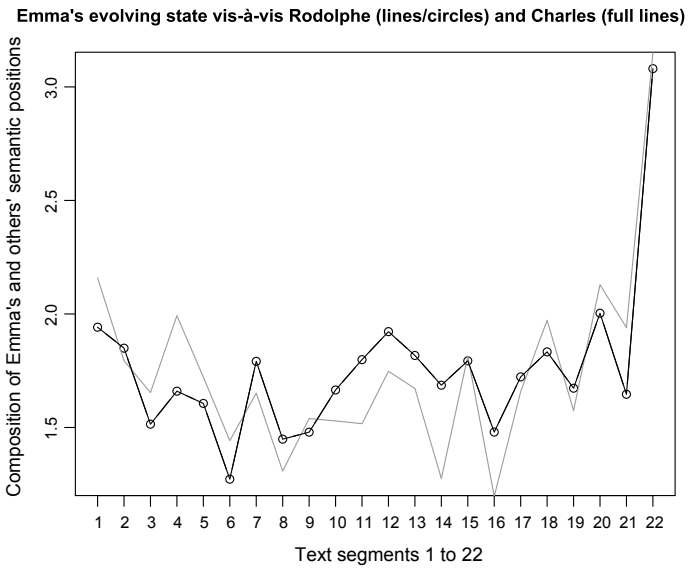


Fig. 3 Gustave Flaubert's Madame Bovary: the relationship of Emma to Rodolphe (more black lines/circles) and to Charles (full line, gray) are mapped out. The text segments encapsulate narrative chronology, that maps approximately into a time axis. Low or small values can be viewed as emotional attachment

7 Analyses of Mapping of Behavioural or Activity Patterns or Trends

This concerns semantic mapping of Twitter data relating to music, film, theatre, etc. festivals. 75 languages were found to be in use, including Japanese, Arabic and so on, with the majority in Roman script. As indicative association to language, because the labelled language may be partially used or not in fact used, we take the following: English, Spanish, French, Japanese, Portuguese. Here, we consider the days 2015-05-11 to 2016-08-02, with two days removed, due to lack of tweets. The numbers of tweets for these languages were as follows (carried out on 11 August 2016): en, 37681771; es, 9984507; fr, 4503113; ja, 2977159; pt, 3270839.

The tweeters and the festivals are as follows. Tweets characterized as French, 4913781 tweets. (For user, date and tweet content, the file size was: 667 MB.) The following were sought in the tweets: Cannes, cannes, CANNES, Avignon, avignon, AVIGNON. Upper and lower case were retained in order to verify semantic proximity of these variants. These related to the Cannes Film Festival, and the Avignon Theatre Festival. The following total numbers of occurrences of these words were found, and the maximum number of occurrences by a user, i.e. by a tweeter: Cannes, 1230559 and 3388; cannes, 145939 and 4024; CANNES, 57763 and 829; Avignon, 272812 and 4238; avignon, 39323 and 2909; AVIGNON, 14647 and 900.

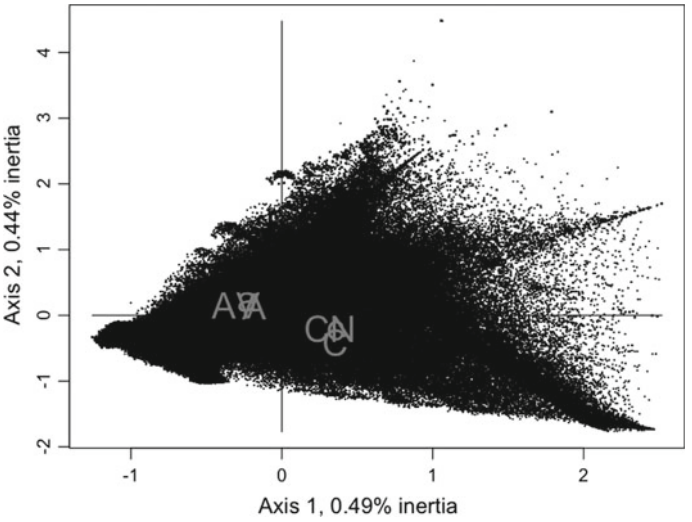


Fig. 4 880664 Twitter tweets projected on the principal factor, i.e. principal axis plane. Attributes here, in gray, are projected

The total number of tweeters, also called users here: 880664; total number of days retained, from 11 May 2015 to 11 Sept. 2016, 481. Cross-tabulated are: 880664 users by 481 days. There are 1230559 retained and recorded tweets. The non-sparsity of this matrix is just: 0.79%.

In Fig. 4, mapped are: C, c, CA (Cannes, cannes, CANNES) and A, a, AV (Avignon, avignon, AVIGNON). They are supplementary variables in the Correspondence Analysis principal factor plane. Semantically they are clustered. They are against the background of the Big Data, here the 880664 tweeters, represented by dots. This demonstrates how semantic association and also disambiguation are achieved in this Correspondence Analysis context, as the data clouds, endowed with the chi squared distance are mapped into the factor space, endowed with the Euclidean distance.

Current considerations, relating to approximately 55 million tweets per year (from May 2015), are as follows. Determine some other, related or otherwise, behavioural patterns that are accessible in the latent semantic, factor space. Retain selected terms from the tweets, and, as supplementary elements, see how they provide more information on patterns and trends. Carry out year by year trend analysis.

For further analyses and description of the data, see [5, 15].

8 Conclusion

Much that is at issue here is close to what is under discussion in [4]. The integral association of methodology and application domain will, of course, have shared and common methodological perspectives. However the application of statistical models, and other analytical stages such as feature selection, data aggregation with the various implications of this, and what is often termed data cleaning or data cleansing, all of these issues require analytical focus, and account to be taken of the analytical context. The latter may well include baselining, or benchmarking in an operational manner. In a sense, we might state that combinatorial inference is so paramount because of its applicability.

A good deal of the case studies reported on here made use of preliminary functionality, now in the R package, `Xplortext`. This package makes use of these R packages, and add greatly to their functionality: `tm`, `FactoMineR`. The software system, `SPAD`, is also extending greatly into support for text processing.

Finally, in finalizing the conclusions, we note again how challenges listed in [6] are being addressed here. For convenience, the focus here is on textual data, so this is being considered, just in this context here, as either administrative (cf. factual and descriptive) and transaction (cf. communication) data. The following citations are from [6].

"*Challenge 10*. Report changes and time series with appropriate measures of uncertainty, so that both the statistical and the substantive significance of changes can be evaluated. The measures of uncertainty should include all sources of uncertainty which can be identified."

"As is well known, observational data present challenges in establishing causality." The issue here is that the data is all that we have, while we must have further, where necessary, interaction with the contextual framework, or basis, of our data. As expressed on page 16, "To establish causality, we need to intervene to break all possible causal links except the link that we wish to test ... The most common way to do this is via a properly controlled experiment involving randomization." All in all here, we are involved with the visualization and verbalization of data.

Section 6. "Combining data from different sources". This involves: data integration, information fusion, and, in the analytics, complement, supplement, add to accuracy. This is so very relevant, for all aspects of validation and verification of the methodology at issue.

References

1. Bécue-Bertaut, M., Kostov, B., Morin, A., Naro, G.: Rhetorical strategy in forensic speeches: multidimensional statistics-based methodology. J. Classif. **31**, 85–106 (2014)
2. Bienaise, S., Le Roux, B.: Combinatorial typicality test in geometric typicality test in geometric data analysis. Stat. Appl. Italian J. Appl. Stat. **29**(2–3), 331–348 (2017)

3. Blasius, J., Greenacre, M. (eds.): Visualization and Verbalization of Data. Chapman and Hall/CRC Press, Boca Raton (2014)
4. Gelman, A., Hennig, C.: Beyond subjective and objective in statistics. J. R. Stat. Soc. Ser. A **180**(Part 4), 1–31 (2017)
5. Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., SanJuan, E.: Overview of the CLEF 2016 cultural micro-blog contextualization workshop. In: Fuhr, N., Quaresma, P., Goncalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, 5–8 September 2016, Proceedings. Lecture Notes in Computer Science, vol. 9822, pp. 371–378 (2016)
6. Hand, D.J.: Statistical challenges of administrative and transaction data. J. R. Stat. Soc. Ser. A **181**(3), 1–24 (2018). Including F. Murtagh comments
7. Hernández, D.M., Bécue-Bertuat, M., Barahona, I.: How scientific literature has been evolving over the time? A novel statistical approach using tracking verbal-based methods. In: JSM Proceedings, 2014, Section on Statistical Learning and Data Mining. American Statistical Association, pp. 1121–1132 (2014)
8. Keiding, N., Louis, T.A.: Perils and potentials of self-selected entry to epidemiological studies and surveys. J. R. Stat. Soc. A **179**(Part 2), 319–376 (2016) Including F. Murtagh comments
9. Legendre, P., Legendre, L.: Numerical Ecology, 3rd edn. Elsevier, Amsterdam (2012)
10. Le Roux, B.: Analyse Géométrique des Données Multidimensionelles. Dunod, Paris (2014)
11. Le Roux, B., Lebaron, F.: Idées-clefs de l'analyse géometrique des données (Key ideas in the geometric analysis of data). In: Lebaron, F., Le Roux, B. (eds.) La Méthodologie de Pierre Bourdieu en Action: Espace Culturel, Espace Social et Analyse des Données, pp. 3–20. Dunod, Paris (2015)
12. Le Roux, B., Rouanet, H.: Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis. Kluwer, Dordrecht (2004)
13. McKee, R.: Story: Substance, Structure, Style, and the Principles of Screenwriting. Methuen, London (1999)
14. Murtagh, F.: Multidimensional Clustering Algorithms. Physica-Verlag, Würzburg (1985)
15. Murtagh, F.: Semantic mapping: towards contextual and trend analysis of behaviours and practices. In: Balog, K., Cappellato, L., Ferro, N., MacDonald, C. (eds.) Working Notes of CLEF 2016 – Conference and Labs of the Evaluation Forum, Évora, Portugal, 5–8 September 2016, pp. 1207–1225 (2016). http://ceur-ws.org/Vol-1609/16091207.pdf
16. Murtagh, F.: Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics. Chapman and Hall, CRC Press, Boca Raton (2017)
17. Murtagh, F., Farid, M.: Contextualizing geometric data analysis and related data analytics: a virtual microscope for big data analytics. J. Interdiscip. Methodol. Issues Sci. **3** (2017). arXiv:1611.09948v3
18. Murtagh, F., Ganz, A.: Pattern recognition in narrative: tracking emotional expression in context. J. Data Min. Digit. Humanit. **2015** (2015)
19. Murtagh, F., Ganz, A., McKie, S.: The structure of narrative: the case of film scripts. Pattern Recognit. **42**, 302–312 (2009)
20. Murtagh, F., Spagat, M., Restrepo, J.A.: Ultrametric wavelet regression of multivariate time series: application to Colombian conflict analysis. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **41**, 254–263 (2011)
21. Murtagh, F., Pianosi, M., Bull, R.: Semantic mapping of discourse and activity, using Habermas's theory of communicative action to analyze process. Qual. Quant. **50**(4), 1675–1694 (2016)
22. Murtagh, F., Orlov, M., Mirkin, B.: Qualitative judgement of research impact: domain taxonomy as a fundamental framework for judgement of the quality of research. J. Classif. **35**(1), 5–28 (2018)

A Three-Way Data Analysis Approach for Analyzing Multiplex Networks

Giancarlo Ragozini, Maria Prosperina Vitale and Giuseppe Giordano

Abstract In the present contribution, the use of factorial methods for three-way data is proposed to visually explore the structure of multiplex networks, that is, in presence of more relationships measured for a common set of nodes. Specifically, the DISTATIS technique, an extension of multidimensional scaling to three-way data, is used to analyze multiplex one-mode networks. In this procedure different types of relationships are represented in separate spaces and in a compromise space. A well-known dataset in the related literature is considered to illustrate how this procedure works in practice.

Keywords Social network analysis · Factorial methods · DISTATIS · Lazega lawyers network data

1 Introduction

Multilayer networks [11, 17] arise when there exist two or more relationships for a common set or different sets of nodes. A multiplex network is a special case of a multilayer network that consists of a fixed set of nodes that interact through different relationships. For instance, in social sciences the researcher can collect network data on friendship, neighbors, kinship, trust, and advice relationships among the same set of actors.

For this kind of network data structure, usually the proposed approaches consist of dealing with multiple relationships separately or flattening the information embedded

G. Ragozini
Department of Political Science, University of Naples Federico II, Napoli, Italy
e-mail: giragoz@unisa.it

M. P. Vitale (✉) · G. Giordano
Department of Political and Social Studies, University of Salerno, Fisciano, Italy
e-mail: mvitale@unisa.it

G. Giordano
e-mail: ggiordan@unisa.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_5

in all layers. This latter procedure reduces the complexity of multiplex data and may lead to a loss of relevant information. To cope with this issue, it could be useful to adapt multivariate factorial methods to multiplex network data. In this regard, factorial methods have been proposed in the network analysis framework to explore different network structures [12, 14, 21], including attributes of nodes and events in two-mode networks [15], or to analyze network-derived measures [19]. In the case of multiplex networks, several methods have been proposed. More specifically, canonical correlation analysis was adopted to identify dimensions along which two networks are related to each other [8], and analytical procedures were recently introduced for dimension reduction using correspondence analysis [28] or cluster analysis [26].

Within this framework, the present contribution aims at extending the use of the DISTATIS approach [2] to visually explore the hidden structure of multiplex networks preserving the inherent complexity. The proposed method aims at analytically and visually exploring (*i*) the network structure in terms of nodes' similarity in each single layer, (*ii*) the common structure of all layers, (*iii*) the nodes' variation across layers, and (*iv*) the similarity among the structure of the layers.

The contribution is organized as follows. In Sect. 2, a brief review of multiplex networks is presented. Section 3 describes in detail the analytic procedure for handling multiplex network data using the DISTATIS approach. Section 4 discusses the main results of the proposed procedure for the analysis of a set of relationships among lawyers described in Lazega [18]. Section 5 concludes with suggestions for future lines of research.

2 Multiplex Networks: A Brief Review

Multilayer networks are based on multiple kinds of relationships among either a different sets of nodes or among a unique set of nodes, as in multiplex networks. More formally, a multilayer network $\mathscr{M}$ is a pair $(\mathscr{G}, \mathscr{E})$, with $\mathscr{G} = \{G_k\}_{k=1,\ldots,K}$, the collection of K networks; $\mathscr{E} = \{E_{kk'}\}_{(k,k'=1,\ldots,K)}$, the collections of intra-layer $(k = k')$ and inter-layer $(k \neq k')$ edges.

Note that in the case of pure multiplex networks, such as the case under analysis, the set of nodes is fixed, that is, $V_1 = V_2 = \cdots = V_K = V$, and the inter-layer edges are constant and indicate only that the nodes are present in the different layers [17]. In each layer, $G_k = (V, E_{kk})$, with $V = (v_1, \ldots, v_n)$ the set of n nodes of each network, and $E_{kk} \subseteq V \times V$ the set of edges. For $k = 1, \ldots, K$, let us consider from the network $G_k \in \mathscr{G}$ the corresponding adjacency matrix $\mathbf{A}_k = (a_{ijk})$, with $a_{ijk} = 1$ if $(v_i, v_j) \in E_{kk}$, and $a_{ijk} = 0$ otherwise.

The methodological approaches developed to handle multiplex social networks are mainly based on exponential random graph models [20], blockmodeling analysis and relational algebras [27], and factorial methods [8, 26, 28]. In addition, empirical studies on real-world multiplex network data have been proposed in several sci-

entific fields [3, 7, 25], presenting sophisticated network visual analytics [23] and developments in software tools [9].

Here, to handle these kinds of complex network systems the adaptation of factorial methods designed for a three-way data structure is considered. The K adjacency matrices give rise to a three-way matrix $\mathbb{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_K)$ that can be analyzed using the DISTATIS technique [1, 2].

3 Using the DISTATIS Procedure for Multiplex Network Data

Several statistical methods have been presented to deal with three-way data [16]. In the present contribution, starting from the characteristics of the adjacency matrices describing a binary one-mode multiplex network, the DISTATIS method is adopted. It represents a generalization of multidimensional scaling (MDS) in the STATIS approach [13] introduced to analyze a set of distance matrices. The method analyzes the network structure embedded in each layer, as well as the global structure derived as a linear combination of the layers by considering data-driven weights. Therefore, this method provides a rich set of analytical and graphical results in which the different relationships can be considered as facets of a common underlying relational structure (corresponding to the *compromise* space).

To illustrate how DISTATIS could be adapted and applied to the analysis of multiplex networks, three points must be taken into account: *(i)* how to derive a three-way distance matrix from a multiplex network, *(ii)* how to apply this procedure to the derived three-way distance matrix, and *(iii)* how to interpret the results and the factorial graphical representations in terms of network structures.

First, a three-way distance matrix $\mathbb{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_K)$ is derived from the multiplex adjacency matrix $\mathbb{A}$, with $d_{ijk} = geo_k(v_i, v_j)$ being the geodesic distance[1] between the nodes v_i and v_j in the layer k. If two nodes cannot reach each other, i.e. one, or both, are isolated, or they belong to different unconnected subgraphs, the geodesic distance is not defined. In such a case, the distance can be set as $d_{ijk} = p * \max[geo_k(v_i, v_j)]$ in the kth layer, where p is a suitable constant.

Second, from the matrix $\mathbb{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_K)$, the DISTATIS algorithm can be synthesized through the following steps:

(1) Compute for each layer $k = 1, \ldots, K$ the cross-product matrices $\tilde{\mathbf{S}}_k$ according to the so-called double-centering transformation of MDS, as follows: $\tilde{\mathbf{S}}_k = -\frac{1}{2}\mathbf{C}\mathbf{D}_k\mathbf{C}^T$, where $\mathbf{C}$ is the centering operator; $\mathbf{C} = \mathbf{I} - \mathbf{1}\mathbf{n}^T$, where $\mathbf{I}$ is an

[1] The geodesic distance between pairs of nodes in a network is a graph-theoretic distance and consists of the shortest path between the given nodes. To evaluate the geodesic distances, we use the modified breadth-first search algorithm by Brandes [6] implemented in the R package *sna* [5]. An alternative graph-theoretic distance measure can be find in Cohen [10], and alternative distance/dissimilarity measures for binary data can be found in Batagelj [4].

n-dimensional identity matrix, $\mathbf{1}$ is an n-dimensional unit vector, and $\mathbf{n}$ is an n-dimensional vector with elements equal to $1/n$, that is, the mass of each node;

(2) Compute the normalized cross-product matrix $\mathbf{S}_k$ dividing $\tilde{\mathbf{S}}_k$ by its first eigenvalue;

(3) Evaluate the similarity matrix among layers $\mathbf{H} = \{h_{k,k'}\}$ by using the *RV* coefficient [22] computed on the normalized cross-product matrices $\mathbf{S}_k$, i.e., $h_{k,k'} = \frac{\mathbf{s}_k^T \mathbf{s}_{k'}}{||\mathbf{s}_k|| ||\mathbf{s}_{k'}||}, \quad k, k' = 1, \ldots, K$ with $\mathbf{s}_k$ the vectorization of $\mathbf{S}_k$. This matrix is indicated as the *RV* matrix;

(4) Compute the eigenvalues and the eigenvectors of the matrix $\mathbf{H}$; that is, $\mathbf{H} = \mathbf{P}^T \mathbf{\Lambda} \mathbf{P}$. The first two columns of the matrix $\mathbf{P}\mathbf{\Lambda}^{1/2}$ are the coordinates of a factorial map representing the similarities among the different kinds of relationships building the multiplex network, each one represented as a point in the factorial plan. With this map, the similarity among the K layers can globally evaluated. The first eigenvector $\mathbf{p}_1$ is used to determine a set of weights α_k used in the next step to compute the compromise: the α's that reflect the similarities among the normalized cross-product matrices and are defined as $\alpha_k = \frac{p_{k1}}{||\mathbf{p}_1||}$, and p_{k1} the coordinate of the k layer on it. Finally, a measure of the quality of the compromise τ can be derived by the first eigenvalue λ_1 over the trace of $\mathbf{\Lambda}$; i.e., $\tau = \frac{\lambda_1}{\sum_{k=1}^{K} \lambda_k}$;

(5) Compute the *compromise* matrix $\mathbf{S}$ as the weighted sum of the normalized cross-product matrices, $\mathbf{S} = \sum_{k=1}^{K} \alpha_k \mathbf{S}_k$;

(6) Perform the eigenvalue decomposition of the compromise matrix $\mathbf{S}, \mathbf{S} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$, to obtain the factorial coordinates for plotting the nodes in the common space $\mathbf{F} = \mathbf{V}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{S}\mathbf{V}\mathbf{\Sigma}^{-\frac{1}{2}}$. Furthermore, the coordinates can be used to compute the contribution of the i-th each node on l-th factorial axes, as follows: $crt_{il} = \frac{f_{il}^2}{\lambda_l}$;

(7) Represent the cross-product matrices $\mathbf{S}_k$ in the space of the compromise by projecting the matrices as supplementary points. The coordinates can be easily computed as $\mathbf{F}_k = \mathbf{S}_k \mathbf{V}\mathbf{\Sigma}^{-\frac{1}{2}}$, and are called partial scores. They represent the position of each node in each layer, and all the coordinates have a common reference space.

The derived compromise matrix represents a weighted average of the distance matrices using a double system of weights. The first eigenvalues of the cross-product matrices express the relative importance of the layers in terms of their inertia; whereas the α_k coefficients measure such importance with reference to the similarity among the layers.

Given the analytical results above, four factorial maps are defined to graphically analyze different aspects of the multiplex data structure. In the first one, it is possible to represent each layer as a point on a two-dimensional map obtained by the eigenvectors of the similarity matrix between the cross-product matrices (*between-layers map*). Here, if two points are close, it implies that the global relational patterns of the corresponding layers are similar. In the second one (*compromise score map*), each node is represented as a point by using the first coordinates obtained decomposing the compromise matrix. In this map, if two points are close, the corresponding nodes have similar relational patterns in almost all layers. By considering the partial scores,

all the nodes can be represented in separate factorial maps (one for each layer). Here, the relative position of the points shows the similarities of the relational pattern of the corresponding nodes, layer by layer (*partial score maps*). The compromise scores and the partial scores can be also used to obtain a joint representation (*joint map*) in which each node is represented by $K + 1$ points, one for each layer plus one for the compromise. The points of the compromise are the barycenters of the points representing the layers, as the compromise is a weighted average of the layers. Connecting the compromise point to the layer points, a star shape describes each actor. The size and the shape of the stars provide information about the variability of the relational patterns over the layers actor by actor. The joint map allows us to appreciate the variability of the relational patterns, node by node, in the different layers.

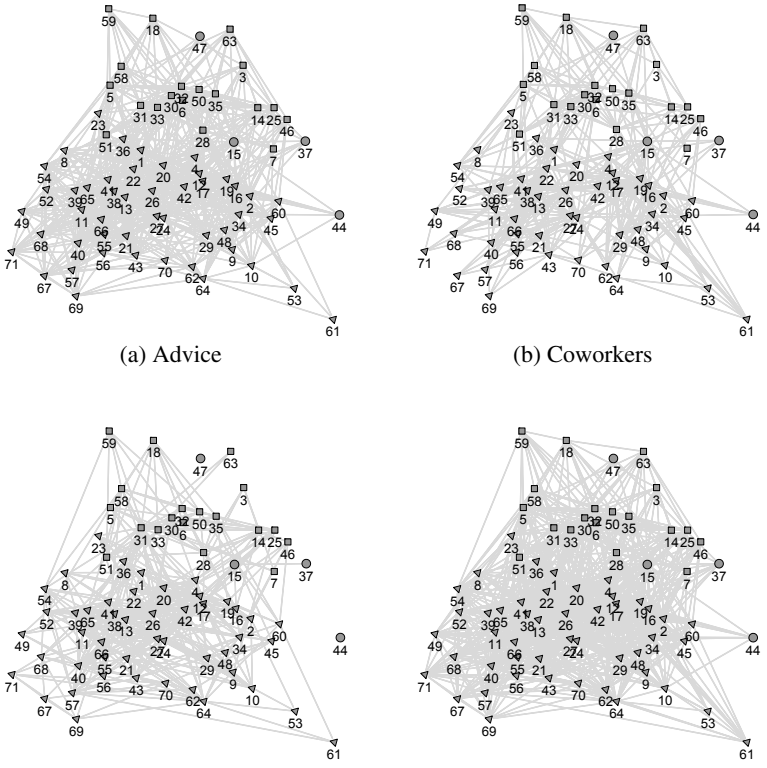
4 The Lazega Lawyers Data

Many multiplex networks are available in online archives[2] and can be used to demonstrate the advantages of the proposed method.

Some examples presented in the network literature [24] are based on relations between lawyers in a corporate law firm in New England collected by Lazega [18]. Here, this data set is specifically used to illustrate the usefulness of the DISTATIS approach for the treatment of multiplex networks. The information refers to *advice*, *coworkers*, and *friendship* networks between 71 lawyers (partners and associates). Various members' attributes are also part of the dataset. Among others, the affiliation with the *office* in which they work (Boston, Hartford, and Providence) is taken into account in the following.

Spring-embedding graph representations of the three networks and the flattened multiplex network are shown in Fig. 1. The picture displays networks enhanced by using different symbols for each actor-lawyer according to the *office* in which the lawyers work. The three networks are quite dense with few disconnected nodes (8, 44, 47) for the two relationships coworkers and friendship, respectively. For an in-depth description of the networks' characteristics, readers should refer to Lazega [18]. The graphs in Fig. 1 appear a quite dense reproducing the so-called "hair ball effect". When the office affiliation is considered, the presence of two groups in the networks' structure appears.

[2]For instance, see the Manlio De Domenico homepage with datasets released for reproducibility: https://comunelab.fbk.eu/data.php.



(a) Advice (b) Coworkers

(c) Friendship (d) Flattened Multiplex

Fig. 1 Lazega lawyers data. Spring-embedding graph representations: **a** *Advice* network; **b** *Coworkers* network; **c** *Friendship* network; **d** Flattened multiplex network. Node symbols are set according to the office in which the lawyers work (Triangle = Boston; Square = Hartford; Circle = Providence)

4.1 The Lazega Data with DISTATIS: Analytical Results and Interpretation

According to the underlying social processes described in Lazega [18], such as bounded solidarity, lateral control, quality control, knowledge sharing, balancing powers, and regulation among lawyers, the DISTATIS procedure is performed on the three symmetrized adjacency matrices. Even if each singleton dimension of collaboration among the lawyers is observed, with the proposed approach a unifying dimension of the underpinning concept as a whole is derived. At the same time, every dimension (layer) gives information about local phenomena that can be analyzed and described in terms of an actor's position in the network.

Some derived measures of the procedure are summarized in Table 1. First, the RV's coefficient matrix is displayed, showing that the two layers *advice* and *coworkers* present a higher degree of similarity, as well as the the two layers *advice* and

Table 1 DISTATIS analytic results derived from the matrix **H**: RV's coefficients matrix among layers; factors' scores, eigenvalues, relative (τ), and cumulative percentage of explained inertia, eigenvectors and α weights

RV matrix	*Advice*	*Coworkers*	*Friendship*
Advice	1.00	0.54	0.41
Coworkers	0.54	1.00	0.29
Friendship	0.41	0.29	1.00
Factor scores	dim 1	dim 2	dim 3
Advice	0.85	−0.15	0.51
Coworkers	0.79	−0.46	−0.41
Friendship	0.70	0.70	−0.16
Eigenvalues			
λ_k	1.83	0.72	0.45
τ_k	61.00	24.00	15.00
Cum.	61.00	85.00	100.00
Eigenvectors			
Advice	0.63	−0.18	0.76
Coworkers	0.59	−0.54	−0.61
Friendship	0.51	0.83	−0.23
	Advice	*Coworkers*	*Friendship*
α **weights**	0.36	0.34	0.30

friendship show a certain degree of similarity. The *friendship* and *coworkers* layers present a low degree of similarity. Then, we can conclude that the *advice* relationship resembles both a formal relationship (*coworkers*) and an informal relationship (*friendship*). These coefficients highlight the presence of "bounded solidarity" as discussed in Lazega [18].

The factorial map in Fig. 2 graphically displays the factor scores in Table 1. The plot shows the role played by each layer in determining the final compromise space. Whereas every layer has an important role in weighting the final configuration (given by the high coordinates on the first axis), the second axis reveals the real shape of the configuration. On the top, the *friendship* layer is separated by the two formal contacts of the *advice* and *coworkers* layers. The relative position of each layer determines the different weight and role they play in building the compromise space.

Table 1 reports the eigenvalues and the corresponding percentage of relative inertia (τ values) along with their cumulative percentages. The quality of compromise is quite satisfactory: the first τ is 61%, while the first two dimensions account for 85% of the dissimilarity of the inter-layers. The eigenvectors and the α weights indicate the quite similar importance of the three layers in defining the compromise matrix.

Based on the representation of the lawyers in the DISTATIS compromise factorial plan (Fig. 3), two groups are clearly separated. They are consistent with the office in which they work: the lawyers from Boston and Providence mixed up compared to

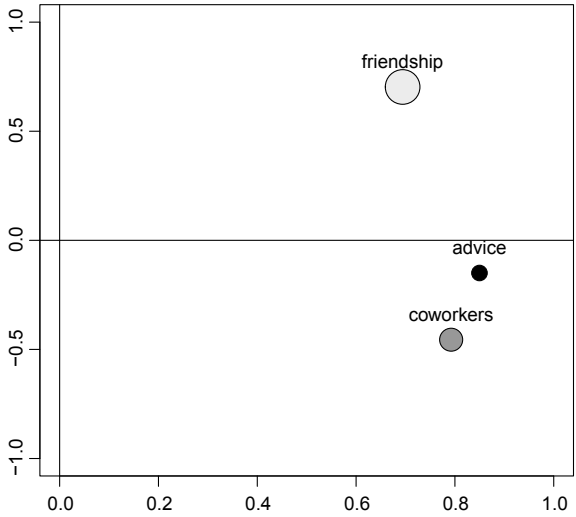


Fig. 2 Lazega lawyers data. Between-layers map. Layer size is proportional to its contribution. Layers are colored in grayscale

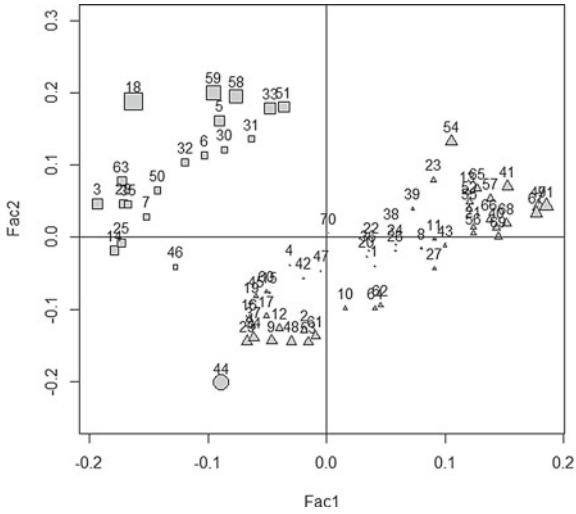
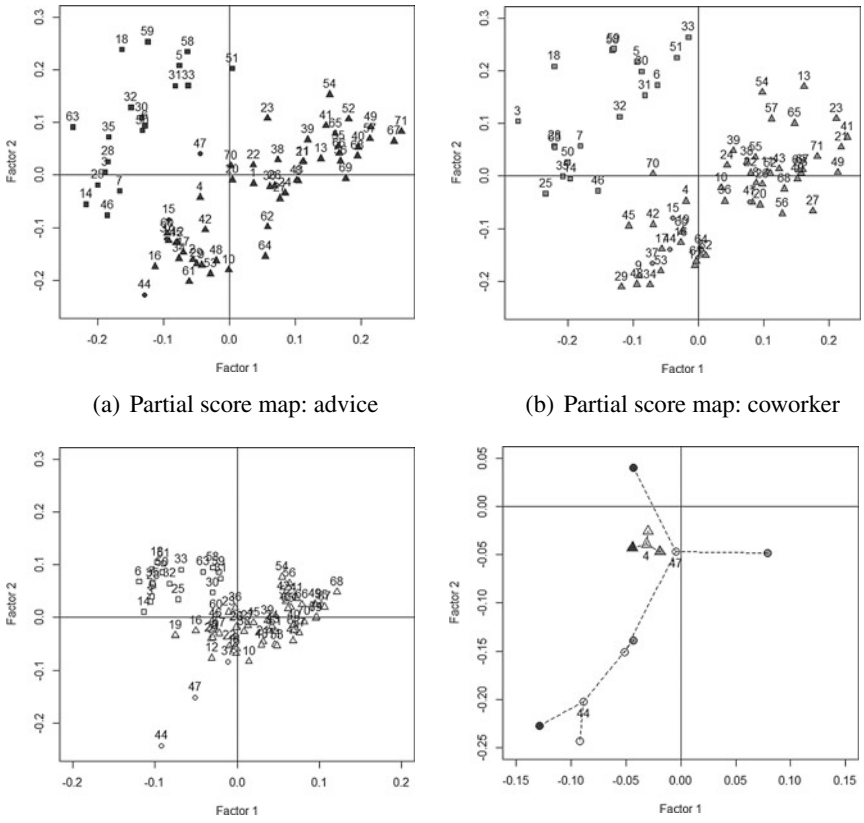


Fig. 3 Lazega lawyers data. Compromise score map: points represent lawyers in the compromise space. Node symbols are set according to the office in which the lawyers work (Triangle = Boston; Square = Hartford; Circle = Providence); point size is proportional to its contribution

those affiliated with Hartford office. In addition, the two main groups can be divided into subgroups.

Finally, the positions of the lawyers in the three layers in the common reference space given by the compromise are reported in Fig. 4a–c. The structures of the first two layers are very similar, while the *friendship* layer appears a little bit different. It is also possible to analyze how a single lawyer changes his/her position in the joint representation appreciating the variability of its relational patterns in the different layers. In Fig. 4d, three lawyers (4, 44, and 47) are represented as an example. Each actor is represented by four points, one for each layer plus one for the compromise.



(a) Partial score map: advice

(b) Partial score map: coworker

(c) Partial score map: friendship

(d) Joint map for actors 4, 44, and 47

Fig. 4 Lazega lawyers data. **a** Partial score maps for the advice layer; **b** Partial score maps for the coworker layer; **c** Partial score maps for the friendship layer; **d** Joint map of actors 4, 44, and 47. Node symbols are set according to the office in which the lawyers work (Triangle = Boston; Square = Hartford; Circle = Providence), points representing the partial scores are colored in grayscale according to the three layers; and points representing the compromise scores—at centers of the stars—are white colored

Looking at our example, 44 and 47 show high variability, while actor 4 presents very low variability. Actor 47 is a young associate man in the Providence office who has worked in litigation practice for three years, perfectly integrated in a subgroup of coworkers in Boston. He plays a central bridging position in the *advice* network in connecting the two groups but is isolated in the *friendship* network (the corresponding point is pulled out from the center of map). Actor 44 is an older female associate in the Providence office who has worked in corporate practice for five years, with high layers' similarities for *advice* and *friendship*, being quite isolated. However she is very well connected in the *coworkers* network (her point is in a more central position). In contrast, Actor 4 (a 59-years-old man from the Boston office, who has

31 years of work experience in litigation practice) has strong connections in all the three layers. Thus, his star is quite small denoting a stable relational pattern in formal and informal relationships with his colleagues.

5 Concluding Remarks

In the present contribution, the use of the DISTATIS approach to treat multiplex network data is considered. As general findings, the procedure analyzed the network structures embedded in one-mode multiplex networks, the similarities among actors in the compromise space, and the global similarities among the layers. As DISTATIS is an extension of MDS to three-way data matrix, the graphical and analytical results are easily interpretable by social network analysts which are already used to MDS spring-embedding visualizations.

The results of the illustrative example showed the high explicative power of the proposed analytic procedure in capturing similarities among layers. The possibility of measuring the inter-dissimilarity between layers allows the definition of a suitable subspace where comparisons at the layer and node levels can be made.

These findings suggest new lines of research in performing a simulation study to assess the stability of the procedure and in deriving measures for multiplex networks. Moreover, as network data allows for several ways of computing distances, a comparison of how different distance measures affect the results and the visualization of compromise space should also be addressed. The analyzed real-world example considers binary one-mode networks, and the attribute data, such as the office affiliation. These latter has been here only exploited in the graphical representations. The inclusion of attribute data in the analytical procedure will be considered as future work, following the approach in Giordano and Vitale [15].

References

1. Abdi, H., O' Toole, A. J., Valentin, D., Edelman, B.: DISTATIS: The analysis of multiple distance matrices. In: Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition, San Diego (CA) US, pp. 42–47 (2005)
2. Abdi, H., Valentin, D., Chollet, S., Chrea, C.: Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. Food Qual. Prefer. **18**, 627–640 (2007)
3. Barbillon, P., Donnet, S., Lazega, E., Bar-Hen, A.: Stochastic block-models for multiplex networks: an application to a multilevel network of researchers. J. R. Stat. Soc. Ser. A. **180**, 295–314 (2016)
4. Batagelj, V., Bren, M.: Comparing resemblance measures. J. Classif. **12**, 73–90 (1995)
5. Butts, C.T.: Social network analysis with SNA. J. Stat. Softw. **24**, 1–51 (2008)
6. Brandes, U.: Faster Evaluation of Shortest-Path Based Centrality Indices. Konstanzer Schriften in Mathematik und Informatik **120** (2000)
7. Bródka, P., Chmiel, A., Magnani, M., Ragozini, G.: Quantifying layer similarity in multiplex networks: a systematic study. Roy. Soc. Open Sci. **5**, 171747 (2018)

8. Carroll, C.: Canonical correlation analysis: assessing links between multiplex networks. Soc. Netw. **28**, 310–330 (2006)
9. De Domenico, M., Porter, M.A., Arenas, A.: MuxViz: a tool for multilayer analysis and visualization of networks. J. Compl. Netw. **3**, 159–176 (2015)
10. Cohen, J.D.: Drawing graphs to convey proximity: an incremental arrangement method. ACM T. Comput-Hum. Int. **4**, 197–229 (1997)
11. Dickison, M.E., Magnani, M., Rossi, L.: Multilayer Social Networks. Cambridge University Press, Cambridge (2016)
12. D'Esposito, M.R., De Stefano, D., Ragozini, G.: On the use of multiple correspondence analysis to visually explore affiliation networks. Soc. Netw. **38**, 28–40 (2014)
13. Escoufier, Y.: Objectifs et procedures de l'analyse conjointe de plusieurs tableaux de donnees. Statistique et analyse des donnees **10**, 1–10 (1985)
14. Faust, K.: Using correspondence analysis for joint displays of affiliation networks. In: Carrington, P., Scott, J., Wasserman, S. (eds.) Models and Methods in Social Network Analysis, pp. 117–147. Cambridge University Press, New York (2005)
15. Giordano, G., Vitale, M.P.: On the use of external information in social network analysis. Adv. Data Anal. Classif. **5**, 95–112 (2011)
16. Kiers, H.A., Mechelen, I.V.: three-way component analysis: principles and illustrative application. Psychol. Methods **6**, 84–110 (2001)
17. Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. J. Compl. Netw. **2**, 203–271 (2014)
18. Lazega, E.: The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership. Oxford University Press, Oxford (2001)
19. Liberati, C., Zappa, P.: Dynamic patterns analysis meets Social Network Analysis in the modeling of financial market behavior. In: Proceedings 59th ISI World Statistics Congress, 25-30 August 2013, Hong Kong, pp. 2447–2452 (2013)
20. Pattison, P., Wasserman, S.: Logit models and logistic regressions for social networks: II. Multivariate relations. Brit. J. Math. Stat. Psychol. **52**, 169–193 (1999)
21. Ragozini, G., De Stefano, D., D'Esposito, M.R.: Multiple factor analysis for time-varying two-mode networks. Netw. Sci. **3**, 18–36 (2015)
22. Robert, P., Escoufier, Y.: A unifying tool for linear multivariate statistical methods: the RV-coefficient. J. R. Stat. Soc. C-Appl. **25**, 257–265 (1976)
23. Rossi, L., Magnani, M.: Towards effective visual analytics on multiplex and multilayer networks. Chaos Soliton. Frac. **72**, 68–76 (2015)
24. Snijders, T.A.B., Pattison, P.E., Robins, G.L., Handcock, M.S.: New specifications for exponential random graph models. Sociol. Methodol. **36**, 99–153 (2006)
25. Snijders, T.A.B., Lomi, A., Torl, V.J.: A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. Soc. Netw. **35**, 265–276 (2013)
26. Voros, A., Snijders, T.A.B.: Cluster analysis of multiplex networks: defining composite network measures. Soc. Netw. **49**, 93–112 (2017)
27. White, H.C., Boorman, S.A., Breiger, R.L. Social structure from multiple networks. I. Blockmodels of roles and positions. Am. J. Sociol. **81**, 730–780 (1976)
28. Zhu, M., Kuskova, V., Wasserman, S., Contractor, N.: Correspondence analysis of multirelational multilevel networks. In: Lazega, E., Snijders, T.A.B. (eds.) Multilevel Network Analysis for the Social Sciences, pp. 145–172. Springer International Publishing, Switzerland (2016)

Comparing FPCA Based on Conditional Quantile Functions and FPCA Based on Conditional Mean Function

Mariantonietta Ruggieri, Francesca Di Salvo and Antonella Plaia

Abstract In this work functional principal component analysis (FPCA) based on quantile functions is proposed as an alternative to the classical approach, based on the functional mean. Quantile regression characterizes the conditional distribution of a response variable and, in particular, some features like the tails behavior; smoothing splines have also been usefully applied to quantile regression to allow for a more flexible modelling. This framework finds application in contexts involving multiple high frequency time series, for which the functional data analysis (FDA) approach is a natural choice. Quantile regression is then extended to the estimation of functional quantiles and our proposal explores the performance of the three-mode FPCA as a tool for summarizing information when functional quantiles of different order are simultaneously considered. The methodology is illustrated and compared with the functional mean based FPCA through an application to air pollution data.

Keywords FPCA · Conditional quantile functions · Conditional mean function

1 Introduction

In this paper we model data moving from mean functions towards quantile functions. The idea underlying this proposal depends on the type of data we deal with, that is air pollution data, presenting peaks and high variability, thereby our attention is focused on particular features of their distribution, like the tails behavior.

The aim of this work is to provide an appropriate synthesis of the estimated functional quantiles by means of a small number of principal components, exploiting simultaneously the information given by functional quantiles of different order. This idea is suggested by a previous work [15], where a three-mode FPCA was proposed in

M. Ruggieri (✉) · A. Plaia
Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy
e-mail: mariantonietta.ruggieri@unipa.it

F. Di Salvo
Department of Agricultural Sciences, Food and Forestry, University of Palermo, Palermo, Italy
e-mail: francesca.disalvo@unipa.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_6

order to take into account the whole variability, exploiting the functional covariances; moreover, the increasing demand of statistical tools for FDA encourages us to extend the idea to the quantile functions for infinite-dimensional data.

One of the earlier paper studying functional quantile regression is [11]; starting with a linear quantile regression, in which the response is a scalar while the covariate is a function, and expanding the covariate and the slope function in terms of their principal components (PCs), the model is transformed into a quantile regression model with an infinite number of regressors.

First results concerning the conditional quantile estimation, adapted to functional data, were obtained by [1] by using a B-spline approach for the representation of the response. A deep study of nonparametric kernel estimator of the conditional quantile is in [7]; it has been adapted to the functional context in [5], where a prove of the almost complete consistency of the estimator is presented. References [2, 6] address asymptotic properties and asymptotic distribution of the functional kernel regression estimate; [10] introduces the class of the L_1 local linear estimator of the quantile regression function as a generalization of the local constant (kernel) estimator.

Generalized quantile regression in [9] refers to a broad family, including conditional quantiles as special cases, concerning the conditional distribution of a response, given a set of explanatory variables. This family collects many contributions proposed in literature, motivated by applications sharing high variability and extreme fluctuations; in order to overcome the problem of insufficiency of data, despite their high variability at the tails of the distributions, [9] proposed a functional data analysis approach to obtain efficient estimation of regression quantiles with fixed order, using multiple data set: conditional quantiles, given a probability level, are considered as functions of time and are estimated non-parametrically on the basis of multiple data set, assuming the Karhunen-Loéve expansion for each quantile, as the sum of an overall mean function plus a linear combination of principal functions.

In this paper, the proposed method aims to the simultaneous estimation of a collection of quantiles at different probability levels, using the method of penalized splines [4, 17], and then reduces the dimensionality by mean of a functional principal component analysis. In the following, we describe the main characteristics of the FDA and the quantile regression approach in Sect. 2; in Sect. 3 we present the proposed methodology. After introducing the data set in Sect. 4, in Sect. 5 we compare our proposal with the traditional procedure. Finally, we show the obtained results and draw some conclusions in Sects. 6 and 7.

2 The Functional Quantile Regression Approach

2.1 The FDA Approach: Smoothing in Time

FDA approach is suggested by the functional structure of our data: air pollution data, even if recorded at discrete times, can be considered as realizations of random curves. According to FDA, observed data are affected by errors and the functional

data are estimated through a P-spline smoothing model, expressed in terms of a linear combination of basis functions, spanning the time interval, and coefficients, capturing the temporal dynamics [13, 14].

In this context, the minimization of the penalized residual sum of squares, with a penalty matrix expressed in terms of second-order derivatives, corresponds to estimate the mean of the conditional distributions at each time point. On the basis of this estimated functions, the principal component analysis is the most used tool in order to reduce the dimensionality and synthesize the variability.

An extensive literature reports that functional principal component analysis is applied for different purposes [3, 13]. In the next paragraph, we generalize this basic idea involving other features of the functional variability.

2.2 The Functional Quantile Approach

As an alternative to the classical approach, that models the mean of the conditional distribution of the response variable, the quantile regression approach [12] models the conditional quantiles, providing information on particular features of the conditional distributions, for example the tail behavior. Moreover, it allows to deal with data in presence of model mis-specification.

As known, given a real valued random variable X, the αth ($0 < \alpha < 1$) quantile function is essentially defined as the inverse of its cumulative distribution function F [8, 9]:

$$Q_X(\alpha) = F_X^{-1}(\alpha) = inf\{x \in R : F(x) \geq \alpha\}; \tag{1}$$

one prominent quantile value is the median of X ($\alpha = 0.5$).

When a vector of covariates is associated to X, the interest could be in studying the conditional (or regression) quantile as a function of these covariates. In particular, if the covariate is time, the αth quantile is function of t:

$$Q_{X|T}(\alpha|t) = F_{X|t}^{-1}(\alpha) = l_\alpha(t). \tag{2}$$

Here we are interested in a collection of regression quantiles, at different probability levels, from a collection of multiple time series; Cross-validation and generalized cross-validation are adapted to select a common smoothing parameter for all sample curves with the roughness penalty approaches; in a FDA framework a representation in terms of linear combination of smooth functions is considered for the functional quantile at each unit i, with $i = 1, \ldots, N$:

$$l_{\alpha,i}(t) = \sum_{k=1}^{K} \theta_{k,i}^{\alpha} \phi_k(t) = \Phi(t)^T \theta_i^{\alpha}, \tag{3}$$

where:

- $\Phi(t) = [\phi_1(t), \ldots, \phi_K(t)]^T$ is a K-vector of B-spline basis functions;
- $\theta_i^{\alpha} = \left[\theta_{i,1}^{\alpha}, \ldots, \theta_{i,K}^{\alpha}\right]^T$ is a K-vector of coefficients.

In order to estimate the coefficients of B-splines, separately for each probability level α, an expected loss function is minimized:

$$L(\Theta^{\alpha}) = E\left[\rho_{\alpha}\left(X - \Phi\Theta^{\alpha}\right)\right], \tag{4}$$

where $\rho_{\alpha}(.)$ is an asymmetric loss function and Θ^{α} is the $K \times N$ matrix of parameters to be estimated; in this paper, we focus our attention on the loss function defined as the weighted sum of absolute residuals:

$$\rho_{\alpha}\left(X - \Phi\Theta^{\alpha}\right) = \mathbf{w}_{(\alpha)}|X - \Phi\Theta^{\alpha}|. \tag{5}$$

The elements of the vector of weights, $\mathbf{w}_{(\alpha)} = \left[w_{1_{(\alpha)}}, \ldots, w_{N_{(\alpha)}}\right)\Big]^T$, are defined as follows:

- $w_{i_{(\alpha)}} = \alpha, \qquad$ if $X_i > \alpha\Phi_i\Theta^{\alpha}$;
- $w_{i_{(\alpha)}} = 1 - \alpha, \qquad$ if $X_i \leq \alpha\Phi_i\Theta^{\alpha}$.

Using the method of penalized splines, a penalized average empirical loss is also considered in [9], where the introduced penalty term penalizes the roughness of the fitted quantile function:

$$L(\Theta^{\alpha}) = E\left[\rho_{\alpha}\left(X - \Phi\Theta^{\alpha}\right)\right] + \lambda\Theta^{\alpha}\mathbf{H}\Theta^{\alpha}. \tag{6}$$

The estimation of quantile regression is well described in [9, 12].

3 The Proposed Procedure: Three-Mode FPCA

Assuming that quantiles, estimated at different values of probability, share some common features, our purpose is identifying the principal directions along which summarizing their temporal dynamics by a small number of functional PCs.

More precisely, we follow [9], representing the collection of quantile functions, for a relevant set of probability values, $\alpha = \left[\alpha_1, \alpha_2, \ldots \alpha_Q\right]$, as realizations of a stochastic process $l_{\alpha}(t)$ with mean function $\mu_{\alpha}(t)$; then they can be represented in terms of the Karhunen-Loéven expansion:

$$l_{\alpha,i}(t) = \mu_{\alpha}(t) + \sum_{h=1}^{\infty} \psi_{h,i}\xi_h(t), \tag{7}$$

where:

- $E\left[l_\alpha(t)\right] = \mu_\alpha(t)$;
- $\psi_{h,i}$ are the PC scores and $\xi_h(t)$ are the eigenfunctions.

After estimating the quantile functions, they can be expanded by a reduced-rank model:

$$l_{\alpha,i}(t) = \mu_\alpha(t) + \sum_{h=1}^{H} \psi_{h,i}\xi_h(t), \tag{8}$$

where $\sum_{h=1}^{H} \psi_{h,i}\xi_h(t)$ is the linear combination of the first H PC scores and eigenfunctions, for a suitable choice of H on the basis of the explained variability.

From the functional quantiles, estimated by (3) and (4), principal components can be obtained by the three-mode eigen-analysis, that is well described in [3, 15], as the functions are expressed in terms of a linear combination of the basis matrix Φ and a three way array of coefficients, whose slices, one for each α, are the matrices Θ^α. An interesting result is the decomposition of the functions into two sets: the set of principal scores, one for each of the original curves, resuming variability along time and accounting for all the quantiles, and the set of corresponding time varying eigenfunctions, one for each probability level α, resuming variability among quantile functions of the same order.

The proposed methodology is implemented in R, using *splines*, *quantreg* and *fda* packages (http://cran.r-project.org), and it is applied to air pollution data.

4 The Air Pollution Data Set

Data concern concentrations of $PM10$ recorded during a year (2011) at different monitoring stations dislocated along the California state. Raw data are available at: http://www.epa.gov/airdata/ad_data_daily.

The whole network of monitoring stations contains 513 sites but, unfortunately, do not monitor all the pollutants. In particular, $PM10$ is recorded at 141 sites, but we retain only those 59 stations with at least 75% of data available per day.

The map of all the 141 sites, where $PM10$ is recorded (black dots), and of the 59 sites (red dots), is reported on the left of Fig. 1. In order to highlight the results of the procedure, a subset of 7 sites is selected; the map of the 7 selected sites (colored dots) for the 59 chosen sites (gray dots) is reported on the right of Fig. 1. A preliminary analysis has been carried out, in order to obtain daily syntheses: data have been aggregated by time, at each site, using daily average (24 h average concentration), according to EPA guidelines. Daily average is not computed with more than 25% of missing values on a day.

Data have been also standardized by linear interpolation, taking into account long term adverse health effects, as shown in a previous work [16]; EU instead of US EPA

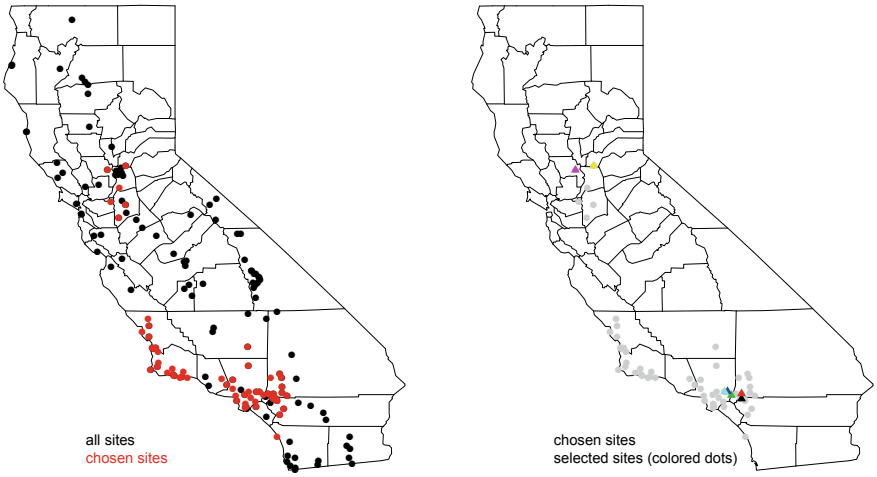


Fig. 1 Monitoring network for $PM10$

breakpoints are considered, allowing us to obtain values ranging in [0, 100], with threshold value corresponding to 50. After preprocessing, standardized observed data have been converted into functional.

5 Comparing FQ and FM Procedures

Functional quantiles and functional mean are estimated; in particular, for each station:

- five quantile regression curves for different values of α (0.1, 0.25, 0.5, 0.75, 0.9) (FQ procedure);
- the conditional mean function (FM procedure) is also estimated by minimizing the penalized residual sum of squares.

For FQ and FM procedures, we consider the P-spline approach, choosing the cubic B-spline basis system with equally spaced knots, a number of bases equal to 21 and a smoothing parameter equal to 20. The basis system (cubic B-splines) is assumed to be unique for all the considered sites. The number of knots is selected by means of generalized cross validation criterion (GCV). The use of the cubic B-spline basis system with equally spaced knots allows us to capture seasonal, monthly and weekly variations, but also events that occur irregularly and that cannot be expected periodically repeated.

The chosen value for the smoothing parameter ($\lambda = 20$) appears to be a fair compromise between what can be suggested by an automatic method, such as the GCV, and a subjective choice, that aims at smoothing rough data without hiding

their variability linked to possible peaks. In other words, our choices seem to be a good trade-off in smoothing between the removal of measurement error and the preservation of information.

Then, FPCA is performed on the curves obtained by both the procedures.

6 Results

Observed and functional data are reported, for each station, in Fig. 2; functional data are estimated by quantile regression for each value of α. In particular, in Fig. 2a the set of the N observed curves are represented; in Fig. 2b–f the estimated quantile functions for different probability values, from 0.1 to 0.9, synthesize the specific pattern of the respective curves. The gray lines represent all the stations; a subset of 7 curves, related to the selected sites reported in Fig. 1, are colored in order to point out the results of the procedure. The same sites are highlighted with the same colors in Figs. 4 and 7.

The explained variance resulting from Three-mode FPCA is reported in Table 1. Looking at the proportion of total variability explained by the PCs, the first three PCs explain 90% of the total variability among stations; in particular, the first PC (PC1) explains 79% of the variability, while the second mode of variation (PC2) explains 7% and the third (PC3) only 4%. A deeper understanding about the meaning of the PCs is obtained by looking at the proportion of variability accounted for by variations of each quantile, corresponding to each value of α; as it is highlighted in bold in Table 1, the most significant quantile for PC1 is the last ($Q_{0.9}$), since it explains 35% of variability, while the major contribution to PC2 and PC3 is given by the first one ($Q_{0.1}$). The curves obtained by the FQ procedure and by the FM

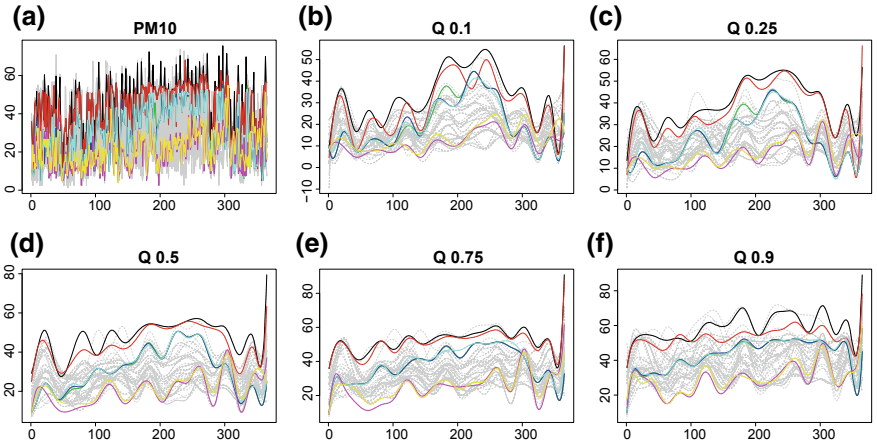


Fig. 2 Observed data (**a**) and estimated quantile regression curves (**b**)–(**f**)

Table 1 Main results from three-mode FPCA: explained variance

	$Q_{0.1}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.90}$	
PC1 0.79	0.11	0.14	0.19	0.22	**0.35**	1.00
PC2 0.07	**0.52**	0.15	0.02	0.09	0.22	1.00
PC3 0.04	**0.27**	0.17	0.21	0.16	0.19	1.00

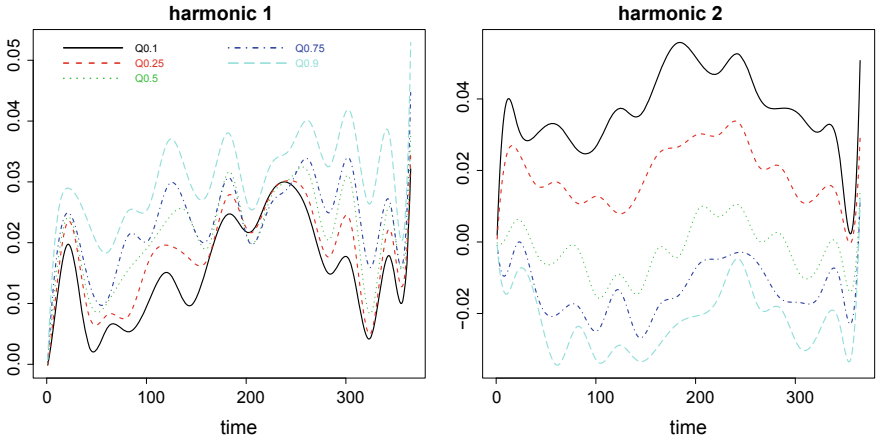


Fig. 3 Plots of the first two harmonics by FQ

procedure are projected in the space of the first two PCs. The plots of the first two harmonics are reported in Fig. 3. The eigenfunction, named harmonic in the FDA context, can be split into Q sub-eigenfunctions, one for each value of α. For the first harmonic, which is always positive, we can observe that weights are higher during the summer, especially for the first two quantiles (first two values of α; this means that low concentrations of $PM10$ are more affected by this trend, that is by this seasonal variation (see also Fig. 2b, c), with respect to high concentrations, presenting less variability (see also Fig. 2f). The second harmonic explain a low percentage of variance, so it is negligible.

Figure 4f shows the projections of the quantile functions, representing the sites, in the space of the first two PCs, together with the proportion of variance explained (0.793 and 0.073, respectively); Fig. 4a–e are the partial scores, that is those ones related to each value of α. We can observe that the two functional PCs retain the most information, almost the 80% of the original curves, moreover, curves with similar pattern have similar scores. In particular, we can distinguish three groups of sites. As most of the variation is explained by the first PC, then the distribution of monitoring stations along the PC1 axis is of greatest interest. According to what we say about the first harmonic, as we expect, the stations colored red and black have scores higher than other stations especially for low values of α, that is for the first quantiles. In general, looking at Fig. 4f, we say that sites with high scores on PC1 show a greater

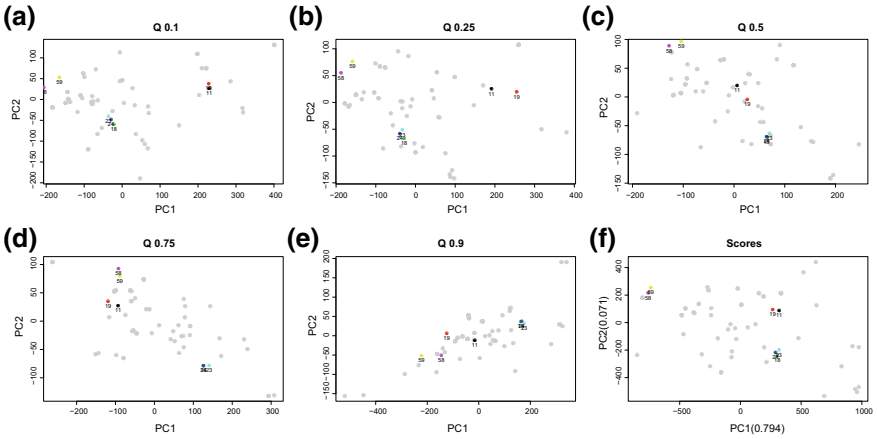


Fig. 4 Projection of the curves in the space of the first two partial (**a**)–(**e**) and total (**f**) PCs by FQ

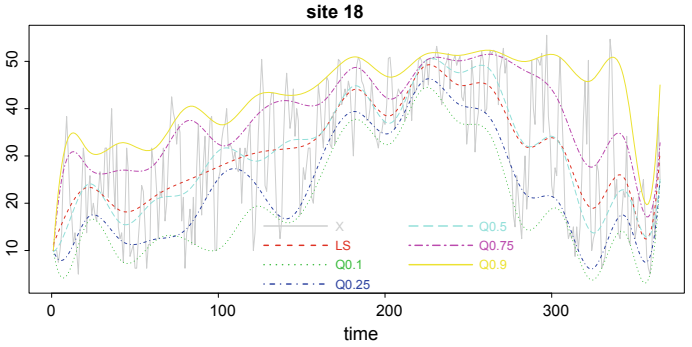


Fig. 5 Observed (X) and functional data, estimated by FM and FQ

variability in pollution levels (PM10) over time (sites colored red, black, blue and green; this can be observed also in Fig. 2a). Actually, as the last quantile ($Q_{0.9}$) weighs more than other quantiles on PC1, the first PC should represent the average level of air pollution accounted for PM10; then, the most polluted stations have the highest scores on PC1, while the least polluted stations have the lowest scores on PC1. The remaining stations record a level of pollution near the mean of the town pollution (origin of axes).

In Fig. 5, the five estimated quantile regression curves for different values of α (0.1, 0.25, 0.5, 0.75, 0.9) and the least squares estimate of the conditional mean function are shown for one selected station. The mean is colored red, while the median is colored cyan, the first and the fourth quartile are colored blue and magenta, respectively. In gray we report observed data. As we can observe, the quantile functions capture better than the mean function the variability of data.

In Fig. 6 the same results obtained for some other selected stations are reported.

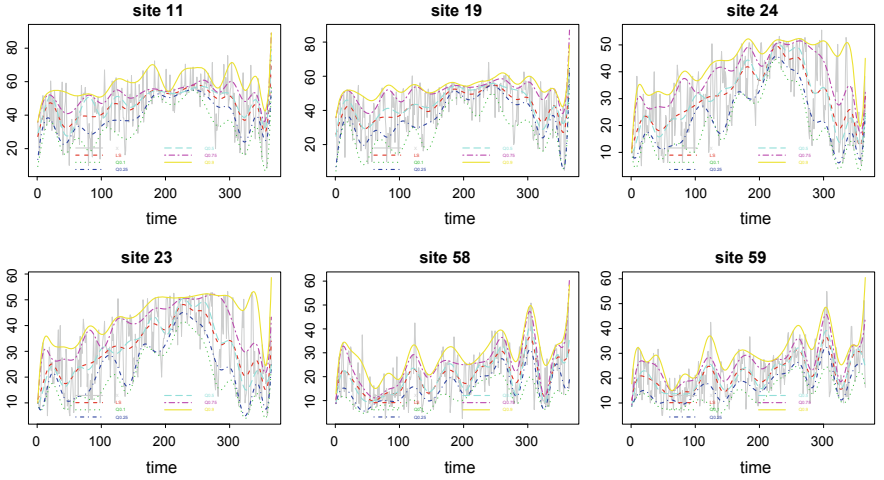


Fig. 6 Observed (X) and functional data, estimated by FM and FQ

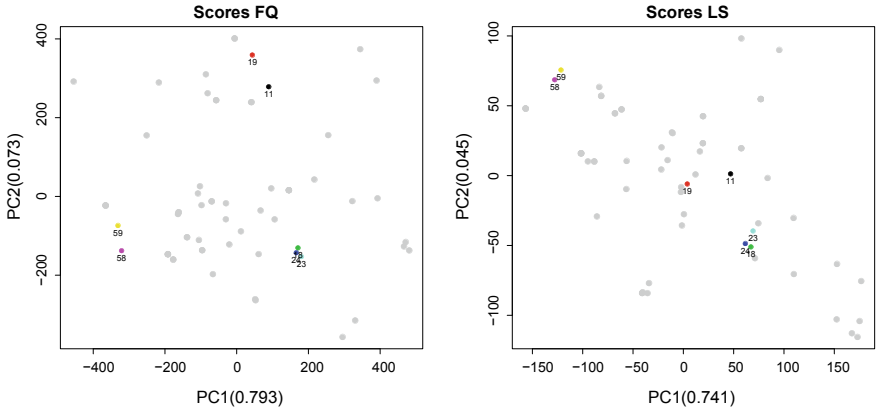


Fig. 7 Projection of the curves in the space of the first two PCs by FQ (left) and by FM (right)

In Fig. 7 the projections of the quantile functions (on the left) and the projections of the mean functions (on the right) are compared in the space of the first two PCs (the projections of the sites in the space of the PCs). The proportion of variance explained by the two PCs is also reported. As we can observe, the FQ procedure outperforms the FM procedure. In fact, although curves with similar pattern have similar scores for both cases, the functional PCs related to the quantile functions retain more information than the functional PCs related to the mean functions: the proportion of variance explained by the first PC in FQ increases up to 90% (80% in FM) while the second PC becomes negligible.

7 Conclusions and Further Development

In this work the simultaneous estimation of quantile functions is proposed in the FDA framework; using the method of penalized splines, the functions are represented in terms of linear combinations of bases and coefficients. Exploiting the covariance functions of the coefficients matrix, a small number of functional PCs synthesize common features of the estimated quantiles at different values of probability. Functional PCs, based on quantile functions of different order, are also compared with the functional PCs derived through the classical approach (FM procedure), that alternatively identify the directions of highest variability of the curves with respect to their functional mean. We show that our proposal, called FQ procedure, outperforms the FM procedure; in fact, the projection of the curves obtained with our estimation procedure, in the space of the first two PCs, allows to catch not only the variability of observed data, as in the traditional approach, but also the difference of the tails behaviour. Obviously, the higher the data variability, the better our procedure with respect to the FM one, especially if the average trend of the curves is very similar but the variability is quite different.

The approach has the advantage of further generalization, such as the inclusion of explanatory variables and distributional assumptions. The proposed approach copes with just one dimension, that is only time is considered as covariate, but it may be appealing to extend it to more than one dimension; for example, data can be modeled as functions of space or space-time jointly. At this aim, while in the classical approach a proper framework could be found in GAMs, allowing to deal with non gaussian data as well, the generalization of the concept of quantile functions to a multidimensional setting is not straightforward, since there is no natural order for R^n when $n \geq 2$.

A huge literature has been devoted to this topic in the last years with different methodological proposals; implication and appealing intuitions could be also borrowed from approaches relied on depth measures, in order to construct basic tools for functional data, as in [8].

Moreover, many consequent applications of the FPCA in quantile regression are motivated by the Karhunen-Loéve theorem, by means of which the random curves find convenient representations in terms of empirical orthogonal functions; the Karhunen-Loéve decomposition could be also useful for reconstruction of data in presence of missing and long gaps.

References

1. Cardot, H., Ferraty, F., Sarda, P.: Spline estimators for the functional linear model. Statistica Sinica **13**, 571–591 (2003)
2. Dabo-Niang, S., Laksaci, A.: Nonparametric quantile regression estimation for functional dependent data. Comm. Statist. Theory Methods **41**, 1254–1268 (2015)
3. Di Salvo, F., Ruggieri, M., Plaia, A.: Functional principal component analysis for multivariate multidimensional environmental data. Environ. Ecol. Stat. **22**(4), 739–757 (2015)

4. Eilers, P., Marx, B.: Flexible smoothing with B-splines and penalties. J. Am. Stat. Assoc. **11**, 89–121 (1996)
5. Ferraty, F., Rabhi, A., Vieu, P.: Conditional quantiles for functionally dependent data with application to the climatic El Nio Phenomenon. Sankhya **67**, 378–399 (2005)
6. Ferraty, F., Laksaci, A., Vieu, P.: Estimating some characteristics of the conditional distribution in nonparametric functional models. Statist. Infer. Stoch. Process **9**, 47–76 (2006)
7. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Springer, New York (2006)
8. Fraiman, R., Pateiro-Lopez, B.: Functional quantiles. In: Ferraty, F., Romain, Y. (eds.) Recent Advances in Functional Data Analysis and Related Topics. Contributions to Statistics, pp. 123–129. Physica-Verlag HD (2011)
9. Guo, M., Zhou, L., Huang, J.Z., Hardle, W.K.: Functional data analysis of generalized regression quantiles. Stat. Comput. **25**(2), 189–202 (2015)
10. Kaid, Z., Laksaci, A.: Functional quantile regression: local linear modelisation. In: Aneiros, G., Bongiorno, E.G., Cao, R., Vieu, P., (eds.) Functional Statistics and Related Fields. Contributions to Statistics. Springer, Cham (2017)
11. Kato, K.: Estimation in functional linear quantile regression. Ann. Stat. **40**(6), 3108–3136 (2012)
12. Koenker, R.: Quantile Regression. Cambridge University Press, New York (2005)
13. Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis. Springer-Verlag (2002)
14. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer-Verlag (2005)
15. Ruggieri, M., Plaia, A., Di Salvo, F., Agró, G.: Functional principal component analysis for the explorative analysis of multisite multivariate air pollution time series with long gaps. J. Appl. Stat. (2013)
16. Ruggieri, M., Plaia, A.: An aggregate AQI: comparing different standardizations and introducing a variability index. Sci Total Environ. (2012)
17. Ruppert, D., Wand, M., Carroll, R.: Semiparametric Regression. Cambridge University Press, Cambridge (2003)

Statistical Archetypal Analysis for Cognitive Categorization

Francesco Santelli, Francesco Palumbo and Giancarlo Ragozini

Abstract Human knowledge develops through complex relationships between categories. In the era of *Big Data*, the concept of categorization implies data summarization in a limited number of well-separated groups that must be maximally and internally homogeneous at the same time. This proposal exploits archetypal analysis capabilities by finding a set of extreme points that can summarize entire data sets in homogeneous groups. The archetypes are then used to identify the best prototypes according to Rosch's definition. Finally, in the geometric approach to cognitive science, the Voronoi tessellation based on the prototypes is used to define categorization. An example using a well-known wine dataset by Forina et al. illustrates the procedure.

Keywords Archetypal analysis · Prototyping · Statistical learning

1 Introduction

"*Knowledge consists basically of categorizations and corrections of categorizations so that we can adapt ourselves to our environment*" [31]. Humans can learn new concepts quickly by building complex relationships between a set of complex items or categories. Whilst the total number of objects considered should remain limited to five or six, these objects can be described by several features that define a high grade of complexity. Categories are stored in our long-term memory, and it has been demonstrated that we recall these categories in our working memories, developing

F. Santelli
Dept. of Social Sciences, Università degli Studi di Napoli Federico II, Naples, Italy
e-mail: francesco.santelli@unina.it

F. Palumbo (✉) · G. Ragozini
Dept. of Political Sciences, Università degli Studi di Napoli Federico II, Naples, Italy
e-mail: fpalumbo@unina.it

G. Ragozini
e-mail: giancarlo.ragozini@unina.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_7

connections among them that improve our knowledge [7]. In other words, a few examples of a new concept are often sufficient for us to grasp the concept's meaning. On the contrary, we are often overwhelmed by large amounts of data and information.

With the explosion of Big Data, statistical learning has become a very hot field in many scientific areas as well as in marketing, finance, and other environmental and behavioral disciplines. The huge amount of stored data represents an incredible source of knowledge, provided that it can be summarized in a (small) number of categories that are consistent with human cognitive capabilities.

In the present paper, we parallel the cognitive process of categorization through statistical learning techniques, relying on the conceptual space framework [18] in which conceptual spaces are geometric structures and the categorization mainly consists in partitioning the conceptual spaces. The paper is structured in six sections following this introduction: Sect. 2 describes how developments in cognitive science have evolved into conceptual space theory. Section 3 discusses the relationship between statistical learning and the construction of categorizations in cognitive science. Section 4 lists a consolidate formalization [1] of objects in the topological conceptual space. Section 5 presents the prototype identification after the archetypal analysis; through a real data-based example, Sect. 6 presents the Voronoi tessellation [35] beginning with the prototypes as a tool for deriving a categorization in the conceptual space, and the last section presents several concluding remarks and possible directions for future research.

2 Conceptual Space Framework in Cognitive Representation

The theoretical framework field in cognitive science mainly defines the ways in which learning is developed given a set of hypotheses about the fixed structures of the mind and how the different components work together. This complex system and the way it works is usually defined as cognitive architecture. It can be related to both the human mind and artificial systems. Currently, the three most common approaches to the model learning process are considered to be symbolism, connectivism and conceptual space theory [19]. The first approach (symbolism) makes the assumption that learning processes can be properly described by means of Turing's machine, which processes symbols according to a table of rules without taking into account the semantic context. It mainly aims to model high-level abstract entities, performing inference to figure out them using mostly first-order logical predicates. Starting with the associationism theory (for Locke and Hume, learning consists of associations among perceptions), the second theory revived in recent years developed into *connectivism*. This theory began to have more space year by year thanks to its innate relationships with the increase in the availability of a huge amount of data due to technology development [34]. From a statistical point of view, the arising system was called artificial neuron (or neural) network. Lastly, as introduced by Gärdenfors [18], the third approach is the formalization of information structures made by a number of quality dimensions embedded in a topological space called the conceptual space.

In this space, it is possible to carry out an analysis considering its metric nature. The concept of similarity between entities becomes, as a result, closely related to the metric distance between them, given the quality dimensions under investigation. In this framework, the natural property in a domain is a convex region [36]; therefore, the focal points of each region are prototypes of the categories, and all entities close enough to the prototype belong to the same category.

3 Statistical Learning and Cognitive Categorization

Statistical and machine learning can significantly speed up human knowledge development, helping to determine the basic categories in a relatively short amount of time. Exploratory data analysis (EDA) can be considered the forefather of statistical learning; it relies on the mind's ability to learn from data and, in particular, it aims to summarize datasets through a limited number of interpretable latent features or clusters offering cognitive geometric models to define categorizations. It can also be understood as the implementation of the human cognitive process extended to huge amounts of data: "*Big Data*" [20]. Factorial models belong to the former approach, they permit the representation of the original data into a reduced space by replacing the original variables with a reduced number of linear mixtures of independent components. These methods include principal component analysis (PCA), independent component analysis (ICA), and independent vector analysis (IVA), when dealing with multiple datasets. On the other hand, fuzzy and crisp clustering methods allow us to represent each statistical unit as a weighted sum of the means of the groups that minimize overall model error.

However, EDA itself cannot answer to the questions: "*How many, and what are the categories to retain?*" and "*What are the observations that can represent a category better than others. in human cognitive processes?*". In cognitive science, according to Rosch [32, 33], the best observation is related to the concept of typicality; in other words, we must look for those elements that can represent a category better than others. From a general perspective, in a cognitive science domain, categorization is assumed to be a set of processes of determining units that belong together according to a criterion. A category is a group or class of stimuli or entities that bear a physical similarity among them. Concepts are thought to be the knowledge that facilitates the categorization process [3], and in the conceptual space, there are convex regions for more than one domain (therefore, natural property, considered for only one domain,is a special and simpler case of a concept).

We call *prototypes* those elements that are able to represent a category and measure their representativeness degree using a distance function to a salient entity of the category [15, 29]. These objects can be observed or unobserved (abstract), and they can be represented by a single value or by interval-valued variables. In many cases, in classification and clustering, and more generally in cognitive sciences, the concept of *prototype* has been unknowingly adopted to synthesize and represent categories [4, 6]. However, regarding Big Data, the role of prototypes has become more and more

relevant, thus giving rise to a wide variety of studies in the literature on prototype-based clustering methods (see [21, Chap. 13]).

Identifying groups that can be connected to a related prototype does not fulfill the categorization process. Without proper description, prototypes cannot be advantageous to learning. D'Esposito et al. (2012, 2013) [9, 10] and Ragozini et al. (2016) [29] considered the archetypal analysis, as proposed by Cutler and Breiman [8], to identify prototypes from a geometric perspective. According to the idea of symbolic object [12], in [10], D'Esposito et al. (2013) proposed the prototype description in terms of symbolic objects. The present proposal grounds on the conceptual space framework and starting from the geometric properties of the proposed prototypes exploits the Voronoi tessellation to obtain a data-driven categorization; i.e. a partition of the conceptual space in convex regions centered on the prototypes. This procedure can be summarized in a proposal to achieve a categorization in two steps: (1) a data-drive prototype analysis and (2) the ensuing Voronoi tessellation based on the identified prototypes.

4 Formalization of Objects in a Conceptual Space

In the conceptual space framework, some authors have proposed the integration/creation of a comprehensive algebra. Given that conceptual spaces are based on the paradigm of cognitive semantics [23], they are dynamic systems under the assumption that algebraic operations between concepts or entities are allowed. To allow them, formal definitions of the *objects* embedded in this space are needed. Going through the hierarchical classification proposed by Adams [1], the base element is the *quality dimension* tool that measures and orders entities in the space according to a specific feature/characteristic. The quality dimension is, in turn, made of three factors: a measurement level or scale (ratio, interval, or ordinal, the range of the dimension (in which the boundaries are minimum and maximum values), and whether it is circular. A *quality domain*, on the other hand, is a finite set of quality dimensions. Therefore, latitude and longitude, for example, are two distinct quality dimensions; however, once brought together, they form a quality domain of coordinates. *Instances* are a finite set of points in one or more domains; a specific point is a vector of the values assumed by the quality dimensions. These values represent an instrument for measuring and ordering different quality values of objects in the space. A bounded intersection of half-spaces is a method (H-polytope representation) of building a *convex region*; in this layered structure, a *concept* is a finite set of convex regions.

5 Prototype Identification

In statistical literature, numerical techniques to find prototypes in given multivariate datasets have been proposed and are based on several different criteria. The most widely used techniques are generally based on non-hierarchical clustering algorithms

[11, 22]. However, in this proposal, we present some recent results on the prototypes definition through an archetypal analysis (AA). AA was first introduced by Cutler and Breiman [8]. It is mainly a matrix factorization method of a generic $n \times p$ data matrix $\mathbf{X}$ such that $\min_{\boldsymbol{\Gamma}\mathbf{A}} \{||\mathbf{X} - \boldsymbol{\Gamma}\mathbf{A}||_F\}$, where $\boldsymbol{\Gamma}$ and $\mathbf{A}$ represent the factorization matrices of order $n \times k$ and $k \times p$, respectively, with $\mathbf{A} = \mathbf{B}\mathbf{X}$ and $||\cdot||_F$ states for the Frobenius norm. Matrices $\mathbf{B}$ and $\boldsymbol{\Gamma}$ have nonnegative entries and must satisfy the following constraints: (i) $\mathbf{B}\mathbf{1}_n = \mathbf{1}_k$ and (ii) $\boldsymbol{\Gamma}\mathbf{1}_k = \mathbf{1}_n$, where $\mathbf{1}$ is a vector of ones. The $k \times p$ matrix $\mathbf{A} = \mathbf{B}\mathbf{X}$ represents the k archetypes, where k is assumed as a priori defined. It is worth noting that the matrix $\boldsymbol{\Gamma}$ defines a fuzzy allocation rule of each data point to the k archetypes; let us indicate with γ_{ij} the general term of $\boldsymbol{\Gamma}$, with $i = 1, \ldots n$ and $j = 1, \ldots k$. Additionally $\sum_j \gamma_{ij} = 1$, γ_{ij} represents the membership degree of $\mathbf{x}_i$ to the archetype $\mathbf{a}_j$. The quantity to be minimized by the algorithm is the residual sum of squares (*RSS*), and it generally does not have a closed form solution. It could be solved by means of general-purpose, non-linear constrained least squares; however, a consolidate approach is to use an alternating least square algorithm [5, 8]. It starts from the whole RSS, then it is divided into two quantities (in the first one, it finds the best γ_{ij} given the set of archetypes, and in the second one, it finds the best β_{ij} given the recalculated archetypes) and solves them using an iterative procedure, finding a local minimum for the criterion.

Setting up structural constraints makes learning more efficient. In other words, one can constrain the learning process in a convex space. However, adding structural constraints often means that some form of information about the relevant domains or other dimension-generating structures is added. Consequently, this strategy presumes a conceptual level in the construction of the prototypes. AA exploits redundancies in input data; it finds the number of archetypes in the input data that can be used to represent (approximate) all data points. It is worth noting that AA constraints ensure symmetrical relationships between archetypes and data points; archetypes are convex combinations of data points and data points are approximated in terms of the convex combinations of archetypes. The first constraint ensures that the archetypes to be found will lie on the convex hull of the data cloud, giving them the peculiar trait of being extremal points.

In this view, we propose a geometric approach that allows prototype identification to be the most typical object within a group or a category. A prototype is the member within a group that best represents the other members (i.e.,in terms of internal resemblance) and that at the same time differs from the members of the other groups or categories (i.e., an external dissimilarity). This double semantics related to centrality and extremeness can been operationalized through a typicality index $T(\cdot, \cdot)$ [17, 24, 25, 30].

Formally, given a set of n objects $\Omega = \{\mathbf{x}_i\}_{i=1,\ldots,n}$, $\mathbf{x}_i \in \Re^p$ and a partition $C = (C_1, \ldots, C_k)$ of Ω in k groups, an internal resemblance measure $R(\mathbf{x}_i, C_h)$ of $\mathbf{x}_i$ w.r.t. $\mathbf{x}_{i'} \in C_h$, an external dissimilarity measure $D(\mathbf{x}_i, \overline{C_h})$ of $\mathbf{x}_i$ w.r.t. $\mathbf{x}_{i'} \notin C_h$, and a mixing function $\Phi(\cdot)$ that combines both measures, and a typicality index $T(\mathbf{x}_i, C_h)$ of $\mathbf{x}_i$ with respect to the class C_h is given by:

$$T(\mathbf{x}_i, C_h) = \Phi(R(\mathbf{x}_i, C_h); D(\mathbf{x}_i, \overline{C_h})). \tag{1}$$

The set of prototypes $\mathscr{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is then defined as:

$$\mathscr{P} = \{\mathbf{p}_h \in \Re^p | \mathbf{p}_h = \arg\max_{\mathbf{x}_i} T(\mathbf{x}_i, C_h), h = 1, \ldots, k\}. \tag{2}$$

It is clear that in this framework and setting, the prototype identification depends on the ways in which the dissimilarity and resemblance are measured and on the partition assumed in advance. The main proposals in this direction for prototype identification assume that both resemblance and dissimilarity measures are based on the Euclidean distance. The semantic of prototypes is also strongly affected by the choice of the mixing or aggregating function $\Phi(\cdot, \cdot)$. If one considers only the internal resemblance, the prototypes will be the central elements of the groups; on the other hand, if one takes into account only the external dissimilarity, the prototypes will be the most extreme points. The mixing function $\Phi(\cdot, \cdot)$ yields a compromise between these two instances. In this framework, the proposal to identify prototypes through the archetypes is made in order to have well-separatedand informative points that represent categories. The procedure can be described in three steps. Prototypes in the beginning of the procedure are identified as the archetypes, maximizing the criterion of external dissimilarity and seeking to a principle of pureness in the categories. Therefore, clusters around the archetypes are built in space spanned by these archetypes, and the centers of these clusters are the new prototypes, achieving the internal resemblance purpose. In the last step, the two previous solutions are combined in the original space to determine the final prototypes; these are, in the end, a compromised solution between the archetypes and the centers of clusters around these archetypes.

Specifically, archetypes can be considered first-step prototypes. However, because archetypes belong to the data convex hull, they lie on the boundary of data scatter; as such, they are extreme points with respect to the other points, and they maximize the external dissimilarity. To improve the internal resemblance of the archetypes, we revert to the space where the archetypes are the vertices of a K-dimensional simplex, i.e., $\mathscr{S}^k$, and each data point $\mathbf{x}_i'$ is represented as a point with barycentric coordinates $\boldsymbol{\gamma}_i'$ [28]. In this simplex, we obtain a partition $C = (C_1, \ldots, C_k)$ of the data set by clusterizing the data around the archetypes, exploiting the properties of the $\boldsymbol{\gamma}_i$ coefficients. If γ_{ih} is close to 1, the point $\mathbf{x}_i$ is very close to the archetype $\mathbf{a}_h$. If γ_{ih} is close to 0, $\mathbf{x}_i$ lies far from $\mathbf{a}_h$. As classifiers, we can adopt a crisp allocation rule (or nearest neighbor rule) where

$$C_h = \{\mathbf{x}_i : \arg\max_j \gamma_{ij} = h\}, \qquad h = 1, \ldots, k, \tag{3}$$

or a fuzzy allocation rule where

$$C_h^\tau = \{\mathbf{x}_i : \gamma_{ih} > \tau\}, \qquad 0 < \tau < 1, \quad h = 1, \ldots, k. \tag{4}$$

Given the partition $C = (C_1, \ldots, C_k)$, we maximize the internal resemblance within each group of the partition, or equivalently, we minimize the internal dissim-

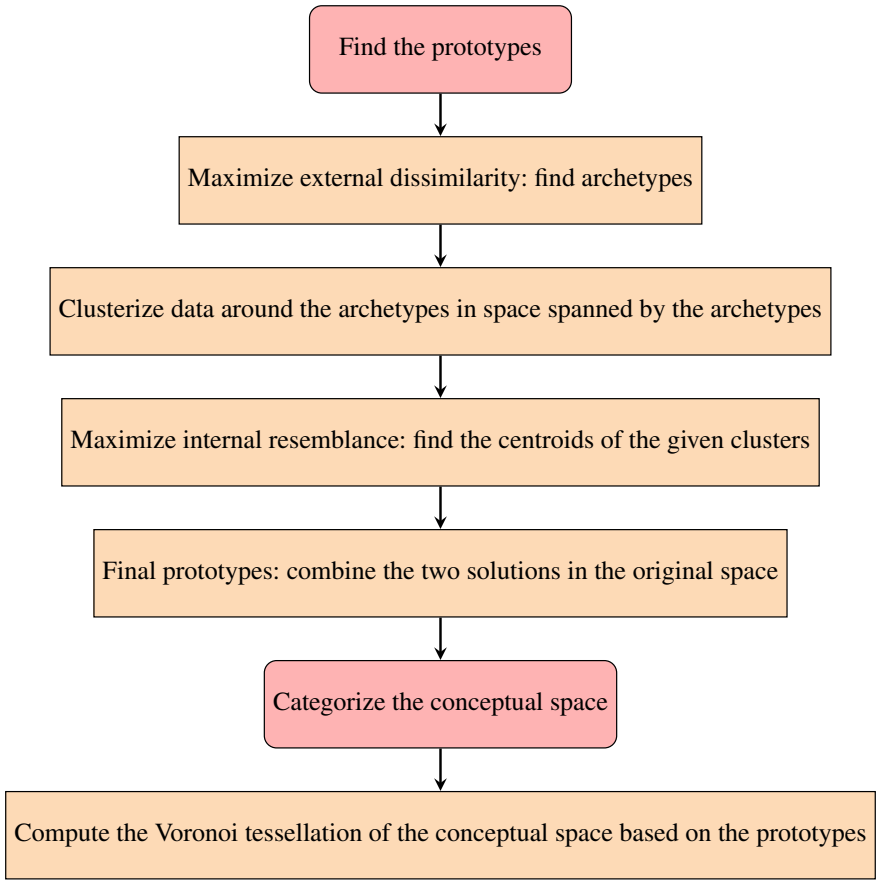


Fig. 1 Flowchart of the entire procedure, from the prototype identification to the Voronoi tessellation

ilarity within each cluster, determining the centroids $(\mathbf{c}_1, \ldots, \mathbf{c}_k)$ of the clusters by solving the following minimization problem:

$$\min_{(\mathbf{c}_1,\ldots,\mathbf{c}_k)} \sum_{\mathbf{x}'_i \in C_h} d(\boldsymbol{\gamma}_i, \mathbf{c}_h) \forall h \tag{5}$$

where $d(\cdot, \cdot)$ is an appropriate dissimilarity measure in the space $\mathscr{S}^k$.

The centroids $(\mathbf{c}_1, \ldots, \mathbf{c}_k)$ can be assumed to be prototypes in the space $\mathscr{S}^k$. The final prototypes $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ in the space of the data points are then obtained by reverting to the $\Re^p$ space:

$$\mathbf{p}_h = \mathbf{c}_h \mathbf{A}(h); \tag{6}$$

that is, each $\mathbf{p}_h$ is a convex combination of the archetypes $\mathbf{A}(h)$ withcoefficients $\mathbf{c}_h$.

The last step of the categorization procedure consists of the partitioning of the conceptual space, starting from prototypes. Given the triple $\Delta(\mathscr{P}, d, \mathscr{C})$ where $\mathscr{P}$ is a set of given prototypes and d is a distance measure defined on a conceptual space $\mathscr{C}$, the tessellated region $c(\mathbf{p}_h)$ is defined such that:

$$\{x \mid d(\mathbf{p}_h, x) \leq d(\mathbf{p}_{h'}, x)\},$$

$\forall h \neq h'$, where x is a generic data point belonging to $\mathscr{C}$ and $c(\mathbf{p}_h)$ is the category generated by $\mathbf{p}_h$.

When the conceptual space is assumed to be the Euclidean one, the categories $c(\mathbf{p}_h)$ obtained through this procedure correspond to the Voronoi cells derived by the Voronoi tessellation [13] based on the prototypes. Thus, the categories are convex regions of the conceptual space, covering it, and allowing for the easy classification of all the other points belonging to the conceptual space, both observed and unobserved.

The entire proposed procedure, from AA to the categorization through to Voronoi tessellation, is presented in the following flow chart (Fig. 1).

6 Categorization Using Voronoi Tessellation: The `wine` Dataset

In the conceptual space framework, the categorization problem can be solved by a partitioning of the space through the Voronoi tessellation, starting with a given set of prototypes. In our approach, we provide a way to derive prototypes from data [29]. We note that the geometrical properties of our prototypes are congruent with the conceptual space approach; then, we propose the use of our data-driven prototypes for the Voronoi tessellation in order to obtain a categorization. In addition, in cognitive science, it is often assumed that the number of prototypes and typologies in the data is a priori known. However, in any real world cognitive study, things are completely different and the *true* number of typologies must be inferred by studying the groups in the data. However, to decide on the number of groups is one of the most widely addressed problems in cluster analysis, and most likely has no satisfactory solution that can be generalized in any category of problem. By dealing with extreme data points, AA allows us to choose the number of archetypes according to the behavior of the loss functions evaluated at different numbers of archetypes. The loss function is plotted on a Cartesian coordinate system where the x-axis represents the number of archetypes and the y-axis represents the value of the loss function (decreasing by definition); the optimal number of archetypes should be revealed by an elbow of the function (graphically: the loss function begins parallel to the x-axis). However, the presence of multivariate outliers or highly correlated variables could mask the *true number* in favor of redundant or unstable solutions. Deeper investigations based on computationally intensive studies can reveal such situations.

In this section, we consider the `wine` dataset. First presented by Forina et al. [16], it contains data pertaining to 178 wines produced from three different Italian

Table 1 List of labels and variable names of the the `wine` dataset

Labels	Variable name
Alc	Alcohol
Mal	Malic acid
Ash	Ash
Alk	Alkalinity of ash
Mag	Magnesium
Phe	Total phenols
Fla	Flavanoids
NFla	Non-flavanoid phenols
Pro	Proanthocyanidins
Col	Color intensity
Hue	Hue
Dil	OD280/OD315 of diluted wines
Prol	Proline

Table 2 Wine data: archetypes as the first solution

	Alc	Mal	Ash	Alk	Mag	Phe	Fla	NFla	Pro	Col	Hue	Dil	Prol
a_1	14.19	1.97	2.51	16.45	114.63	3.24	3.40	0.26	2.21	6.68	1.05	3.28	1316.07
a_2	13.22	3.78	2.48	22.12	97.47	1.56	0.65	0.49	1.05	7.69	0.63	1.51	621.94
a_3	11.79	1.41	2.07	20.04	86.50	2.26	1.97	0.34	1.61	2.15	1.20	3.08	406.40

cultivars (*barbera, barolo, and grignolino*) and described by the 13 features that refer to organoleptic and chemical categories (Table 1).

As the three different varieties of wine are recognized as having their specific properties, we assume that each of them represents a category and can be summarized by a prototype.

The first step of the entire procedure consists of the archetype identification. The `archetypes` package [14], available at the `CRAN` repository, permits the identification of the optimal number of archetypes. Here, we set the number of archetypes to three. We refer interested readers to [29] for a more detailed description of the choice of the number of prototypes. Table 2 reports the three archetypes described by their 13 original variables (expressed in their own original scales).

The second step consists of grouping the points around the archetypes in the space defined by the matrix $\boldsymbol{\Gamma}$. In this example, a crisp classification has been taken into account. The fuzzy allocation rule can also be taken into account; it can ensure a higher "*purity*" degree in the groups and (generally) produces an extra group with respect to the number of archetypes. The three groups, corresponding to the three archetypes, are visualized in the space spanned by the three columns of $\boldsymbol{\Gamma}$ in Fig. 2.

The group's centroids are identified by the generalized compositional geometric mean of the group, computed from the γ_{ij} membership scores. Exploiting the relationship between the geometric basis spanned by the archetypes and the original space [2], prototypes can be represented in the original variable space.

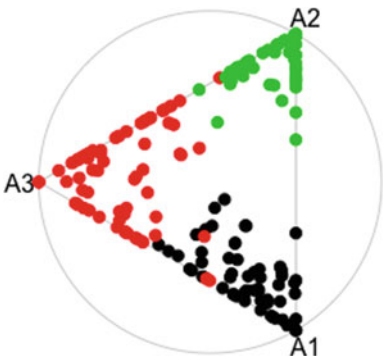
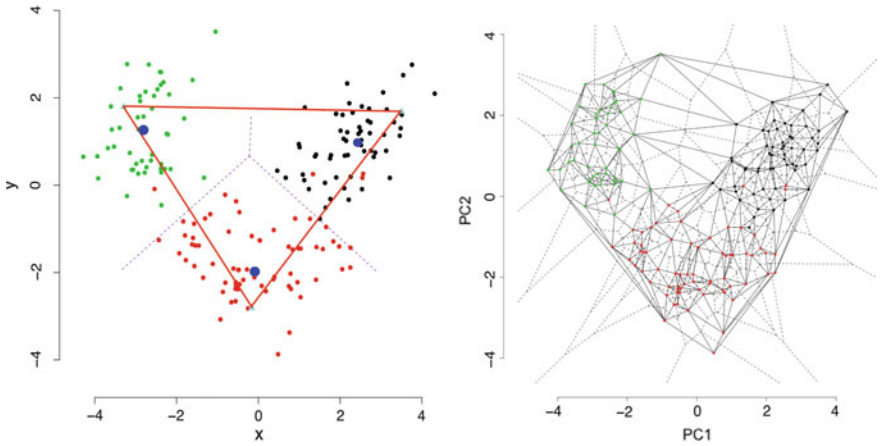


Fig. 2 Wine data set: groups around the archetypes obtained by the crisp allocation rule



(a) Archetypes, prototypes and Voronoi convex regions

(b) Voronoi tesselation

Fig. 3 Wine data set: plots **a** and **b** represent the Voronoi tessellation and the convex geometric region on the first two principal components. In figure **a**, the red triangle vertices represent the archetypes, the blue points refer to the prototypes, and the dashed lines represent the edges of the convex regions that correspond to the three categories

It has been shown that in a metric space, representations of properties are obtained as convex regions. Let us consider the set of prototypes $\mathscr{P} = \{p_1, p_2, \ldots, p_K\}$]; their representation in any conceptual space implies (according to the definition of "prototype" itself) that they are the central points in the categories they represent. The distance between any prototype point p and p' represents their *external dissimilarity*. If we assume that any generic point x_i belongs to the same category as the closest prototype, it has been shown that this rule will generate a partitioning of the space into convex regions [19, 26]. This partition/categorization is given by the Voronoi tessellation of the conceptual space based only on the prototypes. Note that this

approach also has computational advantages. The tessellation is performed using only a few points, i.e., the prototypes; thus, given the geometric properties of the Voronoi tessellation, the allocation on new instances in a given category can be done in a very easy and efficient way.

The two plots in Fig. 3a, b represent the Voronoi tessellation on the first two principal components (29% of the total variance). Figure 3a summarizes the entire categorization process: (*i*) the triangle vertices represent the three archetypes; (*ii*) the blue points (larger than the other points) refer to the prototypes; and (*iii*) the dashed lines converging in the center define the convex regions associated with the three categories, i.e., the Voronoi cells associated with the three wine prototypes. It is worth noting that the prototypes appear more internal with respect to the corresponding archetypes.

Figure 3b, on the right hand side, shows the entire tessellation around the three prototypes that developed with respect to the 178 observed points. It is easy to notice that the categorization given by the tessellation reproduces the three wine typologies well.

7 Conclusion

Several alternative cognitive approaches are grounded in the geometric representation between *properties* and *concepts* in convex conceptual spaces. Based on the connection between statistical learning and cognitive categorization, our method allows the partitioning of a convex conceptual space into convex regions corresponding to the categories through the joint use of Voronoi tessellation and prototype identification. Thus, assuming that a Euclidean metric is defined on the subspace that is subject to categorization, a set of prototypes will generate a unique partition of the subspace into convex regions using this method. In this way, the Voronoi tessellation and archetypes provide a constructive geometric answer for how a similarity measure and a set of prototypes determine a set of categories.

Finally, the proposed procedure can also work in the case of conceptual spaces with different metrics. For example, in the case of interval-valued data, prototypes can be derived using the Hausdorff distance [9], and a coherent Voronoi tessellation should be adopted [27]. In this case, however, the convexity properties and the corresponding cognitive interpretations should be carefully checked.

References

1. Adams, B., Raubal, M.: A metric conceptual space algebra. In: Spatial Information Theory. Springer (2009)
2. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J., Pawlowsky-Glahn, V.: Logratio analysis and compositional distance. Math. Geol. **32**(3), 271–275 (2000)

3. Barsalou, L.W.: Deriving categories to achieve goals. In: Psychology of Learning and Motivation. Elsevier (1991)
4. Bien, J., Tibshirani, R.: Prototype selection for interpretable classification. Ann. Appl. Stat. 2403–2424 (2011)
5. Bauckhage, C., Thurau, C.: Making archetypal analysis practical. In: DAGM-Symposium, pp. 272–281 (2009)
6. Chang, F., Lin, C.C., Lu, C.J.: Adaptive prototype learning algorithms: theoretical and experimental studies. J. Mach. Learn. Res. **7**, 2125–2148 (2006)
7. Cowan, N.: The magical mystery four: how is working memory capacity limited, and why? Curr. Dir. Psychol. Sci. **19**(1), 51–57 (2010)
8. Cutler, A., Breiman, L.: Archetypal analysis. Technometrics **36**(4), 338–347 (1994)
9. D'Esposito, M., Palumbo, F., Ragozini, G.: Interval archetypes: a new tool for interval data analysis. Stat. Anal. Data Min. **5**(4), 322–335 (2012)
10. D'Esposito, M.R., Palumbo, F., Ragozini, G.: Archetypal Symbolic Objects, pp. 41–49. Springer, Berlin (2013)
11. Diday, E.: Optimization in non-hierarchical clustering. Pattern Recognit. **6**(1), 17–33 (1974)
12. Diday, E.: Categorization in symbolic data analysis. In: Cohen, H., Lefebvre, C. (eds.) Handbook of Categorization in Cognitive Science, pp. 845–867. Elsevier Science Ltd, Oxford (2005)
13. Edelsbrunner, H.: Algorithms in Combinatorial Geometry, vol. 10. Springer-Verlag, Berlin (1987)
14. Eugster, M., Leisch, F., Seth, S.: Archetypes: archetypal analysis. In: R Package Version, pp. 2–2 (2014)
15. Fordellone, M., Palumbo, F.: Prototypes definition through consensus analysis between fuzzy c-means and archetypal analysis. Ital. J. Appl. Stat. **26**(2), 141–162 (2014)
16. Forina, M., Armanino, C., Castino, M., Ubigli, M.: Multivariate data analysis as a discriminating method of the origin of wines. Vitis **25**, 189–201 (1986)
17. Frigui, H., Nasraoui, O.: Unsupervised learning of prototypes and attribute weights. Pattern Recognit. **37**(3), 567–581 (2004)
18. Gärdenfors, P.: Conceptual Spaces: The Geometry of Thought. A Bradford Book, vol. 3, p. 16. MIT Press (2000)
19. Gärdenfors, P.: Concept Learning and Nonmonotonic Reasoning, pp. 823–843. Elsevier (2005)
20. Grolemund, G., Wickham, H.: A cognitive interpretation of data analysis. Int. Stat. Rev. **82**(2), 184–204 (2014)
21. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2011)
22. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. (CSUR) **31**(3), 264–323 (1999)
23. Lakoff, G.: Cognitive semantics. Mean. Ment. Represent. **8**(1), 119–154. https://escholarship.org/uc/item/04086580 (1988)
24. Lesot, M., Kruse, R.: Typicality degrees and fuzzy prototypes for clustering. In: Advances in Data Analysis, pp. 107–114. Springer (2007)
25. Lesot, M., Rifqi, M., Bouchon-Meunier, B.: Fuzzy prototypes: from a cognitive view to a machine learning principle. In: Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, pp. 431–452. Springer (2008)
26. Okabe, A., Boots, B., Sugihara, K.: Nearest neighbourhood operations with generalized voronoi diagrams: a review. Int. J. Geogr. Inf. Syst. **8**(1), 43–71 (1994)
27. Papadopoulou, E., Lee, D.T.: The Hausdorff Voronoi diagram of polygonal objects: a divide and conquer approach. Int. J. Comput. Geom. Appl. **14**(06), 421–452 (2004)
28. Porzio, G.C., Ragozini, G., Vistocco, D.: On the use of archetypes as benchmarks. Appl. Stoch. Model. Bus. Ind. **25**, 419–437 (2008)
29. Ragozini, G., Palumbo, F., D'Esposito, M.R.: Archetypal analysis for data-driven prototype identification. Stat. Anal. Data Min.: ASA Data Sci. J. (2016)
30. Rifqi, M.: Constructing prototypes from large databases. In: International conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU'96, Granada, Spain. https://hal.archives-ouvertes.fr/hal-01075383 (1996)

31. Robert, S.: Categorization, reasoning and memory from a neo-logical point of view. In: Cohen, H., Lefebvre, C. (eds.) Handbook of Categorization in Cognitive Science, pp. 700-718. Elsevier (2005)
32. Rosch, E.: Natural categories. Cogn. Psychol. **4**(3), 328–350 (1973)
33. Rosch, E.: Prototype classification and logical classification: the two systems. In: New Trends in Conceptual Representation: Challenges to Piaget's Theory, pp. 73–86 (1983)
34. Siemens, G.: Connectivism: A Learning Theory for the Digital Age. http://er.dut.ac.za/handle/123456789/69 (2014)
35. Watson, D.F.: Computing the n-dimensional delaunay tessellation with application to Voronoi polytopes. Comput. J. **24**(2), 167–172 (1981)
36. Zenker, F., Gärdenfors, P. (ed.): Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation. Springer, Berlin; Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science. **24**(2), 2–10. Springer (2015)

Inferring Rater Agreement with Ordinal Classification

Amalia Vanacore and Maria Sole Pellegrino

Abstract In several contexts ranging from medical to social sciences, rater reliability is assessed in terms of intra (-inter) rater agreement. The extent of rater agreement is commonly characterized by comparing the value of the adopted agreement coefficient against a benchmark scale. This deterministic approach has been widely criticized since it neglects the influence of experimental conditions on the estimated agreement coefficient. In order to overcome this criticism, in this paper an inferential procedure for benchmarking is presented. The proposed procedure is based on non-parametric bootstrap confidence intervals. The statistical properties of the proposed procedure have been studied for two bootstrap confidence intervals via a Monte Carlo simulation. The simulated scenarios differ for sample sizes (i.e. n = 10, 30, 50, 100 items) and rating scale dimensions (i.e. k = 2, 3, 5, 7 categories).

Keywords Rater agreement · Weighted Uniform kappa coefficient · Statistical benchmarking · Monte Carlo simulation

1 Introduction

In many contexts of research (e.g. cognitive and behavioural science, quality science, clinical epidemiology, diagnostic imaging, content analysis), there is frequently a need to assess the accuracy of human instruments (i.e. raters) providing subjective measurements, expressed on a dichotomous, nominal or ordinal rating scale.

ISO 5725 (1994) [21] refers the term *accuracy* to both systematic bias (i.e. trueness) and random error (i.e. precision). Actually, by definition, subjective evaluations lack a reference value for assessing their trueness and the common concept of accuracy cannot be easily operationalized. In such circumstances, the accuracy of

A. Vanacore (✉) · M. S. Pellegrino
Department of Industrial Engineering, University of Naples "Federico II", Naples, Italy
e-mail: amalia.vanacore@unina.it

M. S. Pellegrino
e-mail: mariasole.pellegrino@unina.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_8

subjective evaluations can be related only to precision and assessed as the degree of agreement (i.e. "closeness") between repeated evaluations.

The easiest way of measuring agreement between ratings is to calculate the overall percentage of agreement; nevertheless, this measure does not take into account the agreement that would be expected by chance. A reasonable alternative is to adopt the widespread kappa-type coefficient that was introduced by Cohen in 1960 [9] as a rescaled measure of the probability of observed agreement corrected with the probability of agreement expected by chance.

The extent of a kappa-type coefficient is generally qualified by comparing its estimate against some threshold values [2, 15, 25].

Although commonly adopted by practitioners, this straightforward benchmarking procedure is criticized being based on a single summary measure of agreement that provides limited information. In order to overcome this criticism, researchers recommend to supplement the agreement coefficient with information on statistical uncertainty [24] and suggest the use of the lower confidence bound for agreement benchmarking purpose [26, 29].

The standard methods for building confidence intervals for kappa-type coefficients (e.g. [4, 14]) require large sample sizes [12]; when this condition is not satisfied, that is when the number of items is small (about 30 or less) or moderate (approximately 50 or less), the standard methods perform poorly and the use of bootstrap to build confidence intervals is recommended (e.g. [18, 23, 29]).

This paper focuses on two bootstrap confidence intervals (i.e. percentile and Bias-Corrected and Accelerated) and aims at investigating whether their lower bound can be effectively used to characterize the extent of agreement with small sample sizes, which are not uncommon in agreement studies. An extensive Monte Carlo simulation has been carried out under different scenarios taking into account both null (i.e. chance agreement) and non-null (i.e. positive agreement) inference cases. The performances of both bootstrap confidence intervals have been compared in order to recommend the method that best fits each specific scenario.

The remainder of this paper is organized as follows: in Sect. 2 the weighted Uniform kappa coefficient is introduced; Sect. 3 is devoted to coefficient estimation and inference; in Sect. 4 the simulation design is described and the main results are discussed; finally, conclusions are summarized in Sect. 5.

2 Weighted Uniform Kappa Coefficient

A main issue for the correct definition of a kappa-type coefficient regards the notion of expected proportion of agreement: chance measurements are conceived as blind (that is, uninformative about the rated items) and any distributional assumption for them is likely to be arbitrary. A solution is to adopt the notion of uniform chance measurement [3] that—given a certain rating scale—can be assumed as a reasonable model for the maximally non-informative measurement system. This uniform version of kappa-type coefficient is often referred to as Brennan–Prediger coefficient [6], although it was independently developed by several authors [16, 20, 22].

Table 1 Notation for cell counts in a $k \times k$ contingency table

		Second replication					
	Category	1	...	j	...	k	**Total**
First replication	1	n_{11}	...	n_{1j}	...	n_{1k}	$n_{1\cdot}$
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	i	n_{i1}	...	n_{ij}	...	n_{ik}	$n_{i\cdot}$
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	k	n_{k1}	...	n_{kj}	...	n_{kk}	$n_{k\cdot}$
	Total	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot k}$	n

Let n be the number of items rated two or more times (i.e. replications) on an ordinal rating scale with $k > 2$ classification categories. The data are denoted by Y_{lh}, with $l = 1, \ldots, n$ indexing items and h indexing replications. Of interest for the evaluation of inter/intra-rater agreement is the joint distribution of Y_{lh}.

In the simplest case of two replications (i.e. $h = 1, 2$), Y_{lh} can be arranged in a $k \times k$ contingency table $(n_{ij})_{k\times k}$ where the generic (i, j) cell contains the joint frequency n_{ij} that counts the number of items classified into ith category in the first replication and into jth category in the second replication (Table 1).

In order to consider that on an ordinal rating scale some disagreements are more serious than others (i.e. disagreement on two distant categories are more relevant than disagreement on neighbouring categories), it is necessary to assign a different weight to each proportion of disagreement.

The weighted version of the Uniform kappa coefficient, K_W^U, has been proposed by Gwet [17] and it is formulated as follows:

$$K_W^U = \frac{p_{a_W} - p_{a|c_W}^U}{1 - p_{a|c_W}^U} \tag{1}$$

where p_{a_W} is the probability of observed agreement and $p_{a|c_W}^U$ is the probability of agreement expected by chance. Let π_{ij} be the probability of classifying an item as belonging to category i during the first replication and to category j during the second replication and let w_{ij} be the symmetric (i.e. $w_{ij} = w_{ji}$) agreeing weight a priori assigned to each pair (i, j) of classification categories; p_{a_W} and $p_{a|c_W}^U$ are respectively formulated as:

$$p_{a_W} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} \pi_{ij} \tag{2}$$

$$p^U_{a|c_W} = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} \tag{3}$$

At sample level, Eq. 2 is estimated as follows:

$$\hat{p}_{a_W} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} \frac{n_{ij}}{n} \tag{4}$$

It is worthwhile to pinpoint that although the weights can be arbitrary defined, the most commonly used weighting schemes for kappa-type coefficients are the linear w^L_{ij} [8] and quadratic w^Q_{ij} [13] weights which are formulated as follows:

$$w^L_{ij} = 1 - \frac{|i-j|}{k-1}; \quad w^Q_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \tag{5}$$

All kappa-type coefficients range from -1 to $+1$: when the observed proportion of agreement equals chance agreement, the coefficient is null; when the observed agreement is greater than chance agreement the coefficient returns positive values; when the observed agreement is lower than chance agreement the coefficient takes negative values and it can be interpreted as disagreement.

The weighted Uniform kappa coefficient can be assumed asymptotically normally distributed with mean $\hat{K}^U_W$ and variance $\hat{\sigma}^2_{K^U_W}$ estimated by [17, p. 143]:

$$\hat{K}^U_W = \frac{\hat{p}_{a_W} - p^U_{a|c_W}}{1 - p^U_{a|c_W}} \tag{6}$$

$$\hat{\sigma}^2_{K^U_W} = \frac{1-f}{n(1-p^U_{a|c_W})^2} \left(\sum_{i=1}^{k} \sum_{j=1}^{k} w^2_{ij} \frac{n_{ij}}{n} - \hat{p}^2_{a_W} \right) \tag{7}$$

where $f = n/N$ is the sampling fraction of items from a target population of size N; when N is unknown, f is set equal to 0.

In the special case of missing ratings (i.e. some items were rated only during one replication) the variance is estimated as follows:

$$\hat{\sigma}^2_{K^U_W} = \frac{1-f}{n(1-p^U_{a|c_W})^2} \sum_{h=1}^{n} a^2_l \tag{8}$$

where l refers to the generic rated item and $a_l = \sum_{i,j=1}^{k} w_{ij} \left(\delta^{(l)}_{ij} - n_{ij}/n \right)$ with $\delta^{(l)}_{ij} = 1$ if item l is classified into ith and jth category in the first and second replication, respectively, and $\delta^{(l)}_{ij} = 0$ otherwise.

Without missing ratings, Eqs. 7 and 8 are equivalent.

Table 2 Landis and Koch benchmark scale for kappa-type coefficients

Coefficient	Agreement
$\hat{K}_W^U \leq 0.00$	Poor
$0.00 < \hat{K}_W^U \leq 0.20$	Slight
$0.20 < \hat{K}_W^U \leq 0.40$	Fair
$0.40 < \hat{K}_W^U \leq 0.60$	Moderate
$0.60 < \hat{K}_W^U \leq 0.80$	Substantial
$0.80 < \hat{K}_W^U \leq 1.00$	Almost perfect

3 Characterization of the Extent of Rater Agreement

The approach currently adopted to characterize the extent of agreement is based upon a straight comparison between the estimated coefficient and an adopted benchmark scale. The most widespread benchmark scale for interpreting the magnitude of agreement coefficients was proposed by Landis and Koch [25]. According to this scale, there are six categories of agreement (i.e. Poor, Slight, Fair, Moderate, Substantial and Almost perfect) corresponding to as many ranges of coefficient values (Table 2).

Although benchmark scales are widely adopted for relating the magnitude of the coefficient to the notion of extent of agreement (e.g. [1, 4, 5, 11, 19, 23, 28]), some researchers question their validity and give advice that their uncritical applications may lead to practically questionable decisions [27].

Actually, as argued in [17], the choice of the benchmark scale is less important than the way it is used for characterizing the extent of agreement. As a matter of fact, the straightforward benchmarking does not account for the influence of experimental conditions on the estimated coefficient and, thus, it does not allow for a statistical characterization of the extent of rater agreement. This criticism may be overcome by benchmarking the lower bound of the confidence interval of the agreement coefficient rather than its point estimate.

Assuming the asymptotic normal approximation, the lower $K_{W_l}^U$ and upper $K_{W_u}^U$ bounds of the two-sided $(1-2\alpha)\%$ confidence interval for K_W^U are given by:

$$K_{W_l}^U = \hat{K}_W^U - z_\alpha \hat{\sigma}_{K_W^U}; \quad K_{W_u}^U = \hat{K}_W^U + z_\alpha \hat{\sigma}_{K_W^U} \tag{9}$$

where z_α is the α percentile of the standard normal distribution.

The accuracy of the above confidence interval depends on the asymptotic normality of K_W^U and on the asymptotic solution for $\hat{\sigma}^2_{K_W^U}$ which are both questionable for small sample sizes.

Since resampling adjusts for non-normal distribution, it is generally considered the approach of choice when the assumptions of classical statistical methods are not met.

Among the available resampling methods to build confidence intervals, the percentile bootstrap (hereafter, p) is the simplest and the most popular one. The lower and upper bounds of the two-sided $(1-2\alpha)\%$ p confidence interval are given by:

$$K^U_{W_l} = G^{-1}(\alpha); \quad K^U_{W_u} = G^{-1}(1-\alpha) \tag{10}$$

where G is the bootstrap distribution function for of K^U_W.

On the other hand the Bias-Corrected and Accelerated bootstrap (hereafter, BCa) confidence interval is recommended for severely non normal data [7, 10]. Despite the high computational complexity needed, BCa confidence intervals have generally smaller coverage errors than the others. The lower and upper bounds of the two-sided $(1-2\alpha)\%$ BCa confidence interval are defined as follows:

$$\begin{aligned} K^U_{W_l} &= G^{-1}\left(\Phi\left(b - \frac{z_\alpha - b}{1 + a\,(z_\alpha - b)}\right)\right) \\ K^U_{W_u} &= G^{-1}\left(\Phi\left(b + \frac{z_\alpha + b}{1 + a\,(-z_\alpha - b)}\right)\right) \end{aligned} \tag{11}$$

being Φ the standard Gaussian distribution function, a the acceleration parameter and b the bias correction parameter. The parameters a is proportional to the skewness of the bootstrap distribution and is estimated via the jackknife resampling method; b is related to the proportion of bootstrap estimates that are less than the observed coefficient value. All computational details about the algorithm for building the bootstrap confidence intervals can be found in [7].

4 Simulation Study

In order to investigate the statistical properties of the proposed benchmarking procedure a Monte Carlo simulation study has been developed considering two replications (i.e. two different raters or the same rater in different occasions) of n items into one of k possible ordinal rating categories.

The performances of the benchmarking procedure have been evaluated in terms of statistical significance and power, computed for the cases of null and non null inference on rater agreement (Table 3). The null inference case tests the hypothesis that the rater agreement is positive against the null hypothesis of chance agreement; the non-null inference cases test the hypothesis that the rater agreement is at least Moderate against the null hypothesis of no more than Fair agreement as well as the hypothesis that the rater agreement is at least Substantial against the null hypothesis of no more than Moderate agreement. Specifically, for the null inference case, 5 alternative hypotheses of positive rater agreement (starting from $K^U_W = 0.50$ with step size 0.10) are tested against the hypothesis of chance agreement; for the first case of non-null inference (i.e. at least Moderate agreement), the above 5 alternative

Table 3 Null and non null inference cases

Inference case	H_0	H_1
Null	*Chance agreement*	*Positive agreement*
	$K_W^U = 0.00$	$K_W^U > 0.00$
Non Null	*No more than Fair*	*Moderate*
	$K_W^U \leq 0.40$	$K_W^U > 0.40$
	No more than Moderate	*Substantial*
	$K_W^U \leq 0.60$	$K_W^U > 0.60$

hypotheses are tested against the null hypothesis of no more than Fair agreement; for the second case of non-null inference (i.e. at least Substantial agreement), 4 alternative hypotheses (starting from $K_W^U = 0.70$ with step size 0.10) are tested against the null hypothesis of no more than Moderate agreement.

The statistical properties of the benchmarking procedure have been studied for four different rating scale dimensions (i.e. $k = 2, 3, 5, 7$) by sampling $r = 2000$ Monte Carlo data sets from a multinomial distribution with parameters n and π_{ij}; for each scale dimension the π_{ij} values (with $i, j = 1, \ldots, k$) have been chosen so as to obtain eight true population values of K_W^U (viz. 0, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.00), assuming a linear weighting scheme [8]. Since the simulation has been conducted by varying the sample size from $n = 10$ items to $n = 30$, $n = 50$ and $n = 100$ items (i.e. the most affordable sample sizes in many experimental contexts and also the most critical ones for statistical inference) a total number of $4 \cdot 8 \cdot 4 = 128$ scenarios have been analysed. For each scenario, both p and BCa confidence intervals have been built on 1500 bootstrap replications. The simulation algorithm has been implemented using Mathematica (Version 11.0, Wolfram Research, Inc., Champaign, IL, USA).

4.1 Simulation Results

Simulation results in terms of statistical significance and power are reported for 2, 3, 5 and 7-point scales in Tables 4, 5, 6 and 7, respectively. Specifically, each table contains the results obtained for all the analysed sample sizes and both bootstrap confidence intervals, organized in three distinct sections, corresponding to the three null hypotheses of rater agreement.

The statistical properties of the benchmarking procedure improve as sample size and rating scale dimension increase being satisfactory even for samples of $n = 10$ items and a few-point rating scale.

The statistical significance is generally comparable across p and BCa confidence intervals although it is sometimes slightly closer to its nominal level ($\alpha = 0.025$) when benchmarking the lower bound of the BCa confidence interval. Specifically, the statistical significance decreases with increasing sample size but it grows up for

Table 4 Statistical significance (in bold; $\alpha = 0.025$) and power for different true population values of K_W^U with $k = 2$ rating categories

		$n = 10$		$n = 30$		$n = 50$		$n = 100$	
		p	BCa	p	BCa	p	BCa	p	BCa
$K_W^U = 0.00$	$K_W^U =$ **0.00**	**0.03**	**0.03**	**0.02**	**0.02**	**0.02**	**0.03**	**0.02**	**0.02**
	$K_W^U = 0.50$	0.26	0.52	0.79	0.77	0.97	0.92	1.00	1.00
	$K_W^U = 0.60$	0.38	0.67	0.93	0.92	0.99	0.99	1.00	1.00
	$K_W^U = 0.70$	0.54	0.80	0.99	0.98	1.00	1.00	1.00	1.00
	$K_W^U = 0.80$	0.74	0.92	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.90$	0.91	0.99	1.00	1.00	1.00	1.00	1.00	1.00
$K_W^U = 0.40$	$K_W^U =$ **0.40**	**0.03**	**0.03**	**0.04**	**0.04**	**0.02**	**0.03**	**0.02**	**0.03**
	$K_W^U = 0.50$	0.06	0.13	0.09	0.24	0.25	0.22	0.20	0.23
	$K_W^U = 0.60$	0.11	0.36	0.27	0.32	0.35	0.39	0.63	0.67
	$K_W^U = 0.70$	0.19	0.53	0.51	0.56	0.71	0.74	0.94	0.95
	$K_W^U = 0.80$	0.34	0.72	0.82	0.86	0.95	0.95	1.00	1.00
	$K_W^U = 0.90$	0.59	0.90	0.98	0.98	1.00	1.00	1.00	1.00
$K_W^U = 0.60$	$K_W^U =$ **0.60**	**0.11**	**0.11**	**0.04**	**0.05**	**0.03**	**0.04**	**0.03**	**0.03**
	$K_W^U = 0.70$	0.19	0.19	0.16	0.16	0.17	0.22	0.25	0.29
	$K_W^U = 0.80$	0.34	0.34	0.42	0.42	0.54	0.61	0.80	0.84
	$K_W^U = 0.90$	0.59	0.59	0.81	0.82	0.93	0.96	1.00	1.00
	$K_W^U = 1.00$	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00

Table 5 Statistical significance (in bold; $\alpha = 0.025$) and power for different true population values of K_W^U with $k = 3$ rating categories

		$n = 10$		$n = 30$		$n = 50$		$n = 100$	
		p	BCa	p	BCa	p	BCa	p	BCa
$K_W^U = 0.00$	$K_W^U =$ **0.00**	**0.05**	**0.07**	**0.04**	**0.04**	**0.03**	**0.03**	**0.03**	**0.03**
	$K_W^U = 0.50$	0.50	0.56	0.91	0.91	0.99	0.99	1.00	1.00
	$K_W^U = 0.60$	0.75	0.81	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.70$	0.92	0.95	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.80$	0.97	0.98	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.90$	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
$K_W^U = 0.40$	$K_W^U =$ **0.40**	**0.09**	**0.10**	**0.05**	**0.04**	**0.04**	**0.04**	**0.03**	**0.03**
	$K_W^U = 0.50$	0.16	0.26	0.13	0.13	0.18	0.19	0.25	0.26
	$K_W^U = 0.60$	0.27	0.45	0.38	0.40	0.57	0.61	0.83	0.84
	$K_W^U = 0.70$	0.48	0.69	0.79	0.80	0.94	0.96	1.00	1.00
	$K_W^U = 0.80$	0.69	0.84	0.96	0.96	1.00	1.00	1.00	1.00
	$K_W^U = 0.90$	0.77	0.84	0.99	0.99	1.00	1.00	1.00	1.00
$K_W^U = 0.60$	$K_W^U =$ **0.60**	**0.10**	**0.10**	**0.04**	**0.05**	**0.03**	**0.03**	**0.04**	**0.03**
	$K_W^U = 0.70$	0.22	0.22	0.20	0.26	0.28	0.27	0.48	0.48
	$K_W^U = 0.80$	0.44	0.44	0.59	0.63	0.76	0.74	0.96	0.96
	$K_W^U = 0.90$	0.71	0.71	0.85	0.82	0.96	0.96	1.00	1.00
	$K_W^U = 1.00$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6 Statistical significance (in bold; $\alpha = 0.025$) and power for different true population values of K_W^U with $k = 5$ rating categories

		$n = 10$		$n = 30$		$n = 50$		$n = 100$	
		p	BCa	p	BCa	p	BCa	p	BCa
$K_W^U = 0.00$	$K_W^U =$ **0.00**	**0.07**	**0.07**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**
	$K_W^U = 0.50$	0.65	0.62	0.97	0.95	1.00	1.00	1.00	1.00
	$K_W^U = 0.60$	0.83	0.79	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.70$	0.87	0.84	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.80$	0.94	0.92	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.90$	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
$K_W^U = 0.40$	$K_W^U =$ **0.40**	**0.11**	**0.12**	**0.04**	**0.03**	**0.04**	**0.04**	**0.03**	**0.03**
	$K_W^U = 0.50$	0.20	0.21	0.17	0.15	0.21	0.19	0.35	0.34
	$K_W^U = 0.60$	0.39	0.40	0.50	0.44	0.65	0.62	0.90	0.89
	$K_W^U = 0.70$	0.51	0.51	0.74	0.70	0.90	0.87	1.00	1.00
	$K_W^U = 0.80$	0.64	0.65	0.93	0.91	0.99	0.99	1.00	1.00
	$K_W^U = 0.90$	0.88	0.88	1.00	1.00	1.00	1.00	1.00	1.00
$K_W^U = 0.60$	$K_W^U =$ **0.60**	**0.13**	**0.13**	**0.07**	**0.06**	**0.04**	**0.04**	**0.04**	**0.0**3
	$K_W^U = 0.70$	0.30	0.30	0.26	0.22	0.28	0.27	0.42	0.40
	$K_W^U = 0.80$	0.48	0.51	0.58	0.52	0.71	0.67	0.93	0.92
	$K_W^U = 0.90$	0.75	0.81	0.93	0.90	0.99	0.98	1.00	1.00
	$K_W^U = 1.00$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 7 Statistical significance (in bold; $\alpha = 0.025$) and power for different true population values of K_W^U with $k = 7$ rating categories

		$n = 10$		$n = 30$		$n = 50$		$n = 100$	
		p	BCa	p	BCa	p	BCa	p	BCa
$K_W^U = 0.00$	$K_W^U =$ **0.00**	**0.06**	**0.05**	**0.06**	**0.03**	**0.04**	**0.03**	**0.04**	**0.02**
	$K_W^U = 0.50$	0.60	0.55	0.94	0.92	1.00	1.00	1.00	1.00
	$K_W^U = 0.60$	0.81	0.75	1.00	0.99	1.00	1.00	1.00	1.00
	$K_W^U = 0.70$	0.93	0.91	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.80$	0.99	0.97	1.00	1.00	1.00	1.00	1.00	1.00
	$K_W^U = 0.90$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$K_W^U = 0.40$	$K_W^U =$ **0.40**	**0.09**	**0.08**	**0.06**	**0.05**	**0.04**	**0.03**	**0.04**	**0.03**
	$K_W^U = 0.50$	0.17	0.15	0.19	0.16	0.22	0.19	0.32	0.29
	$K_W^U = 0.60$	0.34	0.31	0.52	0.45	0.64	0.57	0.89	0.86
	$K_W^U = 0.70$	0.58	0.55	0.88	0.83	0.97	0.95	1.00	1.00
	$K_W^U = 0.80$	0.86	0.84	1.00	0.99	1.00	1.00	1.00	1.00
	$K_W^U = 0.90$	0.96	0.95	1.00	1.00	1.00	1.00	1.00	1.00
$K_W^U = 0.60$	$K_W^U =$ **0.60**	**0.16**	**0.16**	**0.07**	**0.05**	**0.06**	**0.05**	**0.05**	**0.04**
	$K_W^U = 0.70$	0.34	0.33	0.30	0.25	0.37	0.32	0.58	0.52
	$K_W^U = 0.80$	0.71	0.70	0.81	0.75	0.93	0.90	1.00	1.00
	$K_W^U = 0.90$	0.91	0.92	0.99	0.97	1.00	1.00	1.00	1.00
	$K_W^U = 1.00$	0.99	0.99	0.99	1.00	0.99	1.00	1.00	1.00

increasing true population value of K_W^U; it is always close to the nominal level in the null inference case and only for $n \geq 50$ in non-null inference cases.

The statistical power, instead, usually differs across p and BCa confidence intervals being slightly higher when benchmarking the lower bound of the BCa confidence interval; differences get smaller for increasing sample size and rating scale dimension, especially when testing high rater agreement level. For $n \leq 30$, the statistical power is less than 80% only when testing hypotheses referring to adjacent agreement categories (e.g. Poor vs Slight, Moderate vs Substantial) with $k > 3$. For $k = 2, 3$ and $n \leq 30$, the statistical power reaches 80% only when testing an almost perfect rater agreement level.

5 Conclusions

The proposed benchmarking procedure can be suitably applied for the characterization of the extent of agreement over a small or moderate number of items evaluated by one rater in two different occasions or simultaneously by more raters.

The procedure adopted for testing agreement shows satisfactory statistical properties also with small and moderate sample sizes and a few-point scale both in null and non-null inference cases. The procedure is adequately powered in detecting differences in the extent of rater agreement that are of practical interest for agreement studies (i.e. differences more than 0.2). For small samples of $n = 10$ items, benchmarking the lower bound of the BCa confidence interval is recommended; with $30 \leq n \leq 50$ the statistical power slightly differs between p and BCa confidence intervals, whereas for $n = 100$ the difference is about 1% even for a 2-point scale, therefore testing agreement via p confidence interval could be suggested because of the less computation burden.

Acknowledgements The authors express their gratitude to the anonymous reviewers for their positive comments and helpful suggestions which contributed to the improvement of this article.

References

1. Altaye, M., Donner, A., Eliasziw, M.: A general goodness-of-fit approach for inference procedures concerning the kappa statistic. Stat. Med. **20**(16), 2479–2488 (2001)
2. Altman, D.G.: Practical Statistics for Medical Research. CRC Press, Boca Raton (1990)
3. Bennett, E.M., Alpert, R., Goldstein, A.: Communications through limited-response questioning. Public Opin. Q. **18**(3), 303–308 (1954)
4. Blackman, N.J.M., Koval, J.J.: Interval estimation for Cohen's kappa as a measure of agreement. Stat. Med. **19**(5), 723–741 (2000)
5. Bland, J.: Measurement in health and disease. Cohen's Kappa. Department of Health Sciences, University of York, New YorK, UK (2008)
6. Brennan, R.L., Prediger, D.J.: Coefficient kappa: some uses, misuses, and alternatives. Educ. Psychol. Meas. **41**(3), 687–699 (1981)

7. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. **19**(9), 1141–1164 (2000)
8. Cicchetti, D.V., Allison, T.: A new procedure for assessing reliability of scoring EEG sleep recordings. Am. J. EEG Technol. **11**(3), 101–110 (1971)
9. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
10. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. CRC Press, Boca Raton (1994)
11. Everitt, B.S.: The Analysis of Contingency Tables. CRC Press, Boca Raton (1992)
12. Fleiss, J.L., Cicchetti, D.V.: Inference about weighted kappa in the non-null case. Appl. Psychol. Meas. **2**(1), 113–117 (1978)
13. Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ. Psychol. Meas. **33**(3), 613–619 (1973)
14. Fleiss, J.L., Cohen, J., Everitt, B.: Large sample standard errors of kappa and weighted kappa. Psychol. Bull. **72**(5), 323 (1969)
15. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical methods for rates and proportions. Wiley, New York (2013)
16. Guttman, L.: The test-retest reliability of qualitative data. Psychometrika **11**(2), 81–95 (1946)
17. Gwet, K.L.: Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC (2014)
18. Gwet, K.L.: Testing the difference of correlated agreement coefficients for statistical significance. Educ. Psychol. Meas. **76**(4), 609–637 (2016)
19. Hallgren, K.A.: Computing inter-rater reliability for observational data: an overview and tutorial. Tutor. Quant. Methods Psychol. **8**(1), 23 (2012)
20. Holley, J.W., Guilford, J.P.: A note on the G index of agreement. Educ. Psychol. Meas. **24**(4), 749–753 (1964)
21. International Organization for Standardization (ISO): Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 1: General Principles and Definitions (5725-1). Geneva, Switzerland (1994)
22. Janson, S., Vegelius, J.: On generalizations of the G index and the Phi coefficient to nominal scales. Multivar. Behav. Res. **14**(2), 255–269 (1979)
23. Klar, N., Lipsitz, S.R., Parzen, M., Leong, T.: An exact bootstrap confidence interval for κ in small samples. Journal of the Royal Statistical Society: Series D (The Statistician) **51**(4), 467–478 (2002)
24. Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M., Streiner, D.L.: Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Int. J. Nurs. Stud. **48**(6), 661–671 (2011)
25. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics, 159–174 (1977)
26. Roberts, C., McNamee, R.: A matrix of kappa-type coefficients to assess the reliability of nominal scales. Stat. Med. **17**(4), 471–488 (1998)
27. Sim, J., Wright, C.C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. **85**(3), 257–268 (2005)
28. Watson, P., Petrie, A.: Method agreement analysis: a review of correct methodology. Theriogenology **73**(9), 1167–1179 (2010)
29. Zapf, A., Castell, S., Morawietz, L., Karch, A.: Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med. Res. Methodol. **16**(1), 93 (2016)

Knowledge Based Methods

Bayesian Analysis of ERG Models for Multilevel, Multiplex, and Multilayered Networks with Sampled or Missing Data

Johan Koskinen, Chiara Broccatelli, Peng Wang and Garry Robins

Abstract Social network analysis has typically concerned analysis of one type of tie connecting nodes of the same type. It has however been recognised that people are connected through multiple types of ties and that people in addition are affiliated with multiple types of non-people nodes. Exponential random graph models (ERGM) is a family of statistical models for social networks that at this point allows for a number of different types of network data, including one-mode networks, bipartite networks, multiplex data, as well as multilevel network data. Multilevel networks have been proposed as a joint representation of associations between multiple types of entities or nodes, such as people and organization, where two types of nodes gives rise to three distinct types of ties. The typical roster data collection method may be impractical or infeasible when the node sets are hard to detect or define or because of the cognitive demands on respondents. Multilevel multilayered networks allow us to consider a multitude of different sources of data and to sample on different types of nodes and relations. We consider modelling multilevel multilayered networks using

J. Koskinen (✉) · G. Robins
Melbourne School of Psychological Sciences, The University of Melbourne,
Melbourne, Vic 3010, Australia
e-mail: johan.koskinen@unimelb.edu.au

G. Robins
e-mail: garrylr@unimelb.edu.au

J. Koskinen
The Institute of Analytical Sociology, Linköping University,
Linköping, Sweden

The Alan Turing Institute, British Library, 96 Euston Road,
London NW1 2DB, UK

C. Broccatelli
MRC/CSO Social and Public Health Sciences Unit, University of Glasgow,
200 Renfield Street, Glasgow G2 3QB 9PL, Scotland
e-mail: Chiara.Broccatelli@glasgow.ac.uk

P. Wang
Faculty of Business and Law, Centre for Transformative Innovation,
Swinburne University of Technology, Melbourne, Vic, Australia
e-mail: pengwang@swin.edu.au

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_9

exponential random graph models and extend a recently developed Bayesian data-augmentation scheme to allow for partially missing data. We illustrate the proposed inference procedures for the case of multilevel snowball sampling and sampling with error based on the Noordin Top network.

Keywords ERGM · Exchange algorithm · Missing data · Multilevel networks · Snowball sampling · Social networks

1 Introduction

Graphs have proven a powerful conceptualisation for studying social interaction among individuals. In social network analysis (SNA), graphs can be used to represent people as nodes, with edges connecting nodes that are deemed to be socially connected [44]. The use of SNA has not been limited to the study of ties among the same types of nodes, and there is a long tradition, e.g. [2], of studying how people are affiliated to other types of nodes in so-called two-mode networks. Similarly, there has been extensive analysis of networks of multiple relationships on the same types of nodes (e.g. [22, 46]). Kivelä et al. [10] coined the term 'multilayered networks' as a general framework for jointly designating multiple types of network data, such as one-mode, two-mode, and multiplex networks, where researchers had typically dealt with each instance separately. Lazega et al. [21] defined a 'multilevel network' as a network on two disjoint node sets, comprising three distinct types of ties.

The number of statistical models for modelling different types of network structures is ever growing [38]. Amongst these, exponential random graph models (ERGM) [24], are increasingly recognised as one of the most successful approaches for modelling network structure. Making explicit assumptions about the dependencies among tie-variables [29, 40], ERGM prescribes a log-linear, exponential family distribution with counts of different graph-configurations as their statistics. ERGMs have been defined for undirected [7, 39], and directed one-mode networks [7, 34]; two-mode networks [1, 40, 43]; multiplex networks [22, 28]; one-mode networks with valued ties [20, 32]; as well as fully multilevel networks [42, 45] and network panel data [16]. With increasing complexity of data, the risk of having missing or partially observed data increases, either because the network design mirrors that information has been pulled from many different sources (as in the example of Noordin Top used here), or that data collection becomes increasingly more difficult with many dimensions. ERGMs provide us not only with a principled framework for modelling dependence among tie-variables but by the same token it provides us with a coherent model for data that we may use to account for data imperfections like missing or non-sampled data. In situations where you are likely to have imperfect information on network ties, availing yourself of the full set of tools that may be

derived from a wider framework for networks may prove beneficial. Here we consider the extension of previous Bayesian data-augmentation techniques [17, 18] for partial or patchy multilevel multilayered data.

2 Data Structure

We assume two distinct set of nodes: $A = \{1, \ldots, n\}$ and $B = \{1, \ldots, m\}$ where we might observe ties among all combinations of nodes type. A tie thus belong to either of the sets $\binom{A}{2}$, $A \times B$, or $\binom{B}{2}$. For a set A, we use $\binom{A}{k}$ to denote the set $\{\{i_1, i_2, \ldots, i_k\} : i_1, i_2, \ldots, i_k \in A, i_1 < i_2 < \cdots < i_k\}$ of un-ordered k-tuples. In the sequel we will use AA, AB, and BB as a notational shorthand for these edge-sets, with the corresponding incidence matrices $\mathbf{X}_{AA}$, $\mathbf{X}_{AB}$, and $\mathbf{X}_{BB}$, respectively.

The element $(\mathbf{X}_E)_v$, or $X_{E,v}$ when it is unambiguous, of matrix $\mathbf{X}_E$ is equal to 1 if the edge $v \in E$ belongs to the graph and 0 otherwise. A multilevel network may be represented as a one-mode network with a blocked, symmetric adjacency matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{AA} & \mathbf{X}_{AB} \\ \mathbf{X}_{BA} & \mathbf{X}_{BB} \end{pmatrix}$$

When extending binary one-mode networks to multiple relations (say 'friendship' and 'advice') it is convention to represent this as a collection of graphs or adjacency matrices, one for each relation. For multilevel networks we by definition have different relations for different combinations of node-sets. Let the number of relations be denoted by R_E, for $E = AA, AB, BB$, with incidence matrices being defined as $\mathbf{X}_E^{(r)} = (X_{E,v}^{(r)})$, where $X_{E,v}^{(r)} = 1$ if there is a tie on relation $r = 0, \ldots, R_E - 1$ for edge-set $E = AA, AB, BB$. When the number of relations for $E = AA, AB, BB$ differ, we are not able to unambiguously define the multilayered network as a collection of one-mode network with blocked, symmetric adjacency matrices.

For AA, AB, and BB define the binary indicator matrices $\mathbf{D}_{AA}$, $\mathbf{D}_{AB}$, and $\mathbf{D}_{BB}$, each of which having elements $D_{E,v}$ of $\mathbf{D}_E$ equal to 1 or 0 depending on whether the corresponding tie-variable v is observed or not, respectively. For each $E = AA, AB, BB$ the indicators extend straightforwardly to account for more than one relation. Thus, for example, if $\mathbf{X}_{AA}^{(0)}$ represent friendship ties and $\mathbf{X}_{AA}^{(1)}$ represent advice ties, the corresponding matrices $\mathbf{D}_{AA}^{(0)}$ and $\mathbf{D}_{AA}^{(1)}$ would indicate what friendship and advice ties were observed and which ones were not observed.

We follow the convention [23] of partitioning data $\mathbf{X}$ into observed $\mathbf{X}^{\text{obs}} = \{X_v : D_v = 1\}$ and unobserved $\mathbf{X}^{\text{miss}} = \{X_v : D_v = 0\}$ data, conditional on an outcome $\mathbf{D}$. For a given $\mathbf{D}$ we take $(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}})$ to denote $\mathbf{X}$ reconstructed.

3 Model Formulation

Frank and Strauss [7] derived ERGMs for one-mode networks from the so called Markov dependence assumption that posited that for any two pairs $\{i, j\}$ and $\{k, \ell\}$ of vertices of a graph, the tie-variables $X_{i,j} \perp X_{k\ell} | X_{-(i,j),(k,\ell)}$ if $\{i, j\} \cap \{k, \ell\} = \emptyset$. They proved that the Markov dependence assumption implied a log-linear model for the collection of tie-variables that has as its sufficient statistics counts of different network 'configurations' (incidentally echoing the conclusions drawn by Moreno and Jennings [26]). Snijders et al. [39] elaborated on the Markov model by proposing parameters derived from the so called social circuit dependence assumption. The general form of ERGM is

$$p(\mathbf{X}|\theta) = \exp\{q(\mathbf{X}; \theta) - \psi(\theta)\}$$

where the normalising constant $\psi(\theta) = \sum_{\mathbf{Y} \in \mathcal{X}} \exp\{q(\mathbf{Y}; \theta)\}$ and $q(\mathbf{X}; \theta)$ is a potential dependent on the structure of the network and a $p \times 1$ vector $\theta \in \Theta = \{\theta \in \mathbb{R}^p : \psi(\theta) < \infty\}$ of statistical parameters. This general form is agnostic to the specific dependencies we may hypothesis for a particular type of network object. For undirected one-mode network, the model of Frank and Strauss [7] has the potential written as a weighted sum of sufficient graph statistics

$$q(\mathbf{X}; \theta) = \sum_{r}^{p_s} \theta_{S_r} \sum_{I \in A} \binom{X_{i+}}{r} + \theta_T \sum_{(i,j,k) \in \binom{A}{3}} X_{ij} X_{ik} X_{jk}$$

where the statistics correspond to two distinct categories of statistics, namely stars and triangles (in the expression $X_{i+} = \sum_j X_{ij}$). ERG models have been proposed for two-mode networks [1, 37, 43] and multiplex networks [28]. The modelling family has also been extended to the joint analysis of ties between different types of nodes [45] and for fully defined multilevel networks by Wang et al. [42]. Wang et al. [42] factor the function $q(\mathbf{X}; \theta)$

$$\begin{aligned} q(\mathbf{X}; \theta) = {} & \theta_{AA}^\top z(\mathbf{X}_{AA}) + \theta_{BB}^\top z(\mathbf{X}_{BB}) + \theta_{AB}^\top z(\mathbf{X}_{AB}) \\ & + \theta_{AA,BB}^\top z(\mathbf{X}_{AA}, \mathbf{X}_{BB}) + \theta_{AA,AB}^\top z(\mathbf{X}_{AA}, \mathbf{X}_{AB}) + \theta_{BB,AB}^\top z(\mathbf{X}_{BB}, \mathbf{X}_{AB}) \\ & + \theta_{AA,BB,AB}^\top z(\mathbf{X}_{AA}, \mathbf{X}_{BB}, \mathbf{X}_{AB}) \end{aligned}$$

to explicitly allow for different dependencies depending on what edge-sets are considered. For example, $z(\mathbf{X}_{AA})$ only involves statistics calculated on AA while $z(\mathbf{X}_{AA}, \mathbf{X}_{BB})$ involves crossed statistics, calculated for ties in $\binom{A}{2} \times \binom{B}{2}$. With multiple relations, statistics can be further partitioned, so that the linear predictors take into account dependencies between different types of ties between different types of nodes. Considering for example the interactions between ties in AA and AB, we

have

$$\theta_{AA,AB}^{\top} z(\mathbf{X}_{AA}, \mathbf{X}_{AB}) = \sum_{s=0}^{R_{AA}-1} \sum_{t=0}^{R_{AB}-1} \theta_{AA,AB,st}^{\top} z(\mathbf{X}_{AA}^{(s)}, \mathbf{X}_{AB}^{(t)})$$

The interpretation is that a tie of type s among pairs in AA may depend on affiliation of nodes in A with nodes in B of type t. For a recent treatment of properties of ERGM see [15, 19, 35, 36].

3.1 Observation Process

Conditional on a realisation $\mathbf{X}$, we assume an observation process

$$f(\mathbf{D}|\mathbf{X}, \zeta)$$

where the parameter ζ is distinct [23] from θ. The observation process may be thought of equivalently as a missing data generating mechanism or a sampling design, such as snowball sampling, for purposes of inference [9]. If we assume that tie-variables are observed conditionally independently conditional on $\mathbf{X}$, $f(\cdot)$ can be modelled as a regular log-linear model with a standard link function. Given that $\mathbf{D}$ has the same range-space $\mathcal{X}$ as $\mathbf{X}$, the observation indicators can also be modelled using an ERGM. Inference for an informative, missing not at random (MNAR) [23] process will however be contingent on informative priors.

4 Estimation

The posterior distribution of θ given $\mathbf{X}$ is doubly intractable as $\psi(\theta)$ in the likelihood as well as the normalising constant of the posterior are intractable. Markov chain Monte Carlo (MCMC) methods do not require that the normalising constant of the posterior is analytically tractable but need the likelihood to be available in closed form. Numerically, the likelihood can be evaluated by estimating the ratios of normalising constants [11] in each update of the parameters. Møller et al. [25] demonstrated that such an algorithm can be a proper MCMC even in the case of the importance sample being of size 1. It has been shown, however, that an improved sampler is required for ERGs [12]. Caimo and Friel [5] proposed a modified, approximate exchange algorithm [27] adopted to ERG models. This is an 'exact' MCMC when samples from the ERGM can be sampled perfectly [30].

We build on a recently proposed Bayesian data-augmentation scheme for doing inference for one-mode ERGM under the assumption of 'missing at random' (MAR) [18] (for a definition of MAR see [23]). A Markov chain Monte Carlo (MCMC) scheme is constructed by drawing from the joint posterior of $(\theta, \mathbf{X}^{\text{miss}}, \zeta)$ using

updating steps that update from $(\theta^{(t-1)}, \mathbf{X}^{\text{miss},(t-1)}, \zeta^{(t-1)})$ to $(\theta^{(t)}, \mathbf{X}^{\text{miss},(t)}, \zeta^{(t)})$. Conditional on $\mathbf{D}$, θ is updated using the approximate exchange sampler [5]:

(a) Draw η from $h(\eta|\theta^{(t-1)})$
(b) Draw $\mathbf{Y}$ from $p(\mathbf{Y}|\eta) = \exp\{q(\mathbf{Y};\eta) - \psi(\eta)\}$
(c) With probability $\min\{1, H\}$, set $\eta^{(t)} := \theta^{(t-1)}$ and $\theta^{(t)} := \eta$ where

$$\begin{aligned} H &= \frac{p(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss},(t-1)}|\eta)\pi(\eta)h(\theta^{(t-1)}|\eta)p(\mathbf{Y}|\theta^{(t-1)})}{p(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss},(t-1)}|\theta^{(t-1)})\pi(\theta^{(t-1)})h(\eta|\theta^{(t-1)})p(y|\eta)} \\ &= \exp\{q(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss},(t-1)};\eta) + q(\mathbf{Y};\theta^{(t-1)}) \\ &\quad -q(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss},(t-1)};\theta^{(t-1)}) - q(\mathbf{Y};\eta)\}\pi(\eta)/\pi(\theta^{(t-1)}) \end{aligned}$$

otherwise $\eta^{(t)} := \eta$ and $\theta^{(t)} := \theta^{(t-1)}$

In (a), $h(\cdot)$ is a symmetric proposal distribution, typically a multivariate Gaussian distribution. In the exchange sampler [27], updating steps (a) and (b) are performed by drawing directly from the conditional distributions in a Gibbs update. Generally for ERGM (b) will have to be performed through MCMC, meaning that the algorithm for drawing from the posterior is not a proper MCMC scheme. For a related auxiliary variable MCMC, convergence was monitored through running multiple, parallel, coupled chains [12], approximating a perfect sampler [30]. While Butts [4] demonstrate that a proper perfect sampling scheme may be constructed for ERGMs, mixing time may be considerably longer compared to common heuristics for determining burnin (see e.g. [14]). Here we chose the latter, in which case the MCMC scheme is approximate in the sense that (b) is not guaranteed to be a proper draw.

Koskinen et al. [18] propose to update $\mathbf{X}^{\text{miss}}$ under the assumption of missing at random (MAR) for one-mode networks. Whereas MAR implies

$$f(\mathbf{D}|\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}}, \zeta) = f(\mathbf{D}|\mathbf{X}^{\text{obs}}, \zeta),$$

we relax the assumption of MAR and allow for $\mathbf{D}$ to depend on all of $\mathbf{X}$. The modification of the updating-step for missing data is to draw $\mathbf{X}^{\text{miss}}$ given the rest from the full conditional posterior

$$\pi(\mathbf{X}^{\text{miss}}|\mathbf{X}^{\text{obs}}, \theta) = \frac{\exp\{q(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}};\theta) - \psi(\theta)\}f(\mathbf{D}|\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}}, \zeta)\pi(\zeta)}{\sum_{\mathbf{Y}^{\text{miss}}} \exp\{q(\mathbf{X}^{\text{obs}}, \mathbf{Y}^{\text{miss}};\theta) - \psi(\theta)\}f(\mathbf{D}|\mathbf{X}^{\text{obs}}, \mathbf{Y}^{\text{miss}}, \zeta)\pi(\zeta)}$$

The conditional distribution of $\zeta^{(t)}$ simplifies to a distribution proportional to $f(\mathbf{D}|\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}}, \zeta)\pi(\zeta)$. If the distribution $f(\cdot)$ is not fully tractable, draws of ζ cannot be made directly. Assuming that it is straightforward to draw $\mathbf{D}$ from $f(\cdot)$, ζ can be updated using steps (a), (b) and (c), with $f(\cdot)$ playing the role of $p(\cdot)$.

5 Empirical Illustration

We provide a brief empirical case-study using the 'Noordin Top' Terrorist Network [6]. The node set A consists of $n = 79$ individuals where $R_{AA} = 5$ for the relations 'classmates', 'communication', 'kinship', 'friendship', and 'soulmates'. The affiliation nodes B may be construed in a number of ways as a number of different types of affiliations are reported. We can treat the $R_{AB} = 9$ different types of affiliations as multiplex ties, with the interpretation that an affiliation node may be identified by the type of tie. For example, if $\mathbf{X}_{AB}^{(s)}$ represent people's affiliations to meetings, $\mathbf{X}_{AB}^{(t)}$ may represent people's religious affiliations and the number of columns for s and t do not necessarily need to be the same. A multilevel multilayered configuration could for example be the extent to which peoples' religious and meeting affiliations align with their friendships r, $\sum_{i<j}(\mathbf{X}_{AA}^{(r)})_{ij}\sum_k\sum_\ell(\mathbf{X}_{AB}^{(s)})_{ik}(\mathbf{X}_{AB}^{(s)})_{jk}(\mathbf{X}_{AB}^{(t)})_{i\ell}(\mathbf{X}_{AB}^{(t)})_{j\ell}$.

For the purposes of illustration we set $R_{AA} = 1$ and use the friendship ties reported in Everton for AA. We furthermore collapse the three relations meetings, training, and operations [31], creating a single relation $\mathbf{X}_{AB}$ with $m = 50$. To construct ties BB among events, we have elaborated on the time-stamped version of Broccatelli, Everett and Koskinen [3] and coded the explicitly mentioned connections between different events and operations in the International Crisis Group Report [8]. For the purposes of illustration, the event-by-event network is considered fixed and exogenous. Furthermore, we condition on the overall activity of the network, fixing the number of ties in both AA and BB. Consequently, all analyses have to be interpreted conditionally on the overall number of event participations and total number of friendship ties. We denote by $d_{i,E} = \sum_{j\neq i}(\mathbf{X}_E)_{ij}$ the degree of node i on the relation E. The resulting network is graphed in Fig. 1.

We fit a model to the Noordin Top network that has as statistics $z(\cdot)$ counts of the configurations in Fig. 2 (these are described in more detail in Wang et al. [41]). These statistics have the form:

(a) ASA: $\sum_{k=2}^{n-1}(-1)^k S_k/\lambda_s^{k-2}$, $S_k = \sum_{1\leq i\leq n}\binom{d_{i,AA}}{k}$.
(b) ATA: $3T_1 - T_2/\lambda_t + \cdots + (-1)^{n-3}T_{n-2}/\lambda_t^{n-3}$, $k = 2,\ldots,n-1$, $T_k = \sum_{i<j}(\mathbf{X}_{AA})_{ij}\binom{L_{2ij}}{k}$, $L_{2ij} = \sum_{h\neq i,j}(\mathbf{X}_{AA})_{ih}(\mathbf{X}_{AA})_{hj}$
(c) star2AX: $\sum_{1\leq i\leq n} d_{i,AA}d_{i,AB}$
(d) star2BX: $\sum_{1\leq j\leq m} d_{j,AB}d_{j,BB}$
(e) TriangleXBX: $\sum_{1\leq i\leq n}\sum_{1\leq j<h\leq m}(\mathbf{X}_{AB})_{ij}(\mathbf{X}_{AB})_{ih}(\mathbf{X}_{BB})_{jh}$
(f) XACB: $\sum_{1\leq i<j\leq n}\sum_{1\leq h<\ell\leq m}(\mathbf{X}_{AB})_{ih}(\mathbf{X}_{AB})_{jh}(\mathbf{X}_{AB})_{i\ell}(\mathbf{X}_{AB})_{j\ell}$

For the completely observed network, summaries of the posteriors for the corresponding parameters are provided in Table 2. Typical for one-mode network we find strong support for triadic closure (the 95% CI for ATA is (0.341, 1)) but also strong support for people taking part in events that are functionally related to other events that they take part in (the 95% CI for TriangleXBX is (0.786, 1.859)). For the affiliations, we note that being central in the one-mode friendship network is associated with taking part in many events (Star2AX). In addition, individuals tend to cluster

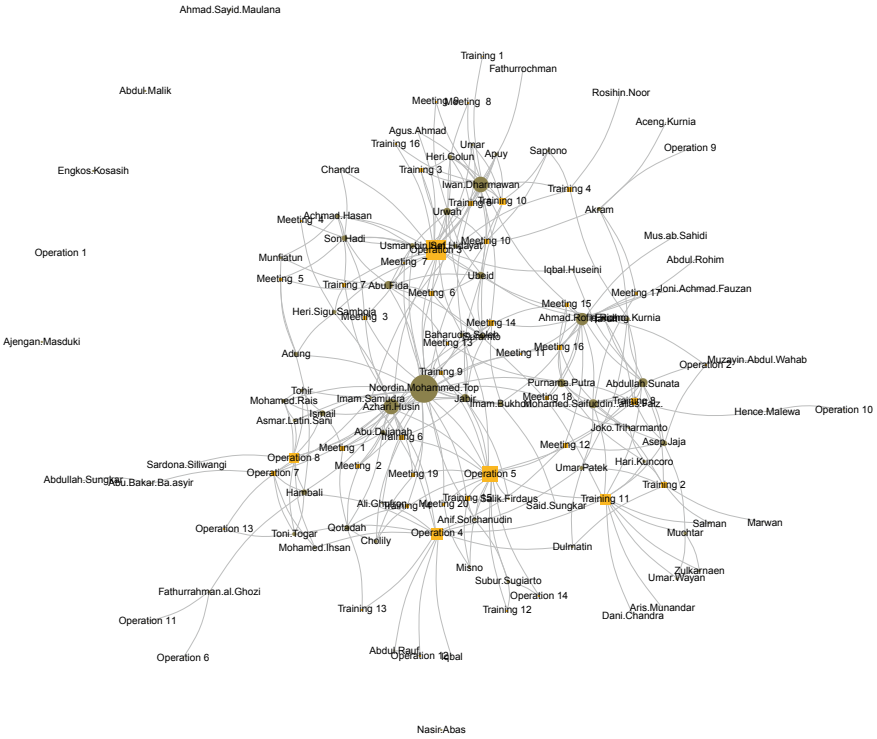
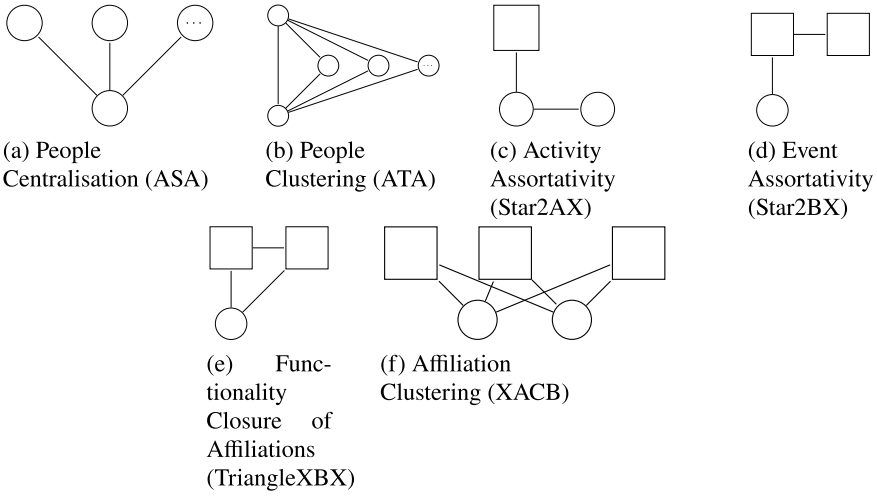


Fig. 1 The multilevel network of Noordin Top. Nodes are people and events. Node size is proportional to degree



(a) People Centralisation (ASA)

(b) People Clustering (ATA)

(c) Activity Assortativity (Star2AX)

(d) Event Assortativity (Star2BX)

(e) Functionality Closure of Affiliations (TriangleXBX)

(f) Affiliation Clustering (XACB)

Fig. 2 Configurations of multilevel ERGM for Noordin Top (configurations a, b, and f are geometrically weighted)

around events in four-cycles, something that might reflect recruitment processes [13, 33] or simply reflect team structure.

5.1 Multilevel Snowball Sampling

We now consider a hypothetical data collection scenario, where we aim to learn the ties of the network through multilevel snowball sampling. We snowball using Operation 3 as our seed (this is the 2004 Australian embassy bombing that took place on 9 September 2004 in Jakarta, Indonesia, killing 9–11 people and injuring more than 150 people). The tacit assumption in this hypothetical data collection scenario is that the observer starts exploring the network by recording anyone involved in this bombing. Anyone who participated in this operation is defined as being in wave 1, and anyone who is not in wave 1 but is tied to anyone in wave 1, belongs to wave 2. Conditional on $\mathbf{X}$, $\mathbf{D}$ is completely determined by the choice of seed node and non-sampled tie-variables are MAR [9, 17, 18].

The snowball sample obtained from using Operation 3 as seed node is illustrated in Fig. 3. Row 38 (from top) are the affiliations of Operation 3. We fit the model to the snowball sample as in Sect. 4, treating non-sampled ties as MAR. The results in Table 2 are qualitatively the same as for the model with completely observed data. Compared to the complete data posteriors, the sampling approach has increased uncertainty somewhat. For example, the 95% credibility interval for TriangleXBX is (0.786, 1.859) for the complete data and (0.769, 1.863) given the sampled data.

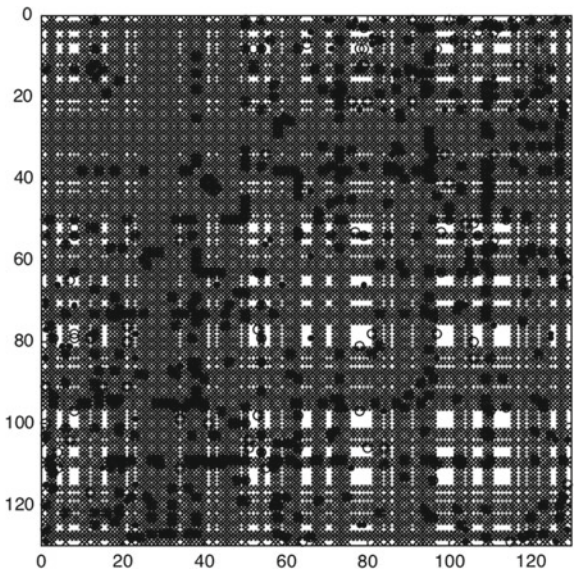


Fig. 3 Adjacency matrix (events: rows 1 trough 50; people: rows 51 through 129) for Noordin Top (filled dot indicating $(\mathbf{X})_{ij} = 1$). For snowball sample $(\mathbf{D})_{ij} = 1$ indicated by $\times$. Circles are ties predicted in one iteration of the MCMC

5.2 Sampling with Error

Another data collection scenario is that information gathering is focused on one or several important individuals. Again, a focal individual can serve as the seed node in a snowball sample, in which case the previous approach applies. Alternatively, we may assume that the analysts is gathering intelligence is such a way that individuals that are close to the focal actor are more visible that individuals that are far from the focal actor. For each tie-variable (i, j), we define independently $\Pr(D_{ij} = 1|\mathbf{X}, \zeta) = \Pr(D_{ij} = 1|h_{ij}(\mathbf{X}), \zeta)$, where $h_{ij}(\mathbf{X}) = \max\{d_i(\mathbf{X}), d_j(\mathbf{X})\}$, where $d_i(\mathbf{X})$ is the distance in $\mathbf{X}$ between $i \in A, B$ and Noordin Top (all ties in BB are assumed fixed and known). We model the probabilities $\Pr(D_{ij} = 1|h_{ij}(\mathbf{X}), \zeta)$ as in Table 1, with the interpretation that ties that are further from the leader Noordin Top are less visible that ties close to him.

In contrast to how $\mathbf{D}$ is defined in snowball sampling, the missingness mechanism now depends explicitly on unobserved data and is therefor MNAR. Consider for example the case of j being connected to a cut point i. The distance $h_{ij}(\mathbf{X})$ clearly depends on $(\mathbf{X})_{ij}$ and if $(\mathbf{D})_{ij} = 0$, then data are MNAR. (To simplify the situation, we do not remove ties between events B, as the event by event network is considered exogenous and fixed).

We fit a model under the assumption that we know probabilities in Table 1. The results in Table 2 indicate that effects corresponding to clustering is attenuated (the CI for ATA is $(0.341, 1)$ for complete data and $(-0.108, 0.727)$ for the MNAR case) but degree-related effects are amplified (with the exception of XASA). These changes are a natural consequence of the observation process respecting distance but not necessarily clustering.

Table 1 Detection bias in MNAR observation mechanism for Noordin Top

$h_{ij}(x)$	1	2	3	4	>4
$n_h(x)$	1122	6360	6090	1190	1750
$\Pr(D_{ij}\|x)$	0.99	0.75	0.5	0.25	0.15

Table 2 Posterior summaries for ERGM fitted to Noordin Top

Effect	No missing		Snowball sample		MNAR	
	Mean	Std	Mean	Std	Mean	Std
ASA	0.162	0.215	0.160	0.229	0.662	0.264
ATA	0.673	0.169	0.637	0.177	0.29	0.201
Star2AX	0.106	0.020	0.106	0.020	0.129	0.021
Star2BX	−0.014	0.046	0.000	0.049	0.022	0.06
TriangleXBX	1.322	0.273	1.299	0.278	1.191	0.293
XASA	0.185	0.205	0.337	0.213	0.037	0.212
XACB	0.106	0.029	0.091	0.035	0.069	0.046

6 Conclusions and Future Directions

We have proposed a statistical approach for analysing the structure of multilevel multilayered networks that account for imperfections in data. We provide an illustrative example of analysis of a multilevel network for three types of observation processes. While the approach is consistent when the observation process is known, a MNAR process requires making a number of untestable assumptions and is most likely of use merely as a sensitivity analysis. Further work is needed in order to systematically investigate the sensitivity of MNAR to different plausible MNAR mechanisms.

Acknowledgements The work of Koskinen and Broccatelli is funded by the Leverhulme Trust Grant RPG-2013-140.

References

1. Agneessens, F., Roose, H.: Local structural properties and attribute characteristics in 2-mode networks: p^* models to map choices of theater events. Journal of Mathematical Sociology **32**(3), 204–237 (2008)
2. Breiger, R.L.: The duality of persons and groups. Soc. Forces **53**, 181–190. http://www.jstor.org/stable/10.2307/2576011 (1974)
3. Broccatelli, C., Everett, M., Koskinen, J.: Temporal dynamics in covert networks. Methodol. Innov. **9**, 1–14 (2016)
4. Butts, C.T.: A perfect sampling method for exponential family random graph models. J. Math. Sociol. 1–20 (2017)
5. Caimo, A., Friel, N.: Bayesian inference for exponential random graph models. Soc. Netw. **33**, 41–55 (2011)
6. Everton, S.F.: Disrupting Dark Networks. Cambridge University Press, New York (2012)
7. Frank, O., Strauss, D.: Markov graphs. J. Am. Stat. Assoc. **81**, 832–842 (1986)
8. Group, I.C.: Terrorism in indonesia: Noordin's networks (2006)
9. Handcock, M.S., Gile, K.: Modeling social networks from sampled data. Ann. Appl. Stat. **4**, 5–25 (2010)
10. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. J. Complex Netw. **2**(3), 203–271 (2014)
11. Koskinen, J.: Bayesian analysis of exponential random graphs-estimation of parameters and model selection. Technical Report, Research Report 2004: 2, Department of Statistics, Stockholm University (2004)
12. Koskinen, J.: The linked importance sampler auxiliary variable Metropolis Hastings algorithm for distributions with intractable normalising constants. Technical Report 1, Department of Psychology, University of Melbourne (2008). (Working paper)
13. Koskinen, J., Edling, C.: Modelling the evolution of a bipartite network - peer referral in interlocking directorates. Soc. Netw. **34**, 309–322 (2012). https://doi.org/10.1016/j.socnet.2010.03.001
14. Koskinen, J., Snijders, T.: Simulation, estimation and goodness of fit. In: Lusher, D., Koskinen, J., Robins, G. (eds.) Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications, pp. 141–166. Cambridge University Press, Cambridge (2013)
15. Koskinen, J.H.: Exponential random graph models. In: Wiley StatsRef: Statistics Reference Online, stat08136. Wiley (in press)

16. Koskinen, J.H., Caimo, A., Lomi, A.: Simultaneous modeling of initial conditions and time heterogeneity in dynamic networks: an application to foreign direct investments. Netw. Sci. **3**(1), 58–77 (2015)
17. Koskinen, J.H., Robins, G.L., Pattison, P.E.: Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. Stat. Methodol. **7**, 366–384 (2010)
18. Koskinen, J.H., Robins, G.L., Wang, P., Pattison, P.E.: Bayesian analysis for partially observed network data, missing ties, attributes and actors. Soc. Netw. **35**(4), 514–527 (2013)
19. Koskinen, J.H., Wang, P., Robins, G.L., Pattison, P.E.: Outliers and influential observations in exponential random graph models. Psychometrika **83**(4), 809–830 (2018)
20. Krivitsky, P.N.: Exponential-family random graph models for valued networks. Electron. J. Stat. **6**, 1100 (2012)
21. Lazega, E., Jourda, M.T., Mounier, L., Stofer, R.: Catching up with big fish in the big pond? Multi-level network analysis through linked design. Soc. Netw. **30**(2), 159–176 (2008)
22. Lazega, E., Pattison, P.E.: Multiplexity, generalized exchange and cooperation in organizations: a case study. Soc. Netw. **21**(1), 67–90 (1999)
23. Little, R., Rubin, D.: Statistical Analysis with Missing Data. Wiley, New York (1987)
24. Lusher, D., Koskinen, J., Robins, G.: Exponential Random Graph Models. Cambridge University Press, Cambridge (2013)
25. Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K.: An efficient markov chain monte carlo method for distributions with intractable normalising constants. Biometrika **93**(2), 451–458 (2006)
26. Moreno, J., Jennings, M.: Statistics of social configurations. Sociometry **1**(3), 342–374 (1938)
27. Murray, I., Ghahramani, Z., MacKay, D.J.C.: MCMC for doubly-intractable distributions. In: Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence, pp. 359–366. AUAI Press, Corvallis (2006)
28. Pattison, P., Wasserman, S.: Logit models and logistic regressions for social networks: II. Multivariate relations. Br. J. Math. Stat. Psychol. **52**, 169–193 (1999)
29. Pattison, P.L., Snijders, T.: Modelling social networks: next steps. In: Lusher, D., Koskinen, J., Robins, G. (eds.) Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications, pp. 287–301. Cambridge University Press, Cambridge (2013)
30. Propp, J., Wilson, D.: Exact sampling with coupled markov chains and applications to statistical mechanics. Random Struct. Algorithms **9**, 232–252 (1996)
31. Roberts, N., Everton, S.F.: Strategies for combating dark networks. J. Soc. Struct. **12**(2), (2011)
32. Robins, G., Pattison, P., Wasserman, S.: Logit models and logistic regressions for social networks, III. Valued relations. Psychometrika **64**, 371–394 (1999)
33. Robins, G.L., Alexander, M.: Small worlds among interlocking directors: network structure and distance in bipartite graphs. Comput. Math. Organ. Theory **10**, 69–94 (2004)
34. Robins, G.L., Pattison, P.E., Wang, P.: Closure, connectivity and degree distributions: Exponential random graph (p^*) models for directed social networks. Soc. Netw. **31**, 105–117 (2009)
35. Schweinberger, M.: Instability, sensitivity, and degeneracy of discrete exponential families. J. Am. Stat. Assoc. **106**(496), 1361–1370 (2011)
36. Schweinberger, M., Krivitsky, P.N., Butts, C.T., Stewart, J.: Exponential-family models of random graphs: Inference in finite-, super-, and infinite-population scenarios. Available at https://arxiv.org/abs/1707.04800 (2018)
37. Skvoretz, J., Faust, K.: Logit models for affiliation networks. Sociol. Methodol. **29**(1), 253–280 (1999)
38. Snijders, T.A.: Statistical models for social networks. Annu. Rev. Sociol. **37**, (2011)
39. Snijders, T.A.B., Pattison, P.E., Robins, G.L., Handcock, M.S.: New specifications for exponential random graph models. Sociol. Methodol. **36**, 99–153 (2006)
40. Wang, P., Pattison, P., Robins, G.: Exponential random graph model specifications for bipartite networks - a dependence hierarchy. Soc. Netw. **35**(2), 211–222 (2013)
41. Wang, P., Robins, G., Pattison, P., Koskinen, J.H.: MPNet: Program for the simulation and estimation of (p*) exponential random graph models for multilevel networks. Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne (2014)

42. Wang, P., Robins, G., Pattison, P., Lazega, E.: Exponential random graph models for multilevel networks. Soc. Netw. **35**, 96–115 (2013)
43. Wang, P., Sharpe, K., Robins, G.L., Pattison, P.E.: Exponential random graph (p*) models for affiliation networks. Soc. Netw. **31**(1), 12–25 (2009)
44. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
45. Wasserman, S., Iacobucci, D.: Statistical modelling of one-mode and two-mode networks: simultaneous analysis of graphs and bipartite graphs. Br. J. Math. Stat. Psychol. **44**(1), 13–43 (1991)
46. White, H.C., Boorman, S.A., Breiger, R.L.: Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. Am. J. Sociol. **81**(4), 730–780 (1976)

Bayesian Kantorovich Deconvolution in Finite Mixture Models

Catia Scricciolo

Abstract This chapter addresses the problem of recovering the mixing distribution in finite kernel mixture models, when the number of components is unknown, yet bounded above by a fixed number. Taking a step back to the historical development of the analysis of this problem within the Bayesian paradigm and making use of the current methodology for the study of the posterior concentration phenomenon, we show that, for general prior laws supported over the space of mixing distributions with at most a fixed number of components, under replicated observations from the mixed density, the mixing distribution is estimable in the Kantorovich or L^1-Wasserstein metric at the optimal pointwise rate $n^{-1/4}$ (up to a logarithmic factor), n being the sample size.

Keywords Dirichlet distribution · Kantorovich metric · Kolmogorov metric · Mixing distribution · Mixture model · Posterior distribution · Rate of convergence · Sieve prior · Wasserstein metric

1 Introduction

The Bayesian analysis of the problem of recovering the unknown mixing distribution in mixture models has recently attracted much attention and stimulated an active discussion encouraging new ideas. Several papers–including [Efron [4], Gao and van der Vaart [5], Heinrich and Kahn [9], Ishwaran et al. [11], Nguyen [14], Scricciolo [18]]–have been devoted to the investigation of this topic, with extensive comparisons with the frequentist solutions. In order to introduce the problem, suppose that $x \mapsto k(x \mid y)$ is a probability density for every $y \in \mathscr{Y} \subseteq \mathbb{R}$, where $(\mathscr{Y}, \mathscr{B})$ is a measurable space. If the mapping $(x, y) \mapsto k(x \mid y)$ is jointly measurable, then

C. Scricciolo (✉)
Dipartimento di Scienze Economiche, Università degli Studi di Verona, Via Cantarane 24, 37129 Verona (VR), Italy
e-mail: catia.scricciolo@univr.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_10

$$p_G(x) := \int_{\mathscr{Y}} k(x \mid y)\, \mathrm{d}G(y), \quad x \in \mathbb{R}, \tag{1}$$

defines a probability density on $\mathbb{R}$ for every probability measure G on $(\mathscr{Y}, \mathscr{B})$, whose collection is indicated by $\mathscr{G}$. The cumulative distribution function of the mixed density in (1) is denoted by

$$F_G(x) = \int_{-\infty}^{x} p_G(u)\, \mathrm{d}u, \quad x \in \mathbb{R}.$$

Suppose we observe n independent random variables $X_1, \ldots, X_n$ identically distributed according to the mixed density

$$p_0(x) \equiv p_{G_0}(x) = \int_{\mathscr{Y}} k(x \mid y)\, \mathrm{d}G_0(y), \quad x \in \mathbb{R}.$$

We denote by F_0 the cumulative distribution function of the density p_0, namely,

$$F_0(x) \equiv F_{G_0}(x) = \int_{-\infty}^{x} p_0(u)\, \mathrm{d}u, \quad x \in \mathbb{R}.$$

The interest is in recovering the unknown mixing distribution $G_0 \in \mathscr{G}$ from observations of the random sample $X^{(n)} := (X_1, \ldots, X_n)$. The formulation of the problem applies to both finite and infinite mixtures, but the focus of this chapter is primarily on the case when the sampling density is a mixture with an unknown, but bounded above number of components.

The problem has been initially studied from the frequenstist perspective by Chen [1], who established that, when p_0 has an unknown number of components d_0 such that $1 \leq d_0 \leq N$, for some fixed integer N, then the optimal rate for estimating the mixing distribution G_0 is only $n^{-1/4}$ and this rate is achievable, under identifiability conditions, by some minimum distance estimator. Even if Theorem 2 in Chen [1], p. 226, is not correct because of Lemma 2 it relies on, an emended version of Lemma 2 has been recently given by Heinrich and Kahn [9] in assertion (21) of their Theorem 6.3, p. 2857, by comparing a fixed mixture with all the mixtures having mixing distributions in an L^1-Wasserstein ball, instead of comparing all possible pairs of mixtures in a ball. As a consequence, Theorem 2 of Chen [1] remains valid by dropping uniformity over an L^1-Wasserstein ball and the statement is weakened to an assertion on the optimal pointwise rate of estimation: for any fixed mixing distribution, say G_0, the minimum distance estimator converges at $n^{-1/4}$-rate, but with a multiplicative constant that may depend on G_0. The first Bayesian analysis of the problem we are aware of traces back to Ishwaran et al. [11], who define a prior law over the space of all mixing distributions with at most N components, the mixture weights being assigned an N-dimensional Dirichlet distribution with a non-informative choice for the shape parameters that are all set equal to α/N for a positive constant α. Under conditions similar to those postulated by Chen [1], which, in particular, employ the

notion of strong identifiability in mixture models, they prove that Bayesian estimation of the mixing distribution in the Kantorovich metric is possible at the optimal rate $n^{-1/4}$, up to a $\log n$-factor. More recently, posterior convergence rates for estimating the mixing distribution in the L^2-Wasserstein metric for finite mixtures of multivariate distributions have been discussed by Nguyen [14] following a different line of reasoning. In this chapter, we show that, by combining the approach of Ishwaran et al. [11], which instrumentally uses posterior contraction rates in the sup-norm for the distribution function and strong identifiability to shift to the Kantorovich distance between mixing distributions, with the current methodology for the study of posterior contraction rates, which can by now count upon many refined results for small ball prior probability estimates, the mixing distribution is estimable in the Kantorovich or L^1-Wasserstein metric at the optimal rate $n^{-1/4}$ (up to a logarithmic factor) for a large class of prior laws over the space of mixing distributions with at most N components, under less stringent conditions than those used in Ishwaran et al. [11] or in Nguyen [14]. Many aspects of this fundamental statistical problem still remain unclear and we hope to contribute to a better understanding of it in a follow-up study.

Before introducing the notation, a remark on the use of the term "Bayesian deconvolution" is in order. This phrase has been recently introduced by Efron [4] to describe a maximum likelihood procedure for estimating the mixing distribution in general mixture models of the form in (1). Even if the mixtures herein considered are not necessarily convolution kernel mixtures, we liked the evocative power of the expression to recall the general inverse problem of recovering the unknown mixing distribution.

Notation. In this paragraph, we set out the notation and recall some definitions used throughout the chapter.

- The symbols "$\lesssim$" and "$\gtrsim$" indicate inequalities valid up to a constant multiple that is universal or fixed within the context, but anyway inessential for our purposes.
- All probability density functions are meant to be with respect to Lebesgue measure λ on $\mathbb{R}$ or on some subset thereof.
- The same symbol, say G, is used to denote a probability measure on $(\mathscr{Y}, \mathscr{B})$ as well as the corresponding cumulative distribution function.
- The degenerate probability distribution putting mass one at a point $y \in \mathbb{R}$ is denoted by δ_y.
- The notation Pf stands for the expected value $\int f \, \mathrm{d}P$, where the integral is understood to extend over the entire natural domain when, here and elsewhere, the domain of integration is omitted. With this convention, for the empirical measure $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$ associated with the random sample $X_1, \ldots, X_n$, namely, the discrete uniform distribution on the sample values that puts mass $1/n$ on each one of the observations, the notation $\mathbb{P}_n f$ abbreviates the formula $n^{-1} \sum_{i=1}^n f(X_i)$.
- For every pair $\mathbf{x}_N, \mathbf{y}_N \in \mathbb{R}^N$, $\|\mathbf{x}_N - \mathbf{y}_N\|_{\ell^1}$ stands for the ℓ^1-distance $\sum_{j=1}^N |x_j - y_j|$.
- For a probability measure Q on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, let q denote its density. For any $\epsilon > 0$,

$$B_{\mathrm{KL}}(P_0;\,\epsilon^2) := \left\{ Q :\, P_0\left(\log\frac{p_0}{q}\right) \le \epsilon^2,\quad P_0\left(\log\frac{p_0}{q}\right)^2 \le \epsilon^2 \right\}$$

denotes a *Kullback-Leibler* type neighborhood of P_0 of radius ϵ^2. Defined, for every $\alpha \in (0,\,1]$, the divergence $\rho_\alpha(P_0\|Q) := (1/\alpha)[P_0(p_0/q)^\alpha - 1]$, see Wong and Shen [21], pp. 351–352,

$$B_{\rho_\alpha}(P_0;\,\epsilon^2) := \left\{ Q :\, \rho_\alpha(P_0\|Q) \le \epsilon^2 \right\}$$

is the ρ_α-neighborhood of P_0 of radius ϵ^2. The definition of ρ_α extends to negatives values of $\alpha \in (-1,\,0)$. In particular, for $\alpha = -1/2$, the divergence $\rho_{-1/2}(P_0\|Q) = -2\int p_0[(q/p_0)^{1/2} - 1]\,\mathrm{d}\lambda = \int (p_0^{1/2} - q^{1/2})^2\,\mathrm{d}\lambda$ is the squared Hellinger distance. We can thus define the following *Hellinger* type neighborhood of P_0 of radius ϵ^2:

$$B_{\rho_{-1/2}\|\cdot\|_\infty}(P_0;\,\epsilon^2) := \left\{ Q :\, \rho_{-1/2}(P_0\|Q) \left\| \frac{p_0}{q} \right\|_\infty \le \epsilon^2 \right\}.$$

– For any real number $p \ge 1$ and any pair of probability measures $G_1,\, G_2 \in \mathscr{G}$ with finite pth absolute moments, the *L^p-Wasserstein distance* between G_1 and G_2 is defined as

$$W_p(G_1,\,G_2) := \left(\inf_{\gamma\in\Gamma(G_1,\,G_2)} \int_{\mathscr{Y}\times\mathscr{Y}} |y_1 - y_2|^p\,\gamma(\mathrm{d}y_1,\,\mathrm{d}y_2) \right)^{1/p},$$

where $\Gamma(G_1,\,G_2)$ is the set of all joint probability measures on $(\mathscr{Y}\times\mathscr{Y}) \subseteq \mathbb{R}^2$, with marginal distributions G_1 and G_2 on the first and second arguments, respectively.

2 Main Results

This section is devoted to expose the main results of the chapter and is split into two parts. In the first one, preliminary results on Bayesian estimation of distribution functions in the Kolmogorov metric, which are valid for a large class of prior laws, are presented and some issues highlighted. In the second part, arguably the most relevant, attention is restricted to finite mixtures with an unknown, but bounded above number of components and Bayesian estimation of the mixing distribution in the Kantorovich metric at the optimal rate $n^{-1/4}$ (up to a logarithmic factor) is discussed.

Posterior Concentration of Kernel Mixtures in the Kolmogorov Metric

The following assumption will be hereafter in force.

Assumption **A**. Let

$$\epsilon_n := \left(\frac{\log n}{n}\right)^{1/2} L_n, \quad n \in \mathbb{N}, \tag{2}$$

where, depending on the prior concentration rate on small balls around P_0, the sequence of positive real numbers (L_n) can be either slowly varying at $+\infty$ or degenerate at an appropriate constant L_0.

Comments on the two possible specifications of (L_n) in connection with the prior concentration rate are postponed to Lemma 1, which provides sufficient conditions on the distribution function F_0 and the prior concentration rate ϵ_n for the posterior to contract at a nearly $\sqrt{n}$-rate on Kolmogorov neighborhoods of F_0. We warn the reader that, unless otherwise specified, in all stochastic order symbols used hereafter, the probability measure $\mathbf{P}$ is understood to be P_0^n, the joint law of the first n coordinate projections of the infinite product probability measure P_0^∞. Also, Π_n stands for a prior law, possibly depending on the sample size, over the space of probability measures $\{P_G,\ G \in \mathscr{G}\}$, with density p_G as defined in (1).

Lemma 1 *Let F_0 be a continuous distribution function. If, for a constant $C > 0$ and a sequence ϵ_n as defined in (2), we have*

$$\Pi_n(B_{\mathrm{KL}}(P_0;\ \epsilon_n^2)) \gtrsim \exp(-Cn\epsilon_n^2), \tag{3}$$

then, for $M_n \gtrsim \sqrt{(C + 1/2)}L_n$,

$$\Pi_n\left(\sqrt{n}\sup_x |(F_G - F_0)(x)| > M_n(\log n)^{1/2} \mid X^{(n)}\right) = o_{\mathbf{P}}(1). \tag{4}$$

Proof The posterior probability of the event

$$A_n^c := \left\{G :\ \sqrt{n}\sup_x |(F_G - F_0)(x)| > M_n(\log n)^{1/2}\right\}$$

is given by

$$\Pi_n(A_n^c \mid X^{(n)}) = \frac{\int_{A_n^c} \prod_{i=1}^n p_G(X_i)\, \Pi_n(\mathrm{d}G)}{\int_{\mathscr{G}} \prod_{i=1}^n p_G(X_i)\, \Pi_n(\mathrm{d}G)}.$$

We construct (a sequence of) tests (ϕ_n) for testing the hypothesis

$$H_0 :\ P = P_0 \quad \textit{versus} \quad H_1 :\ P = P_G,\ G \in A_n^c,$$

where $\phi_n :\equiv \phi_n(X^{(n)};\ P_0) :\ \mathscr{X}^n \to \{0,\ 1\}$ is the indicator function of the rejection region of H_0, such that

$$P_0^n\phi_n \to 0 \quad \text{as } n \to +\infty$$

$$\text{and} \quad \sup_{G \in A_n^c} P_G^n(1-\phi_n) \le 2\exp\left(-2(M_n - K)^2 \log n\right) \text{ for sufficiently large } n,$$

with a finite constant $K > 0$ and a sequence $M_n > K$ for every n large enough. Define the test

$$\phi_n := 1_{R_n}, \quad \text{with} \quad R_n := \left\{x^{(n)} : \sqrt{n} \sup_x |(F_n - F_0)(x)| > K(\log n)^{1/2}\right\},$$

where F_n is the empirical distribution function, that is, the distribution function associated with the empirical probability measure $\mathbb{P}_n$ of the sample $X^{(n)}$. Since $x \mapsto F_0(x)$ is *continuous* by assumption, in virtue of the Dvoretzky–Kiefer–Wolfowitz [3] (DKW for short) inequality, with the tight universal constant in Massart [13], the type I error probability $P_0^n\phi_n$ can be bounded above as follows

$$P_0^n\phi_n = P_0^n(R_n) \le 2\exp\left(-2K^2 \log n\right).$$

Then,

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})\phi_n] \le P_0^n\phi_n \le 2\exp\left(-2K^2 \log n\right), \tag{5}$$

where E_0^n denotes expectation with respect to P_0^n, and

$$\begin{aligned} E_0^n[\Pi_n(A_n^c \mid X^{(n)})] &= E_0^n[\Pi_n(A_n^c \mid X^{(n)})\phi_n] + E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)] \\ &\le 2\exp\left(-2K^2 \log n\right) + E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)]. \end{aligned}$$

It remains to control the term $E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)]$. Defined the set

$$D_n := \left\{x^{(n)} : \int_{\mathscr{G}} \prod_{i=1}^n \frac{p_G}{p_0}(x_i)\, \Pi_n(\mathrm{d}G) \le \Pi_n(B_{\mathrm{KL}}(P_0;\, \epsilon_n^2))\exp\left(-(C+1)n\epsilon_n^2\right)\right\},$$

consider the following decomposition

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)] = E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)(1_{D_n} + 1_{D_n^c})].$$

It is known from Lemma 8.1 of Ghosal et al. [7], p. 524, that $P_0^n(D_n) \le (C^2 n\epsilon_n^2)^{-1}$. It follows that

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)1_{D_n}] \le P_0^n(D_n) \le (C^2 n\epsilon_n^2)^{-1}. \tag{6}$$

By the assumption in (3) and Fubini's theorem,

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)1_{D_n^c}] \lesssim \exp\left((2C+1)n\epsilon_n^2\right) \int_{A_n^c} P_G^n(1-\phi_n)\,\Pi_n(\mathrm{d}G). \tag{7}$$

The following arguments are aimed at finding an exponential upper bound on $\sup_{G\in A_n^c} P_G^n(1-\phi_n)$. By the triangular inequality, over the set R_n^c, for every $G \in A_n^c$,

$$\begin{aligned} M_n(\log n)^{1/2} &< \sqrt{n}\sup_x |(F_G - F_0)(x)| \\ &\le \sqrt{n}\sup_x |(F_G - F_n)(x)| + \sqrt{n}\sup_x |(F_n - F_0)(x)| \\ &\le \sqrt{n}\sup_x |(F_G - F_n)(x)| + K(\log n)^{1/2}, \end{aligned}$$

which implies that

$$\sqrt{n}\sup_x |(F_G - F_n)(x)| > (M_n - K)(\log n)^{1/2}.$$

Since $x \mapsto F_G(x) := \int_{-\infty}^x p_G(u)\,\mathrm{d}u$ is continuous, by applying again the DKW's inequality, we obtain that

$$\begin{aligned} \sup_{G\in A_n^c} P_G^n(1-\phi_n) &\le \sup_{G\in A_n^c} P_G^n\left(\sqrt{n}\sup_x |(F_G - F_n)(x)| > (M_n - K)(\log n)^{1/2}\right) \\ &\le 2\exp\left(-2(M_n - K)^2 \log n\right). \end{aligned}$$

Combining the above assertion with (7), we see that

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)1_{D_n^c}] \lesssim 2\exp\left(-[2(M_n - K)^2 - (2C+1)L_n^2]\log n\right), \tag{8}$$

where the right-hand side of the above inequality tends to zero provided that $(M_n - K) > \sqrt{(C+1/2)}L_n$ for every sufficiently large n. The in-probability convergence in (4) follows from (5), (6) and (8). This concludes the proof. □

Some remarks and comments on Lemma 1 are in order.

- The first one aims at spelling out the assumptions used in the proof, some of which could otherwise erroneously seem to be confined to the context of finite mixture models, as well as at clarifying their role. Given the prior concentration rate ϵ_n as defined in (3), which depends on the prior distribution Π_n and the "point" P_0, the only further assumption used is the *continuity* of the distribution functions F_0 and F_G, which is satisfied for Lebesgue dominated probability measures P_0 and P_G. This condition is used to control the type I and type II error probabilities of the (sequence of) tests (ϕ_n) by the DKW's inequality. It is, instead, in no way used the assumption that the density p_G is modeled as a mixture, so that, even if the result has its origin in the context of finite mixtures, it applies to general dominated models and a nearly parametric (up to a logarithmic factor) prior concentration

rate is the only driver and determinant of posterior contraction.

- Lemma 1 has its roots in Theorem 2 of Ishwaran et al. [11], p. 1324 (see pp. 1330–1331 for the proof), which deals with *finite* mixtures having an *unknown* number of components d_0, yet *bounded* above by an integer N, namely, $1 \leq d_0 \leq N < +\infty$, while the prior is supported over the space of all mixing distributions with at most N components, the mixture weights being assigned an N-dimensional Dirichlet distribution with a non-informative choice for the shape parameters that are all set equal to α/N for a positive constant α. Nonetheless, as previously remarked, Lemma 1 has a broader scope of validity and applies also to *infinite* kernel mixtures with other prior laws for the mixing distribution than the Dirichlet process, which "locally" attain an almost parametric prior concentration rate. This is the case for Dirichlet location or location-scale mixtures of normal densities and, more in general, for location-scale mixtures of exponential power densities with an even integer shape parameter, when the sampling density is of the same form as the assumed model, with mixing distribution being either compactly supported or having sub-exponential tails, see Ghosal and van der Vaart [8], Scricciolo [16], Theorems 4.1, 4.2 and Corollary 4.1, pp. 285–290. In all these cases, the prior concentration rate is (at worst) $\epsilon_n = n^{-1/2} \log n$, where $L_n = (\log n)^{1/2}$. An extension of the previous results to convolution mixtures of super-smooth kernels, with Pitman-Yor or normalized inverse-Gaussian processes as priors for the mixing distribution, for which Lemma 1 also holds, is considered in Scricciolo [17], see Theorem 1, pp. 486–487. Another class of priors on kernel mixtures to which Lemma 1 applies is that of *sieve* priors. For a given kernel, a sieve prior is defined by combining single priors on classes of kernel mixtures, each one indexed by the number of mixture components, with a prior on such random number. A probability measure with kernel mixture density is then generated in two steps: first the model index, i.e., the number of mixture components, is selected; then a probability measure is generated from the chosen model according to a prior on it. When the true density p_0 is itself a kernel mixture, the prior concentration rate can be assessed by bounding below the probability of Kullback-Leibler type neighborhoods of P_0 by the probability of ℓ^1-balls of appropriate dimension. In fact, approximation properties of mixtures under consideration can be exploited to find a good fitting distribution of the sampling density in a proper subclass. More precisely, any finite kernel mixture can be approximated arbitrarily well (in the distance induced by the L^1-norm) by mixtures having the same number of components, the mixture components and weights taking values in ℓ^1-neighborhoods of the corresponding true elements. The number of mixture components is then constant, this leading to the prior concentration rate $\epsilon_n \propto (n/\log n)^{-1/2}$, where $L_n \equiv L_0$. Examples of sieve priors in which, for every choice of the model index, the mixture weights are jointly distributed according to a Dirichlet distribution, are provided by the Bernstein polynomials, see Theorem 2.2 of Ghosal [6], pp. 1268–1269, by histograms and polygons, see Theorem 1 of Scricciolo [15], pp. 629–630 (see pp. 636–637 for the proof). If, as a special case, a single prior distribution on kernel mixtures with a sample size-dependent number $N \equiv L_n$ of mixture components is considered,

then the prior concentration rate is $\epsilon_n = (n/\log n)^{-1/2} L_n$ for any arbitrary slowly varying sequence $L_n \to +\infty$.

We now state sufficient conditions on the kernel density and the prior distributions for the mixture atoms and weights so that the overall prior on kernel mixtures with (at most) N components verifies condition (3) for $\epsilon_n \propto (n/\log n)^{-1/2}$, when the sampling density is itself a kernel mixture with $1 \leq d_0 \leq N$ components. The aim of this analysis is twofold: first, to provide less stringent requirements on the kernel density than those postulated in condition (b) employed in Theorem 2 of Ishwaran et al. [11], p. 1324, which relies on Lemma 4 of Ishwaran [10], pp. 2170–2171; second, to generalize the aforementioned result to a class of prior distributions on the mixture weights that comprises the Dirichlet distribution as a special case. The density p_G is modeled as

$$p_G(\cdot) = \sum_{j=1}^{N} w_j k(\cdot \mid y_j),$$

with a discrete mixing distribution $G = \sum_{j=1}^{N} w_j \delta_{y_j}$. The vector $\mathbf{w}_N := (w_1, \ldots, w_N)$ of mixing weights has a prior distribution $\tilde{\pi}_N$ on the $(N-1)$-dimensional simplex $\Delta_N := \{\mathbf{w}_N \in \mathbb{R}^N : 0 \leq w_j \leq 1,\ j = 1, \ldots, N,\ \sum_{j=1}^{N} w_j = 1\}$ and the atoms $y_1, \ldots, y_N$ are independently and identically distributed according to a prior distribution π. We shall also use the notation $\mathbf{y}_N$ for $(y_1, \ldots, y_N)$. The model can be thus described:

- the random vectors $\mathbf{y}_N$ and $\mathbf{w}_N$ are independent;
- given $(\mathbf{y}_N, \mathbf{w}_N)$, the random variables $X_1, \ldots, X_n$ are conditionally independent and identically distributed according to p_G.

The overall prior is then $\Pi = \tilde{\pi}_N \times \pi^{\otimes N}$. Let the sampling density p_0 be itself a finite kernel mixture, with $1 \leq d_0 \leq N$ components,

$$p_0(\cdot) \equiv p_{G_0}(\cdot) = \sum_{j=1}^{d_0} w_j^0 k(\cdot \mid y_j^0),$$

where the mixing distribution is $G_0 = \sum_{j=1}^{d_0} w_j^0 \delta_{y_j^0}$ for weights $\mathbf{w}_{d_0}^0 := (w_1^0, \ldots, w_{d_0}^0) \in \Delta_{d_0}$ and support points $\mathbf{y}_{d_0}^0 := (y_1^0, \ldots, y_{d_0}^0) \in \mathbb{R}^{d_0}$. A caveat applies: if d_0 is strictly smaller than N, that is, $1 \leq d_0 < N$, then the vectors $\mathbf{w}_{d_0}^0$ and $\mathbf{y}_{d_0}^0$ are viewed as degenerate elements of Δ_N and $\mathbb{R}^N$, respectively, with coordinates $w_{d_0+1} = \cdots = w_N = 0$ and $y_{d_0+1} = \cdots = y_N = 0$.

We assume that

(i) there exists a constant $c_k > 0$ such that

$$\|k(\cdot \mid y_1) - k(\cdot \mid y_2)\|_1 \leq c_k |y_1 - y_2| \quad \text{for all } y_1,\ y_2 \in \mathscr{Y};$$

(ii) for every $\epsilon > 0$ small enough and a constant $c_0 > 0$,

$$\tilde{\pi}_N(\{\mathbf{w}_N \in \Delta_N : \|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1} \leq \epsilon\}) \gtrsim \epsilon^{c_0 N};$$

(iii) the prior distribution π for the atoms has a continuous and positive Lebesgue density (also denoted by π) on an interval containing the support of G_0.

Some remarks and comments on the previously listed assumptions are in order. Condition (i) requires the kernel density $k(\cdot \mid y)$ to be globally Lipschitz continuous on $\mathscr{Y}$. Condition (ii) is satisfied for a Dirichlet prior distribution $\tilde{\pi}_N = \mathrm{Dir}(\alpha_1, \ldots, \alpha_N)$, with parameters $\alpha_1, \ldots, \alpha_N$ such that, for constants a, $A > 0$, $D \geq 1$ and, for $0 < \epsilon \leq 1/(DN)$,

$$A\epsilon^a \leq \alpha_j \leq D, \quad j = 1, \ldots, N.$$

Using Lemma A.1 of Ghosal [6], pp. 1278–1279, we find that $\tilde{\pi}_N(N(\mathbf{w}_N^0; \epsilon)) \gtrsim \exp(-c_0 N \log(1/\varepsilon))$ for a constant $c_0 > 0$ depending only on a, A, D and $\sum_{j=1}^N \alpha_j$.

Proposition 1 *Under assumptions* (i)–(iii), *condition* (3) *is verified for*

$$\epsilon_n \propto (n/\log n)^{-1/2}.$$

Proof For every density p_G, with mixing distribution $G = \sum_{j=1}^N w_j \delta_{y_j}$ having support points $\mathbf{y}_N \in \mathbb{R}^N$ and mixture weights $\mathbf{w}_N \in \Delta_N$, by assumption (i) we have

$$\begin{aligned}\|p_G - p_0\|_1 &\lesssim \sum_{j=1}^N w_j^0 \|k(\cdot \mid y_j) - k(\cdot \mid y_j^0)\|_1 + \sum_{j=1}^N |w_j - w_j^0| \|k(\cdot \mid y_j)\|_1 \\ &\lesssim \|\mathbf{y}_N - \mathbf{y}_N^0\|_{\ell^1} + \|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1}.\end{aligned}$$

Let $0 < \epsilon \leq [(1/2) \wedge (1 - e^{-1})/\sqrt{2}]$ be fixed. For $\mathbf{y}_N \in \mathbb{R}^N$ and $\mathbf{w}_N \in \Delta_N$ such that $\|\mathbf{y}_N - \mathbf{y}_N^0\|_{\ell^1} \leq \epsilon$ and $\|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1} \leq \epsilon$, by LeCam [12] inequalities, p. 40, relating the L^1-norm and the Hellinger metric, the squared Hellinger distance between p_0 and p_G can be bounded above by a multiple of ϵ:

$$\rho_{-1/2}(P_0 \| P_G) = \int (p_G^{1/2} - p_0^{1/2})^2 \, \mathrm{d}\lambda \leq \|p_G - p_0\|_1 \lesssim \epsilon.$$

Then, by Lemma A.10 in Scricciolo [16], p. 305, for a suitable constant $c_1 > 0$,

$$\left\{ p_G : G = \sum_{j=1}^N w_j \delta_{y_j}, \ \|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1} \leq \epsilon, \ \|\mathbf{y}_N - \mathbf{y}_N^0\|_{\ell^1} \leq \varepsilon \right\} \subseteq B_{\mathrm{KL}}\left(P_0; \, c_1 \epsilon \left(\log \tfrac{1}{\epsilon} \right)^2 \right).$$

Next, define the set $N(\mathbf{w}_N^0;\ \epsilon) := \{\mathbf{w}_N \in \Delta_N :\ \|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1} \le \epsilon\}$. For $\epsilon > 0$ small enough, by assumption (ii),

$$\tilde{\pi}_N(N(\mathbf{w}_N^0;\ \epsilon)) \gtrsim \exp\left(-c_0 N \log(1/\varepsilon)\right)$$

with an appropriate constant $c_0 > 0$. Denoted by $B(\mathbf{y}_N^0;\ \epsilon)$ the $\mathbf{y}_N^0$-centered ℓ^1-ball of radius $\epsilon > 0$,

$$B(\mathbf{y}_N^0;\ \epsilon) := \{\mathbf{y}_N \in \mathbb{R}^N :\ \|\mathbf{y}_N - \mathbf{y}_N^0\|_{\ell^1} \le \epsilon\},$$

by condition (iii) the prior probability of $B(\mathbf{y}_N^0;\ \epsilon)$ under the N-fold product measure $\pi^{\otimes N}$ can be bounded below as follows:

$$\begin{aligned}\pi^{\otimes N}(B(\mathbf{y}_N^0;\ \epsilon)) &\ge \prod_{j=1}^N \pi([y_j^0 - (\epsilon/N),\ y_j^0 + (\epsilon/N)]) \\ &= \prod_{j=1}^N \int_{y_j^0-(\epsilon/N)}^{y_j^0+(\epsilon/N)} \pi(y)\,\mathrm{d}y \gtrsim \exp\left(-d_1 N \log(1/\epsilon)\right)\end{aligned}$$

for a positive constant d_1. Therefore, for appropriate constants $c_1,\ d_2 > 0$,

$$\Pi(B_{\mathrm{KL}}(P_0;\ c_1\epsilon|\log\epsilon|^2)) \gtrsim \tilde{\pi}_N(N(\mathbf{w}_N^0;\ \epsilon))\,\pi^{\otimes N}(B(\mathbf{y}_N^0;\ \epsilon)) \gtrsim \exp\left(-d_2 N \log(1/\epsilon)\right).$$

Set $\xi := (c_1\epsilon)^{1/2}\log(1/\epsilon)$, since $\log(1/\epsilon) \lesssim \log(1/\xi)$, we have $\Pi(B_{\mathrm{KL}}(P_0;\ \xi^2)) \gtrsim \exp\left(-c_2 \log(1/\xi)\right)$ for a real constant $c_2 > 0$ (possibly depending on p_0). Replacing ξ with ϵ_n, we get $\Pi(B_{\mathrm{KL}}(P_0;\ \epsilon_n^2)) \gtrsim \exp\left(-c_2 n\epsilon_n^2\right)$ for sufficiently large n, and the proof is complete. □

Inspection of the proof of Lemma 1 reveals that, under the small ball prior probability estimate in (3), we have

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})] = O((n\epsilon_n^2)^{-1}).$$

The assertion of Lemma 1 can be enhanced to have

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})] = O(\exp\left(-B_1 n\epsilon_n^2\right))$$

by employing a small ball prior probability estimate involving stronger divergences. The convergence in (4) then becomes almost-sure. Besides, due to the fact that the posterior probability vanishes exponentially fast, namely, along almost all sample sequences, for a finite constant $B > 0$, we have

$$\Pi_n(A_n^c \mid X^{(n)}) \lesssim \exp\left(-Bn\epsilon_n^2\right) \text{ for all but finitely many } n,$$

the stochastic order of the maximum absolute difference between F_0 and the posterior expected distribution function can be assessed, see Corollary 1 below.

Lemma 2 *Under the conditions of Lemma 1, if the small ball prior probability estimate in (3) is replaced by*

$$\Pi_n(B_{\rho_\alpha}(P_0;\, \epsilon_n^2)) \gtrsim \exp(-Cn\epsilon_n^2), \quad \text{for } \alpha \in (0,\, 1], \tag{9}$$

then, for $M_n \gtrsim \sqrt{(C+1/2)}L_n$,

$$\Pi_n\left(\sqrt{n}\sup_x |(F_G - F_0)(x)| > M_n(\log n)^{1/2} \mid X^{(n)}\right) \to 0 \quad P_0^\infty\text{-almost surely.}$$

Proof The proof is an adaptation of that of Lemma 1. We therefore highlight only the main changes. Taking a sequence $K_n = \theta M_n$ for any $\theta \in (0,\, 1)$, we have

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})\phi_n] \le P_0^n\phi_n \le 2\exp(-2\theta^2 M_n^2 \log n)$$

and

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)1_{D_n^c}] \lesssim 2\exp(-[2(1-\theta)^2 M_n^2 - (2C+1)L_n^2]\log n),$$

with

$$M_n > (1-\theta)^{-1}\sqrt{(C+1/2)}L_n \tag{10}$$

for every sufficiently large n. A straightforward extension of Lemma 2 in Shen and Wasserman [19], p. 691 (and pp. 709–710 for the proof), yields that, for every $\xi \in (0,\, 1)$,

$$P_0^n\big(D_n \le \xi\,\Pi_n(B_{\rho_\alpha}(P_0;\, \epsilon_n^2))\exp(-(C+1)n\epsilon_n^2)\big) \le (1-\xi)^{-1}\exp(-\alpha Cn\epsilon_n^2). \tag{11}$$

Considering $M_n = IL_n$ for a finite constant $I > (1-\theta)^{-1}\sqrt{(C+1/2)}$ so that condition (10) is satisfied, by combining partial bounds we obtain that

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})] \lesssim \exp(-B_1 n\epsilon_n^2)$$

for an appropriate finite constant $B_1 > 0$. For a constant $B > 0$,

$$P_0^n\big(\Pi_n(A_n^c \mid X^{(n)}) \ge \exp(-Bn\epsilon_n^2)\big) \lesssim \exp\big(-(B_1 - B)n\epsilon_n^2\big).$$

Choose $0 < B < B_1$. Since $\sum_{n=1}^\infty \exp(-(B_1 - B)n\epsilon_n^2) < +\infty$, almost sure convergence follows from the first Borel-Cantelli lemma. □

Remark 1 The assertion of Lemma 2 still holds if the small ball prior probability estimate in (9) is replaced by the requirement

$$\Pi_n(B_{\rho_{-1/2}\|\cdot\|_\infty}(P_0;\, \epsilon_n^2)) \gtrsim \exp(-Cn\epsilon_n^2), \tag{12}$$

which involves a Hellinger type neighborhood of P_0. Then, a bound similar to that in (11) is given in Lemma 8.4 of Ghosal et al. [7], pp. 526–527.

As previously mentioned, Lemma 2 allows to derive the stochastic order of the maximum absolute difference between F_0 and its Bayes' estimator

$$F_n^{\mathrm{B}}(\cdot) := \int_{\mathscr{G}} F_G(\cdot)\Pi(\mathrm{d}G \mid X^{(n)}),$$

namely, the posterior expected distribution function.

Corollary 1 *Under the conditions of Lemma 2, we have*

$$\sqrt{n}\sup_x |(F_n^{\mathrm{B}} - F_0)(x)| = O_{\mathbf{P}}(M_n(\log n)^{1/2}).$$

Proof By standard arguments,

$$\begin{aligned}
\sqrt{n}\sup_x |(F_n^{\mathrm{B}} - F_0)(x)| &= \sqrt{n}\sup_x \left| \int_{\mathscr{G}} F_G(x)\, \Pi_n(\mathrm{d}G \mid X^{(n)}) - F_0(x) \right| \\
&\le \int_{\mathscr{G}} \sqrt{n}\sup_x |(F_G - F_0)(x)|\, \Pi_n(\mathrm{d}G \mid X^{(n)}) \\
&= \left(\int_{A_n} + \int_{A_n^c} \right) \sqrt{n}\sup_x |(F_G - F_0)(x)|\, \Pi_n(\mathrm{d}G \mid X^{(n)}) \\
&\le M_n(\log n)^{1/2} + 2\sqrt{n}\Pi_n(A_n^c \mid X^{(n)}) \\
&\lesssim M_n(\log n)^{1/2} \quad \text{for sufficiently large } n
\end{aligned}$$

because condition (9) yields that, with probability one, for a finite constant $B > 0$, the posterior probability $\sqrt{n}\Pi_n(A_n^c \mid X^{(n)}) \lesssim \sqrt{n}\exp(-Bn\epsilon_n^2)$ for all but finitely many n. The assertion follows. □

Posterior Concentration of the Mixing Distribution in the Kantorovich Metric

In this section, we deal with the case where the prior distribution Π is supported over the collection of finite kernel mixtures with at most N components. Sufficient conditions are stated in Theorem 1 below so that the posterior rate of convergence, relative to the Kantorovich or L^1-Wasserstein metric, for the mixing distribution of overfitted mixtures is, up to a slowly varying sequence, (at worst) equal to $(n/\log n)^{-1/4}$, the optimal pointwise rate being $n^{-1/4}$, *cf.* Chen [1], Sect. 2, pp. 222–224.

In order to state the result, we need to introduce some more notation. For every $y \in \mathscr{Y}$, we denote by $K(x \mid y)$ the cumulative distribution function at x of the kernel density $k(\cdot \mid y)$,

$$K(x \mid y) := \int_{-\infty}^{x} k(u \mid y)\,\mathrm{d}u.$$

For clarity of exposition, we recall that F_0 is the distribution function of the mixture density $p_0 \equiv p_{G_0}$ corresponding to the mixing distribution G_0 having an *unknown* number of components d_0 *bounded* above by a fixed integer N.

Theorem 1 *Under the conditions of Lemma 1, if, in addition,*

(a) $\mathscr{Y}$ *is compact,*

(b) for all $x \in \mathbb{R}$, $K(x \mid y)$ *is* 2*-differentiable with respect to* y,

(c) $\{K(\cdot \mid y) : y \in \mathscr{Y}\}$ *is strongly identifiable in the sense of Definition 2 in Chen [1], p. 225, equivalently,* 2*-strongly identifiable in the sense of Definition 2.2 in Heinrich and Kahn [9], p. 2848,*

(d) there exists a uniform modulus of continuity $\omega(\cdot)$ *such that*

$$\sup_x |K^{(2)}(x \mid y) - K^{(2)}(x \mid y')| \leq \omega(|y - y'|) \quad \textit{with } \lim_{h \to 0} \omega(h) = 0,$$

then, for $M_n \gtrsim \sqrt{(C + 1/2)}L_n$,

$$\Pi\left(n^{1/4} W_1(G, G_0) > \sqrt{M_n}(\log n)^{1/4} \mid X^{(n)}\right) = o_{\mathbf{P}}(1).$$

Proof Since Lemma 1 holds, we have

$$\Pi\left(\sqrt{n} \sup_x |(F_G - F_0)(x)| > M_n(\log n)^{1/2} \mid X^{(n)}\right) = o_{\mathbf{P}}(1). \tag{13}$$

Consistently with the notation introduced in Lemma 1, we set

$$A_n := \left\{G : \sqrt{n} \sup_x |(F_G - F_0)(x)| \leq M_n(\log n)^{1/2}\right\}.$$

Under assumptions (a)–(d), assertion (21) of Theorem 6.3 of Heinrich and Kahn [9], p. 2857, holds true, this implying that, for every $G \in A_n$, the Kolmogorov distance between the distribution functions F_G and F_0 is bounded below (up to a constant) by the squared L^1-distance between the mixing distributions G and G_0, respectively: there exists a constant $C_0 > 0$ (possibly depending on G_0) such that, for every $G \in A_n$,

$$C_0 \|G - G_0\|_1^2 < \sup_x |(F_G - F_0)(x)| \leq M_n n^{-1/2}(\log n)^{1/2}. \tag{14}$$

Taking into account the following representation of the L^1-Wasserstein distance

$$W_1(G, G_0) = \|G - G_0\|_1,$$

see, e.g., Shorack and Wellner [20], pp. 64–66, which was obtained by Dall'Aglio [2], the assertion follows by combining (13) with (14). This concludes the proof. □

Some comments on the applicability and consequences of Theorem 1 are in order.

- Theorem 1, like Lemma 1, has its roots in Theorem 2 of Ishwaran et al. [11], p. 1324, which is tailored for finite Dirichlet mixtures. However, thanks to Proposition 1, which implies the conclusion of Lemma 1, meanwhile ensuring applicability to a larger family of prior distributions, under conditions (a)–(d), the assertion that, for sufficiently large constant $M > 0$, the convergence

$$\Pi\left(n^{1/4} W_1(G,\, G_0) > M(\log n)^{1/4} \mid X^{(n)}\right) \to 0 \quad \text{in } P_0^n\text{-probability}$$

  takes place, still holds. The present result differs from that of Theorem 5 in Nguyen [14], pp. 383–384, under various respects: the latter gives an assessment of posterior contraction in the L^2-Wasserstein, as opposed to the L^1-Wasserstein metric, for finite mixtures of multivariate distributions, under more stringent conditions and following a completely different line of reasoning.

- As previously observed on the occasion of the transition from Lemma 1 to Lemma 2, if the small ball prior probability estimate in (3) is replaced with either that in (9) or in (12), then the almost-sure version of Theorem 1

$$\Pi\left(n^{1/4} W_1(G,\, G_0) > \sqrt{M_n}(\log n)^{1/4} \mid X^{(n)}\right) \to 0 \quad P_0^\infty\text{-almost surely}$$

  holds and the rate of convergence for the Bayes' estimator of the mixing distribution can be assessed as follows.

Corollary 2 *Under the conditions of Theorem 1, with the small ball prior probability estimate in (9), we have*

$$W_1(G_n^{\mathrm{B}},\, G_0) = O_{\mathbf{P}}(\sqrt{M_n}(n/\log n)^{-1/4}),$$

where $G_n^{\mathrm{B}}(\cdot) := \int_{\mathscr{G}} G(\cdot)\Pi(\mathrm{d}G \mid X^{(n)})$ *is the Bayes' estimator of the mixing distribution.*

Acknowledgements The author gratefully acknowledges financial support from MIUR, grant n° 2015SNS29B "Modern Bayesian nonparametric methods".

References

1. Chen, J.: Optimal rate of convergence for finite mixture models. Ann. Stat. **23**(1), 221–233 (1995)
2. Dall'Aglio, G.: Sugli estremi dei momenti delle funzioni di ripartizione doppia. (Italian) Ann. Scuola Norm. Sup. Pisa **3**(10), 35–74 (1956)
3. Dvoretzky, A., Kiefer, J., Wolfowitz, J.: Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Stat. **27**(3), 642–669 (1956)
4. Efron, B.: Empirical Bayes deconvolution estimates. Biometrika **103**(1), 1–20 (2016)
5. Gao, F., van der Vaart, A.: Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures. Electron. J. Stat. **10**(1), 608–627 (2016)
6. Ghosal, S.: Convergence rates for density estimation with Bernstein polynomials. Ann. Stat. **29**(5), 1264–1280 (2001)
7. Ghosal, S., Ghosh, J.K., van der Vaart, A.W.: Convergence rates of posterior distributions. Ann. Stat. **28**(2), 500–531 (2000)
8. Ghosal, S., van der Vaart, A.W.: Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann. Stat. **29**(5), 1233–1263 (2001)
9. Heinrich, P., Kahn, J.: Strong identifiability and optimal minimax rates for finite mixture estimation. Ann. Stat. **46**(6A), 2844–2870 (2018)
10. Ishwaran, H.: Exponential posterior consistency via generalized Pólya urn schemes in finite semiparametric mixtures. Ann. Stat. **26**(6), 2157–2178 (1998)
11. Ishwaran, H., James, L.F., Sun, J.: Bayesian model selection in finite mixtures by marginal density decompositions. J. Am. Stat. Assoc. **96**(456), 1316–1332 (2001)
12. LeCam, L.: Convergence of estimates under dimensionality restrictions. Ann. Stat. **1**(1), 38–53 (1973)
13. Massart, P.: The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. Ann. Probab. **18**(3), 1269–1283 (1990)
14. Nguyen, X.: Convergence of latent mixing measures in finite and infinite mixture models. Ann. Stat. **41**(1), 370–400 (2013)
15. Scricciolo, C.: On rates of convergence for Bayesian density estimation. Scand. J. Stat. **34**(3), 626–642 (2007)
16. Scricciolo, C.: Posterior rates of convergence for Dirichlet mixtures of exponential power densities. Electron. J. Stat. **5**, 270–308 (2011)
17. Scricciolo, C.: Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or Normalized Inverse-Gaussian process kernel mixtures. Bayesian Anal. **9**(2), 475–520 (2014)
18. Scricciolo, C.: Bayes and maximum likelihood for L^1-Wasserstein deconvolution of Laplace mixtures. Stat. Methods Appl. **27**(2), 333–362 (2018)
19. Shen, X., Wasserman, L.: Rates of convergence of posterior distributions. Ann. Stat. **29**(3), 687–714 (2001)
20. Shorack, G.R., Wellner, J.A.: Empirical Processes with Applications to Statistics. Wiley, New York (1986)
21. Wong, W.H., Shen, X.: Probability inequalities for likelihood ratios and convergence rates of sieve MLES. Ann. Stat. **23**(2), 339–362 (1995)

Discovering and Locating High-Energy Extra-galactic Sources by Bayesian Mixture Modelling

Andrea Sottosanti, Denise Costantin, Denis Bastieri and Alessandra Rosalba Brazzale

Abstract Discovering and locating gamma-ray sources in the whole sky map is a declared target of the *Fermi* Gamma-ray Space Telescope collaboration. In this paper, we carry out an unsupervised analysis of the collection of high-energy photons accumulated by the Large Area Telescope, the principal instrument on board the *Fermi* spacecraft, over a period of around 7.5 years using a Bayesian mixture model. A fixed, though unknown, number of parametric components identify the extra-galactic emitting sources we are searching for, while a further component represents parametrically the diffuse gamma-ray background due to both, extra-galactic and galactic high-energy photon emission. We determine the number of sources, their coordinates on the map and their intensities. The model parameters are estimated using a reversible jump MCMC algorithm which implements four different types of moves. These allow us to explore the dimension of the parameter space. The possible transitions remove from or add a source to the model, while leaving the background

A. Sottosanti (✉) · A. R. Brazzale
Department of Statistical Sciences, University of Padova,
via Cesare Battisti 241, Padova, Italy
e-mail: sottosanti@stat.unipd.it

A. R. Brazzale
e-mail: alessandra.brazzale@unipd.it

D. Costantin
Center for Astrophysics, Guangzhou University, Guangzhou 510006, China

Department of Education, Astronomical Science and Technology Research Laboratory,
Guangdong, Guangdong Province, People's Republic of China

Department of Statistical Sciences, University of Padova,
via Cesare Battisti 241, Padova, Italy
e-mail: denise.costantin@gmail.com

D. Bastieri
Department of Physics and Astronomy "Galileo Galilei", University of Padova, via Belzoni 7,
Padova, Italy
e-mail: denis.bastieri@unipd.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_11

component unchanged. We furthermore present an heuristic procedure, based on the posterior distribution of the mixture weights, to qualify the nature of each detected source.

Keywords Astrostatistics · Bayesian clustering · Reversible jump MCMC · Signal extraction

1 Motivation and Background

Resolving the γ-ray sky by detecting as yet unidentified sources and accurately measuring the diffuse background emission is a declared key scientific objective of the *Fermi* Gamma-ray Space Telescope collaboration, whose broader aim is to identify and study the nature of high-energy phenomena in the Universe.[1] The target of this contribution is the collection of photon count maps for varying energy bins provided by the Large Area Telescope (LAT), the principal scientific instrument on board the *Fermi* spacecraft, during its almost ten years of operation. In particular, we aim at formulating and fitting a model which allows us to: (i) determine the number of extra-galactic high-energy sources, (ii) measure their intensities, and (iii) pool the individual photon counts into the corresponding clusters.

The discovery of celestial objects is an intrinsically interdisciplinary field which combines both, statistical and astronomical methodology. A main challenge of trying and detecting high-energy phenomena from astronomical data is to separate the signal of the putative emitting source from the diffuse γ-ray background which spreads over the entire area observed by the telescope. Different phenomena contribute to this residual radiation [3]. Broadly speaking, its origins can be brought under two headings: galactic interstellar emission (GIE), that is, the interaction of galactic cosmic rays with gas and radiation fields, and a residual all-sky emission. The latter is commonly called the isotropic diffuse gamma-ray background (IGRB), and includes the γ-ray emission from faint unresolved sources and any residual galactic emission which is approximately isotropic.

Traditionally, the analysis is based on so-called *single-source* models, as described in Sect. 7.4 of [7]. Generally speaking, the application of these models requires the whole sky map to be split into small regions. The presence of a possible new source is assessed on a pixel-by-pixel basis using significance tests. An illustrative example is given in [11], who employ Poisson regression to model the number of photons at each pixel. Further treatments from both, the frequentist and the Bayesian viewpoints, can be found in [6, 10, 13, 14]. *Variable-source-number* models address the problem from a more global perspective, as they simultaneously estimate the number of sources in the whole map without the need to separate the latter into smaller cells and to work on single pixels [7, Sect. 7.3]. A recent proposal, which analyses X-ray count maps according to this approach, is made in [8].

[1] https://fermi.gsfc.nasa.gov/.

Here we propose a Bayesian mixture model with a finite, but unknown, number of components for the known and as yet unidentified extra-galactic high-energy sources plus an additional parametric component to represent the diffuse γ-ray background. The directions of the high-energy photons collected by the *Fermi* LAT over a period of approximately 7.5 years is then used to estimate simultaneously the number of sources in the map, their coordinates and their intensities. As in [8], our algorithm iteratively identifies the sources and pulls the individual photons into the corresponding clusters. It furthermore automatically selects the number of components of the mixture. However, [8] consider only the isotropic diffuse X-ray background, which they model assuming a uniform distribution over the entire map. This assumption is too restrictive if the targets are γ-ray sources, as we cannot neglect the huge contamination due to galactic interstellar emission, but have to suitably model it.

The remainder of the paper is organised as follows. Section 2 presents the *Fermi* LAT data which motivated this contribution. Our proposal of a Bayesian finite mixture model is presented in Sect. 3 and is fit to the *Fermi* LAT data in Sect. 4. In this latter section we furthermore discuss the capability of our model to skim off the signal of the sources from the background radiation. The paper closes with the concluding remarks of Sect. 5.

2 The *Fermi* LAT Data

The data collected by the *Fermi* Large Area Telescope (LAT) contribute uniquely to the study of the most extreme phenomena in our Universe such as active galactic

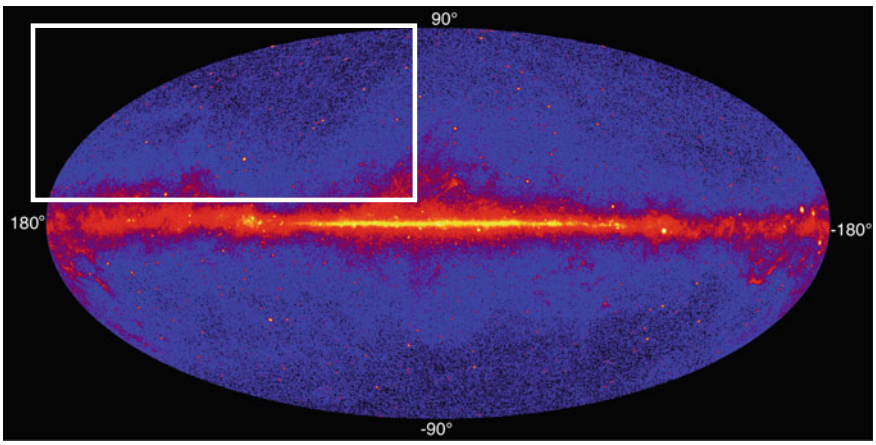


Fig. 1 Whole sky map at γ-ray wavelengths and energies larger than 1 GeV based on data accumulated by the LAT over a period of five years of operation (Image Credit: NASA/DOE/Fermi LAT Collaboration). The region framed in white represents the area analyzed in this paper

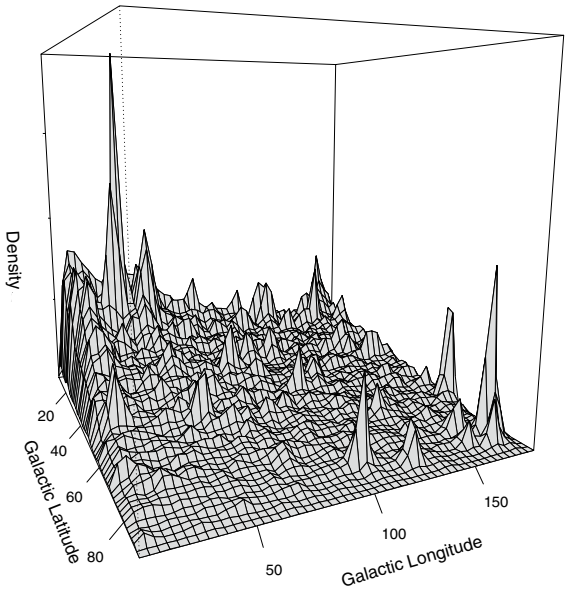


Fig. 2 Nonparametric kernel estimate of the photon density distribution based on the γ-ray count maps accumulated by the *Fermi* LAT over a 7.5 years period. The map is expressed in galactic coordinates and refers to the sky region $[180°, 10°] \times [10°, 90°]$, that is, to the area framed in white in Fig. 1. The spikes represent potential high-energy extra-galactic emitting sources. Both components of the γ-ray background, the GIE and the IGRB, are visible

nuclei, supernova remnants and pulsar wind nebula. Figure 1 represents the Mollweide projection in galactic coordinates of the entire γ-ray sky at energies larger than 1 GeV and is based on the data collected by NASA's *Fermi* LAT over a five years period.[2] The brighter the grey tone, the larger is the intensity of the γ-ray source. The brilliant horizontal stripe which crosses the middle part of the figure to a huge extent conveys the high-energy photon emission of our Milky Way, at whose center we assume a supermassive black hole. The isotropic diffuse γ-ray background is much less evident, while we can clearly identify extra-galactic point and small-area γ-ray emitting sources.

The dataset used in this paper is the collection of photons, generated by different astrophysical events and collected by the LAT over a period of around 7.5 years of observation, whose energy exceeds 10 GeV. The aim of our analysis is to discriminate the signal of extra-galactic γ-ray emitting sources from the various background phenomena, and to reconstruct their direction in the sky map. In particular, for the reasons we will shortly give below, we restrict our attention to a subregion of the sky whose galactic longitude and latitude lie in the intervals $[180°, 10°]$ and $[10°, 90°]$, respectively.[3] This region is framed in white in Fig. 1 and covers broadly the fourth quadrant of the map. In all, 51,000 observations fall in this area. Figure 2 plots the smoothed nonparametric estimate of the photon density distribution. The various spikes identify known and as yet unrevealed high-energy emitting sources.

[2] http://fermi.gsfc.nasa.gov/ssc/.

[3] Here we follow the convention adopted in astronomical whole sky maps to define the longitude on the left at 180° and at −180° on the right.

We decided to test and fine tune our algorithm in a region of the sky map where the diffuse γ-ray background is less prevalent, and possibly dominated by the isotropic diffuse *gamma*-ray background (IGRB) component. Hence, we restricted our analysis to latitudes above 10° to limit the influence of the galactic interstellar emission (GIE). To further reduce and, at least partially remove, the background component radiating from the Galaxy center and from the so-called *Fermi* Bubbles [2], that is, from the two extended regions of excess γ-ray emission located near the galactic center, we only consider longitudes that vary from 180° to 10°. As is evident from Fig. 2, the IGRB is still present though less pronounced as compared to Fig. 1. In Sect. 3.1 we will discuss how to suitably model the diffuse background component. The third catalogue of hard Fermi LAT sources (3FHL, for short) reports 288 high-energy γ-ray emitting sources for the outlined region [4].

3 Bayesian Source Detection

We adopted a flexible Bayesian modelling approach which allowed us to detect catalogued and acknowledged γ-ray sources plus possible new candidates in the sky region of Fig. 2. As in [8] we assembled a finite mixture model whose components were automatically identified using the available data and Bayesian computation. That is, in one go we determined both, the number of sources and their directions in space. The main difference to [8] is the presence of the rather intense background radiation which spreads over the entire map and represents a relevant component of our model. Section 3.1 describes the statistical model for the *Fermi* LAT data; Sect. 3.2 outlines the fitting procedure.

3.1 Statistical Model

Let $x_i \in [180, 10]$ and $y_i \in [10, 90]$ represent the galactic longitude and latitude, respectively, of the n photons detected in the area of the extra-galactic space considered by our analysis. We start off by reconstructing the directions of the γ-ray sources by modelling how the photons they emit scatter around their source.

Assume that photon i was generated by source j whose galactic coordinates are $\mu_j = (\mu_{jx}, \mu_{jy}), j = 1, \ldots, K$. Here K represents the number of sources present in the map. The direction of photon i can then be modeled as

$$(X_i, Y_i) \mid \mu_j \sim PSF(\mu_j), \quad i = 1, \ldots, n, \tag{1}$$

where $PSF(\cdot)$ represents King's established *Point Spread Function* [9]. This function suitably describes how photons cluster around their emitting source. The corresponding density is

$$f(x_i, y_i \mid \mu) = \frac{C}{[1 + \{d(x_i, y_i \mid \mu)/d_0\}^2]^{\eta_0}}, \tag{2}$$

where

$$d(x_i, y_i \mid \mu) = \sqrt{(x_i - \mu_{jx})^2 + (y_i - \mu_{jy})^2/(1 - \varepsilon_0)^2}.$$

Here $d_0 = 0.6$ is the core radius measured in arcsec, $\eta_0 = 1.5$ is the power-law slope and $\varepsilon_0 = 0.00574$ represents the ellipticity; the normalizing constant C is usually determined numerically. The resulting density essentially characterizes a bivariate Student t distribution. The values of the parameters d_0, η_0 and ε_0 are chosen as in [8]. Actually, the *Fermi* LAT collaboration uses an extended version of King's density [1]. In particular, they assume that photons generated from the same source are not identically distributed, but each is characterized by its own dispersion which, in turn, depends on the energy level of the photon. However, for the energy range considered in this paper (>10 GeV), this variation is negligible and model (1) represents a valid approximation.

A different model needs be specified in case the observed photon was not emitted from a specific source but is part of the background radiation. The authors of [8] assume a uniform distribution over the entire map to model the uniquely present isotropic component. We already discussed in Sect. 2 that this assumption is too restrictive for γ-ray counts. Model (1) is hence extended by considering a further bi-dimensional component

$$(X_i, Y_i) \mid \sigma_b \ \sim \ Unif\,(180, 10) \times tExp(\sigma_b). \tag{3}$$

The longitude of a photon stemming from the background is here modelled according to a uniform distribution, while its latitude follows a translated exponential distribution with scale parameter σ_b, that is, an exponential distribution whose support was translated to the interval $[10, +\infty)$. This model well represents the marginal distributions of longitude and latitude for the photons detected by the *Fermi* LAT shown in Fig. 3. Suitable values will be chosen for σ_b so as to guarantee that the

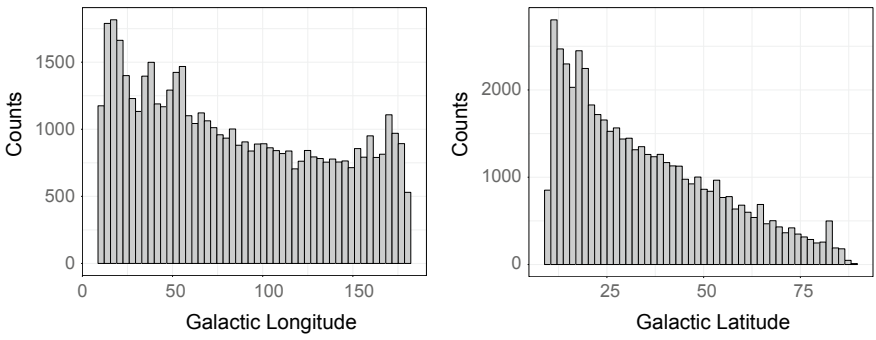


Fig. 3 Distribution of longitudes (left) and of latitudes (right) of the high-energy (>10 GeV) photons detected by the *Fermi* LAT during a 7.5 years period of observation

fitting procedure outlined in the following section generates admissible values for Y_i.

In practice, we have no information whether the photon was emitted from a source or belongs to the background, nor do we know the number of emitting sources and their directions in space. This situation is well represented by a finite mixture model which assumes a fixed, though unknown, number of components to represent the different sources plus an additional component to model the background radiation. This translates into the following marginal model

$$f(x_i, y_i \mid \mu, \sigma_b, \omega) = \omega_0 \, g_b(x_i, y_i \mid \sigma_b) + \sum_{j=1}^{K} \omega_j \, f(x_i, y_i \mid \mu_j), \tag{4}$$

where $g_b(\cdot \mid \cdot)$ represents the distribution of photons from the background as given in (3), while $f(\cdot \mid \cdot)$ models the signal of a specific source according to (1). The vector $\omega = (\omega_0, \omega_1, \ldots, \omega_K)$ contains the mixing proportions ω_j which can be viewed as the intensity ω_0 of the background and of each source, that is, ω_i, $i = 1, \ldots, K$. Our model is hence characterised by a set $\theta_K = \{\mu, \sigma_b, \omega\}$ of $3K + 2$ parameters. Recall, furthermore, that the number K of undetected sources is itself supposed to be unknown and needs be estimated. So, inference will be made on (θ_K, K).

To write down the likelihood function of the statistical model defined at (4), we augment our data as originally proposed in [12] and also advocated in [8]. That is, for each observation $i = 1, \ldots, n$, we introduce the latent group variable Z_i which assumes values in the discrete set $\{0, 1, \ldots, K\}$ with probabilities given by the components of ω. Though actually never observed, this variable conveys useful information as it indicates the source number for photon i. The full data likelihood is then

$$\begin{aligned} L(\theta_K, K \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) &= p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}; \theta_K, K) p(\mathbf{z} \mid \theta_K, K) \\ &= \left[\prod_{i:z_i=0} g_b(x_i, y_i \mid \sigma_b) \prod_{j=1}^{K} \left\{ \prod_{i:z_i=j} f(x_i, y_i \mid \mu_j) \right\} \right] \prod_{j=0}^{K} \omega_j^{n_j}, \end{aligned} \tag{5}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_m)$ and $\mathbf{z} = (z_1, \ldots, z_n)$ are the vectors of observed and latent data, and $n_j = \sum_{i=1}^{n} I(z_i = j)$. As required by Bayes we complete our model definition by eliciting the a priori distributions for the unknown model parameters θ_K and K. Since there is no prior belief on the direction of the sources, a bivariate uniform distribution is used,

$$\mu_{jx} \sim Unif(180, 10) \quad \text{and} \quad \mu_{jy} \sim Unif(10, 90),$$

while the conjugate gamma distribution

$$\pi(\sigma_b \mid \nu, \beta) = \frac{\beta^{\nu}}{\Gamma(\nu)} \sigma_b^{\nu-1} e^{-\beta\sigma_b},$$

Algorithm 1 Reversible jump MCMC – split move

procedure SPLIT j INTO j_1, j_2 WITH PROBABILITY b_k (from k to $k+1$ sources)
  $b_k \leftarrow 0.25$, $d_{k+1} \leftarrow 0.25$
  if $k = \kappa_{min}$ **then** $b_k \leftarrow 0.5$
  $u_1, u_2, u_3 \sim Beta(2,2)$, $v \sim Unif(0,1)$
  $\omega_{j_1} \leftarrow u_1\omega_j$ and $\omega_{j_2} \leftarrow (1-u_1)\omega_j$
  $\mu_{j_1x} \leftarrow \mu_{jx} - u_2\sqrt{\omega_{j_2}/\omega_{j_1}}$ and $\mu_{j_1y} \leftarrow \mu_{jy} - u_3\sqrt{\omega_{j_2}/\omega_{j_1}}$
  $\mu_{j_2x} \leftarrow \mu_{jx} + u_2\sqrt{\omega_{j_1}/\omega_{j_2}}$ and $\mu_{j_2y} \leftarrow \mu_{jy} + u_3\sqrt{\omega_{j_1}/\omega_{j_2}}$
  generate a new vector of labels $\mathbf{z}^*$ using $k+1$ sources
  $p_{k+1} \leftarrow \pi(\theta_{k+1}, k+1 \mid \mathbf{x}, \mathbf{y}, \mathbf{z}^*)$ and $p_k \leftarrow \pi(\theta_k, k \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$
  $g \leftarrow b_{2,2}(u_1)b_{2,2}(u_2)b_{2,2}(u_3)$, where $b_{2,2}(.)$ is the $Beta(2,2)$ density function
  $J \leftarrow \omega_j/[u_1(1-u_1)]$
  $q_k \leftarrow b_k/k$ and $q_{k+1} \leftarrow d_{k+1}/(k+1)$
  if $\arg\min_j \mid\mid \mu_{j_1}, \mu_j \mid\mid = j_2$ **and** $\arg\min_j \mid\mid \mu_{j_2}, \mu_j \mid\mid = j_1$ **then**
    $q_{k+1} \leftarrow 2q_{k+1}$
  $A \leftarrow (p_{k+1}q_{k+1}J)/(p_kq_kg)$
  if $v \leq \min(1, A)$ **then** accept split

with $\nu = 0.02$ and $\beta = 1$, is chosen for the scale parameter σ_b. We, furthermore, assume that the unknown number of components K distributes as a truncated Poisson

$$K \sim tPoi(\kappa \mid \kappa_{min}, \kappa_{max}), \tag{6}$$

where $\kappa = 288$ equals the number of catalogued sources and $[\kappa_{min}, \kappa_{max}] = [250, 400]$. This way the number of detected sources is bound to lay between 254 and 321, that is, ± 2 standard deviations from $\kappa = 288$ wasn't the Poisson truncated. This particular choice allows us to simultaneously detect an 11% of new sources and to reduce the number of false positives. Indeed, because of the rather high background contamination and the limited capability of our parametric formulation to fully capture its rather irregular shapes, we need a prior which limits the upper bound of the support of K. Lastly, conditionally on K, we let ω follow a Dirichlet distribution of size $K+1$ where the $K+1$ parameters are all set to $\alpha = 1$. This corresponds to assigning a priori equal probability to the K putative sources, or differently stated, to assuming that they have the same intensity.

Applying Bayes' theorem, the posterior joint distribution of the unknown model parameters (θ_K, K), conditionally on the latent group variables $\mathbf{z}$, results in

$$\pi(\theta_K, K \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) \propto L(\theta_K, K \mid \mathbf{x}, \mathbf{y}, \mathbf{z})\pi(\theta_K, K). \tag{7}$$

This is the function we will use to estimate the parameters. Note that to obtain the posterior distribution of θ_K and K given only the observed data $(\mathbf{x}, \mathbf{y})$, we would have to sum up (7) over all possible combinations of the latent variables $\mathbf{z} = (z_1, \ldots, z_n)$.

3.2 Model Fitting

We by-passed numerical integration of the posterior kernel (7), as would have been required to compute the normalising constant, using Monte Carlo simulation. However, a further aspect considerably challenges the derivation of the posterior distribution of the model parameters: the dimension of θ_K is itself unknown as it depends on the number of sources K. We implemented a reversible jump Markov chain Monte Carlo algorithm, as proposed in [5], thanks to which we were able to both, reconstruct the posterior distributions of the unknown components of the model and to determine how many there are.

Here we present our algorithm. It consists of a two-stage procedure which iterates two steps: given K, we first update the latent group variables $\mathbf{Z}$ and generate values from the posterior distribution of θ_K; in the second step we redetermine the number of components K. Having written $\mathbf{z}^{(t-1)}$, $\theta_K^{(t-1)}$ and $K^{(t-1)}$ for the values generated at iteration $(t-1)$, the two steps can be summarised as follows:

1. generate $(\mathbf{z}^t, \theta_K^t)$ from the full conditional $\pi(\mathbf{z}, \theta_K \mid K^{(t-1)}; \mathbf{x}, \mathbf{y})$;
2. redefine the dimension of the parameter space, that is, specify a new order of the mixture by generating K^t from $\pi(K \mid \theta_K^t, \mathbf{z}^t; \mathbf{x}, \mathbf{y})$.

An alternative is to have the algorithm iterate Step 1 a given number of times, say 5–10, before proposing the trans-dimensional jump outlined at Step 2. Let us now have a closer look at the two steps.

Step 1

This step implements a Gibbs sampling scheme to update the model parameters θ_K and the latent variables $\mathbf{Z}$ for a fixed number K of components. Let, as above, the superscripts $(t-1)$ and t identify the values generated at iterations $(t-1)$ and t, respectively, and define as k the number of sources detected at iteration $(t-1)$, that is, $K^{(t-1)} = k$. Step 1 of the algorithm develops as follows.

1. For $i = 1, \ldots, n$, generate z_i^t from a multinomial distribution with probabilities

$$\begin{aligned} p(z_i^t = 0 \mid \theta_K^{(t-1)}, K^{(t-1)}; \mathbf{x}, \mathbf{y}) &\propto \omega_0^{(t-1)} g_b(x_i, y_i \mid \sigma_b^{(t-1)}) \\ p(z_i^t = j \mid \theta_K^{(t-1)}, K^{(t-1)}; \mathbf{x}, \mathbf{y}) &\propto \omega_j^{(t-1)} f(x_i, y_i \mid \mu_j^{(t-1)}), \quad j \neq 0. \end{aligned}$$

2. Generate a new vector of mixing probabilities ω^t from the Dirichlet distribution $Dir(n_0^t + \alpha, \ldots, n_k^t + \alpha)$, where $n_j = \sum_{i=1}^n I(z_i^t = j), j = 1, \ldots, k$.
3. Generate $\mu_j^t, j = 1, \ldots, k$, using a Metropolis-Hastings step applied to the full conditional distribution

$$\pi(\mu \mid \sigma_b^{(t-1)}, K^{(t-1)}; \mathbf{x}, \mathbf{y}, \mathbf{z}^t).$$

Use as proposal distribution the bivariate normal distribution centered at $\mu_j^{(t-1)}$ and with covariance matrix the identity matrix rescaled by 0.5^2 so as to guarantee a satisfactory overlapping with King's PSF defined in (1).

Algorithm 2 Reversible Jump MCMC – birth move

procedure GENERATE j^* WITH PROBABILITY b_K (from k to $k+1$ sources)
  $b_k \leftarrow 0.25,\ d_{k+1} \leftarrow 0.25$
  if $k = \kappa_{min}$ **then** $b_k \leftarrow 0.5$
  $\mu_{j^*x} \sim Unif(10, 180),\ \mu_{j^*y} \sim Unif(10, 90)$ and $\omega_{j^*} \sim Beta(1, k+1)$
  rescale the weights using $\omega_j \leftarrow \omega_j(1-\omega_{j^*})$
  generate a new vector of labels $\mathbf{z}^*$ using $k+1$ sources
  $p_{k+1} \leftarrow \pi(\theta_{k+1}, k+1 \mid \mathbf{x}, \mathbf{y}, \mathbf{z}^*)$ and $p_k \leftarrow \pi(\theta_k, k \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$
  $g \leftarrow \pi(\mu_{j^*x})\pi(\mu_{j^*y})b_{1,k+1}(\omega_{j^*})$
  $J \leftarrow (1-\omega_{j^*})^{k+1}$
  $q_k \leftarrow b_k/k$ and $q_{k+1} \leftarrow d_{k+1}/(k+1)$
  $A \leftarrow (p_{k+1}q_{k+1}J)/(p_k q_k g)$
  if $v \leq \min(1, A)$ **then** accept birth

4. Generate σ_b^t from the gamma distribution with scale parameter $\beta + n_0^t$ and shape parameter $\nu + \sum_{i=1}^{n} I(z_i = 0)y_i$.

Further examples can be found in [12, 15].

Step 2

The second step implements the trans-dimensional jump which increases the number of components of the mixture or decreases it by one. The choice is made randomly with equal probabilities. New components are added to the model through either a split or a birth move; a component is removed from the model using a combining or death move [12]. These four steps allow the algorithm to explore the entire map and to search for new sources without affecting the distribution of the background radiation (3). A main difference to [12] is that we allow the algorithm to remove a component from the model using the death move also when it is not empty. This step was introduced to assure interpretability not only of the model parameters, as required by physicist, but also of the steps of the algorithm. Removing a non empty cluster essentially amounts to classifying it as a false positive. This turns out to be quite often the case when we get close to the Galactic equator whose influence is still tangible despite we cut off most of it by limiting the latitudes and longitudes of the explored sky region.

The code boxes of Algorithms 1 and 2 list the pseudo code for the split and the birth moves. Note that they also provide the pseudo code for the combining and the death moves we use to down size by one the number of components of the mixture. So, for instance, to evaluate whether to reduce the number of sources from K to $K-1$ by combining two of them, we interchange $K-1$ and K in the split move outlined in Algorithm 1. The acceptance probability is then $\min\{1, 1/A\}$ instead of $\min\{1, A\}$.

4 Modelling the *Fermi* LAT Data

We applied model (4) to the *Fermi* LAT data described in Sect. 2 and shown in Fig. 2. The corresponding sky region is framed in white in Fig. 1 and covers broadly one fourth of the area observed by the LAT. Recall, furthermore, that the third catalogue of hard *Fermi* LAT sources lists 288 high-energy γ-ray emitting sources for this sector [4]. The 3FHL catalogue will furthermore be used to benchmark the detection capability of our model. We run our reversible jump MCMC algorithm, as described in the previous section, for a total of 20,000 iterations each. The number and directions of the sources present in the 3FHL catalogue were used as starting points for K and μ, respectively. This way, we acknowledge all the a priori available information. The starting points for the mixture weights ω and of the scale parameter σ_b, which characterises the distribution of the background radiation, were randomly drawn from their a priori distributions.

The left panel of Fig. 4 shows the posterior distribution of K, the supposed number of high-energy γ-ray sources present in the analysed region. The posterior mode is $K = 331$, a value which was visited 1,892 times, that is, by around 9.5% iterations. We compared the posterior modes of (μ_{jx}, μ_{jy}), $j = 1, \ldots, K$, for these 331 putative sources with those present in the 3FHL catalogue: appreciably, our algorithm confirmed 255 of the acknowledged ones. The nature of the 76 remaining detections needed be investigated. We will come back to this point shortly. The right panel of Fig. 4 traces the 1,892 values generated for σ_b, and shows a good mixing property of the chain. The posterior mode is 0.0287, slightly higher than what expected on average a priori, with 95% highest posterior density (HPD) interval [0.0284, 0.0289]. These values are also shown in Fig. 4 as solid and dashed horizontal lines, respectively. Most interestingly, however, is the Bayesian estimate of $\omega_0 = 0.9387$ with 95% HPD interval [0.9364, 0.9407]. Remember that this value quantifies the inten-

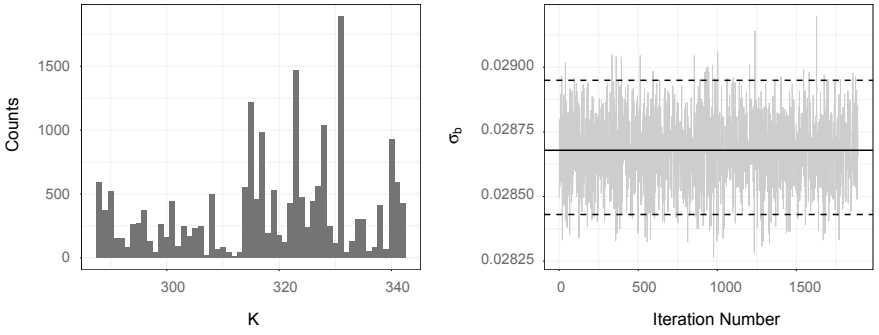


Fig. 4 Summary and diagnostic plots for the fitted model. Left: posterior distribution of K, the putative number of sources present in the analysed sky region. The modal value $K = 331$ is visited 1,892 times out of 20,000. Right: trace plot of the corresponding 1,892 σ_b values. The solid horizontal line at the center represents the posterior mode; the two dashed lines delimit the 0.95% highest posterior density interval

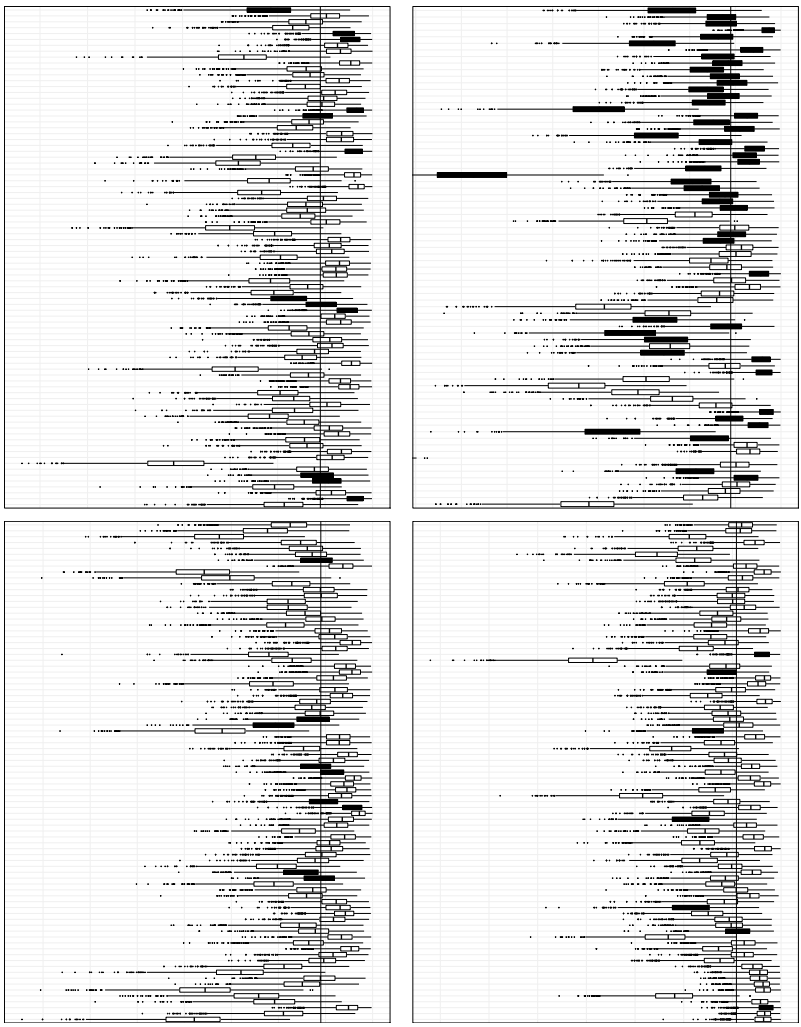


Fig. 5 From left to right and from top to bottom: boxplots of the posterior distribution for the mixing proportions ω_i, $i = 1, \ldots, 331$. In white: catalogued high-energy γ-ray sources; in black: new candidate sources. The solid horizontal lines represents the intensity threshold used to further skim the new detections

sity of the diffuse background radiation: it results that around 94% of the detected photons originated from it. Differently stated, only 6% of the photons were emitted from around 300 sources whose median intensity is 0.000137.

To further discriminate whether the 76 newly identified clusters correspond to real γ-ray emitting sources, we heuristically used the a posteriori available information on their intensities. Figure 5 shows the asymmetric boxplots of the posterior distributions of the 331 mixing proportions ω_i, $i = 1, \ldots, 331$. The white boxes correspond to the 255 already known sources, while the new candidates are drawn in black. Our ad hoc procedure defines the median of the posterior modes for the 255 catalogued sources as the threshold intensity above which we may expect a γ-ray emitting source. We hence qualified the 33 clusters whose posterior modes for ω_i satisfy this criterion as possible undetected sources. Their coordinates are currently being tested as prescribed by the *Fermi* LAT collaboration [4].

5 Conclusions

The results obtained for our model when applied to the *Fermi* LAT data of the limited sky region described in Sect. 2 are rather encouraging. We were able to detect 255 already known sources and to pinpoint possible new candidates. Of the 288 catalogued sources 33 were missed because their signal most likely isn't strong enough to be captured by our model but gets confounded with the prominent and irregularly shaped background radiation which pervades the considered area even after the initial skimming. The opposite holds for the 43 initially identified and successively declassified sources which probably correspond to small areas of excess background intensity. This aspect represents one of the improvements of our model we are currently working on. The proposed parametric formulation for the diffuse background radiation is, in fact, only partially efficient. Using further data provided by the *Fermi* LAT collaboration we are currently developing a more precise background model.

Further future developments focus on both, theoretical and computational aspects. A first aspect regards the distribution used to describe how photons scatter around their emitting source. King's PSF used as approximation in (2) is currently being replaced by the point spread function proposed in [1]. The truncated Poisson distribution (6) could in principle be replaced by a negative binomial distribution, that is, by adding a further level of hierarchy having the Poisson mean being distributed as a gamma with suitable scale and shape parameters. We preferred to take a different route and are currently working on two research strains which correspond to: (i) incorporating into the parametric formulation of the background model all information provided on it by the Fermi LAT project, and (ii) developing a nonparametric formulation based on smoothing splines for the background model. On the computational side, we are replacing the Metropolis-Hastings step used to generate the values of μ with a more efficient Gibbs sampler. Last but not least, the heuristic approach adopted at the end of the previous section to qualify the newly detected sources needs

be replaced by a formal procedure which accounts also for the available, here not used, information on the energy level of each detected photon.

Acknowledgements We would like to thank Dr. Mauro Bernardi for the most helpful discussion of reversible jump MCMC and its extension. We furthermore express our vote of thanks to an anonymous Referee for her/his most careful reading, especially with respect to our model formulation and estimation. This research was supported by SID 2018 grant "Advanced statistical modelling for indexing celestial objects" (BIRD185983) awarded by the Department of Statistical Sciences of the University of Padova. Financial support by Prof. Junhui Fan (grants n. NSFC11733001 and n. NSFC U1531245) is gratefully acknowledged.

References

1. Ackermann, M., et al.: Determination of the point-spread function for the Fermi large area telescope from on-orbit data and limits on pair halos of active galactic nuclei. Astrophys. J. **765**(1), 54 (2013)
2. Ackermann, M., et al.: The spectrum and morphology of the Fermi bubbles. Astrophys. J. **793**, 64 (2014)
3. Ackermann, M., et al.: The spectrum of isotropic diffuse gamma-ray emission between 100 MeV and 820 GeV. Astrophys. J. **799**, 86 (2015)
4. Ackermann, M., et al.: 3FHL: the third catalog of hard Fermi LAT sources. Astrophys. J. Suppl. Ser. **232**, 18 (2017)
5. Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**, 711 (1995)
6. Guglielmetti, F., Fischer, R., Dose, V.: Mixture modeling for background and sources separation in x-ray astronomical images. AIP Conf. Proc. **735**, 111 (2004)
7. Hobson, M.P., Rocha, G., Savage, R.S.: Bayesian source extraction. In: Hobson, M.P., et al. (ed.) Bayesian Methods in Cosmology, p. 167. Cambridge University Press, Cambridge (2010)
8. Jones, D.E., Kashyap, V.L., van Dyk, D.A.: Disentangling overlapping astronomical sources using spatial and spectral information. Astrophys. J. **808**, 137 (2015)
9. King, I.: The structure of star clusters. I. An empirical density law. Astrophys. J. **67**, 471 (1962)
10. Kraft, R.P., Burrows, D.N., Nousek, J.A.: Determination of confidence limits for experiments with low numbers of counts. Astrophys. J. **374**, 344 (1991)
11. Mattox, J.R., et al.: The likelihood analysis of EGRET data. Astrophys. J. **461**, 396 (1996)
12. Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components (with Discussion). J. R. Stat. Soc. Ser. B (Statistical Methodology) **59**, 731 (1997)
13. Stein, N.M., et al.: Detecting unspecified structure in low-count images. Astrophys. J. **813**, 66 (2015)
14. van Dyk, D.A., et al.: Analysis of energy spectra with low photon counts via Bayesian posterior simulation. Astrophys. J. **548**, 224 (2001)
15. Wiper, M., Insua, D.R., Ruggeri, F.: Mixtures of gamma distributions with applications. J. Comput. Graph. Stat. **10**, 440 (2001)

Bayesian Estimation of Causal Effects in Carcinogenicity Tests Based upon CTA

Federico M. Stefanini and Giulia Callegaro

Abstract Despite more than 30,000 chemical substances are currently produced or imported in the European Union in volumes of 1 ton or more per year, they remain widely yet to be tested for carcinogenicity. Cell Transformation Assays (CTAs) are cheap and fast in vitro methods developed to screen chemical substances without resorting to animal-based testing. Here we propose two models for potential outcomes to estimate causal effects of different concentrations of a candidate carcinogen on counts of Type III *foci* growing within Petri dishes. A comparison of our proposals with simpler alternatives suggested in the literature for the BALB/c 3T3 CTA protocol is performed using the LOO information criterion. Here we overcome data manipulations recently proposed in the literature by introducing a flexible class of models based on experts' belief that do not necessitate of: (i) adding fake observations to actual data; (ii) making cumbersome transformations to original counts; (iii) constraining distributions at low concentrations to have a variance larger than the mean. Open issues are discussed in relation to the current practice adopted to perform multi-laboratory experiments on the same substance.

Keywords Causal model · Potential outcomes · Cell transformation assays
In vitro carcinogenicity testing

MSC2010 62F15 · 62P10

F. M. Stefanini (✉)
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze,
Viale Morgagni 59, 50134 Florence, Italy
e-mail: federico.stefanini@unifi.it

G. Callegaro
Dipartimento di Scienze dell'Ambiente e della Terra, Università di Milano Bicocca,
Piazza della Scienza, 20126 Milano, Italy
e-mail: g.callegaro@campus.unimib.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_12

1 Introduction

The attitude of a chemical substance to induce, or to foster the induction of, cancer is called carcinogenicity [18]. European Regulations detail the duties of Member States related to chemical testing, also for carcinogenicity testing. For example, the aim of Registration, Evaluation and Authorisation of Chemicals (REACH) is to systematically evaluate the risks, to human health and environment, of more than 30,000 chemical substances that are produced or imported in the EU in volumes of 1 ton or more per year [17], although they remain widely untested.

Chemical carcinogenicity is firstly assessed by a battery of in vitro and in vivo genotoxicity tests followed by the life-time cancer rodent bioassay [13, 21, 22], the method currently required at regulatory level. Rodent bioassays require more than two years for execution and dozens of animals are involved: it is evident the need of faster and cheaper techniques for screening purposes, like CTAs [5], to allow to cover such a large (and increasing) number of chemicals with carcinogenicity testing. Cell Transformation Assays (CTAs) are test methods increasingly used in the assessment of the carcinogenic potential of compounds. A CTA is an in vitro method in which several Petri dishes are seeded with immortalized cells and later treated with a solution of the chemical substance to test. The endpoint of transformation is the number of fully transformed *foci* grown within each Petri dish, also called Type III *foci*. The type of cell systems and the adopted laboratory protocol jointly define features of a CTA system: the BALB/c 3T3 CTA based on an immortalized cell line derived from embryonal murine fibroblasts is considered in the following because it is able to detect genotoxic carcinogens and some non-genotoxic carcinogens.

In the literature on CTAs, some efforts have been directed towards assessment and improvement of frequentist data analysis. Bretz and colleagues [3] suggested an approach to test possible down turns of expected values at higher doses. Ponti and colleagues [26] emphasized that t-test is not suited to CTA data and they proposed Fisher's exact test instead. The need for statistical analysis tailored to the specific features of CTA relying on BALB/c 3T3 system has been recently recognized by an international expert group at the European Centre for the Validation of Alternative Methods (ECVAM), who formulated recommendations in the EURL ECVAM Prevalidation Report [5, 7]. In a recent proposal [11], negative binomial models were fit to the outcome made by *foci* counts, and in case of a bad fitting a general linear model was considered after transforming raw counts. Unfortunately, it has been empirically found [33] that both classes of models may be inappropriate in a given CTA. As far as we know, a Bayesian analysis of CTAs from the standpoint of the causal framework based on potential outcomes [28] has not been ever proposed in the literature.

In the following sections, Bayesian estimates of causal effects on the number of fully transformed *foci* after treatment with a (candidate) carcinogenic chemical are provided in the potential outcomes framework [28, 29]. Here the number Type III *foci* observed within each Petri dish is compared to counterfactual counts that would have been observed, had the treatment being vehiculus or a different concentration, by

means of Bayesian posterior predictive imputation [12, Sect. 8.4]. In Sect. 2.1 below, we start describing the biological features that characterize a typical CTA protocol, then in Sect. 2.2 a case study is introduced to illustrate the approach. In Sect. 2.3 beliefs and recent recommendations from the literature are listed to make clear the context of the proposed models. In Sect. 2.4, the structure of the causal model is detailed using notation close to [10], while in Sect. 2.5 smooth models are proposed and compared (Sect. 3) to common poisson and negative binomial alternatives. The discussion in Sect. 4 closes this work together with some considerations on future work on CTAs.

2 Cell Transformation Assays: From Features to Statistical Models

In this section we start with the description of typical CTA protocols, then recommendations from the literature are introduced and, finally, smooth models for counts are proposed.

2.1 CTAs: Main Features of Different Protocols

Neoplastic transformation in vitro is a progressive event analogous of in vivo carcinogenesis [1]. Following this principle, cell transformation has been defined as the induction of certain phenotypic alterations in cultured cells that are characteristic of tumorigenic cells [1]. The stepwise process of in vitro transformation leads to several cellular alterations, including:

- the acquisition of infinite life-span (immortalization);
- changes in morphology (e.g. spindle-shape morphology);
- changes in growth pattern (e.g. criss-cross and multilayered growth of the cultured cells);
- aneuploid and genetic instability;
- anchorage-independent growth (e.g. colony formation in soft agar);
- the ability to induce tumours in vivo [2].

Accordingly, the Cell Transformation Assay exploits these concepts as it was developed to mimic the multistage nature of carcinogenesis [36].

The cultured cells suitable to study in vitro transformation must have a low incidence of spontaneous transformation rate and be sensitive to the neoplastic transformation by exposure to a carcinogen. Typically adopted systems are based on rodent cell lines: BALB/c 3T3 and C3H10T1/2 cells, immortalized fibroblasts of rodent origin, and Syrian Hamster Embryo (SHE), that are primary cells [6, 15, 16]. Recently, Bhas 42 cells were established as a clone by the transfection with the *v-Ha-ras* gene

into mouse BALB/c 3T3 A31-1-1 cells and their subsequent selection based on their sensitivity to 12-O-tetradecanoylphorbol-13-acetate (TPA) [31].

Thanks to their promising properties, CTAs gained the attention of the regulatory agencies: in 2007 the OECD published a Detailed Review Paper on Cell Transformation Assays for detection of chemical carcinogens [20], and in 2012 and 2013 the European Union Reference Laboratory for alternatives to animal testing (EURL ECVAM) published two Recommendations for Cell Transformation Assays, using BALB/c 3T3, SHE and Bhas 42 systems [8, 9]. In addition, OECD recently published two Guidance documents for SHE and Bhas 42 CTAs [23, 24].

All the system protocols share the same endpoint: the formation of colonies/*foci* of transformed cells upon treatment in culture with a suspected carcinogen. The transformed cells acquire phenotypic alterations typical of malignant cells and have the ability to form invasive tumours in susceptible animals [1, 14, 27]. In this regard, CTAs have a clear biological connection with cancer.

CTA based on BALB/c 3T3 cell line includes a preliminary cytotoxicity test or dose-range finding phase, followed by the transformation assay. A cytotoxicity test is carried out prior and/or in parallel to the transformation assays to select the optimal range of test chemical concentrations for the transformation assays and to evaluate cytotoxicity of each treatment. The transformation assays must test at least five doses and positive and negative controls must be tested as well. Positive controls are usually 3-methylcholanthrene (MCA) and TPA as tumour promoter, while negative controls should comprise also the vehicle compound. Preferred vessels are 100 mm Petri dish, 10 for each group tested.

The transformation assay starts with the seeding at low density ($2 * 10^4$ cells/100 mm dish), followed after one day by the treatment with the suspected carcinogen. The treatment with the compound can last up to 4 days and it is followed by a long recovery phase when medium is changed regularly. After 27–28 days dishes are methanol-fixed and Giemsa-stained for final microscope observation. Transformed colonies, called *foci*, are visually scored using a stereomicroscope following coded morphological criteria. Three types of *foci* have been distinguished (I, II, III), although it is likely that a continuum of focal phenotypes exists [15]. Type I *foci*, which are more tightly packed than the normal monolayer of cells and only slightly basophilic, are not scored since they do not give rise to neoplastic growths upon injection into irradiated mice. Type II *foci* display massive piling up into virtually opaque multilayers, the cells are moderately polar and criss-crossing is not pronounced. Type III *foci* are highly polar, fibroblastic, multilayered, criss-crossed arrays of densely stained cells. Invasive misoriented cellular projections radiating into the surrounding density-inhibited confluent monolayer of nontransformed cells are sometimes seen in Type III *foci*. Type III *foci* are scored as positive in BALB/c 3T3 CTA. The outcome in CTAs is the number of Type III *foci* per dish.

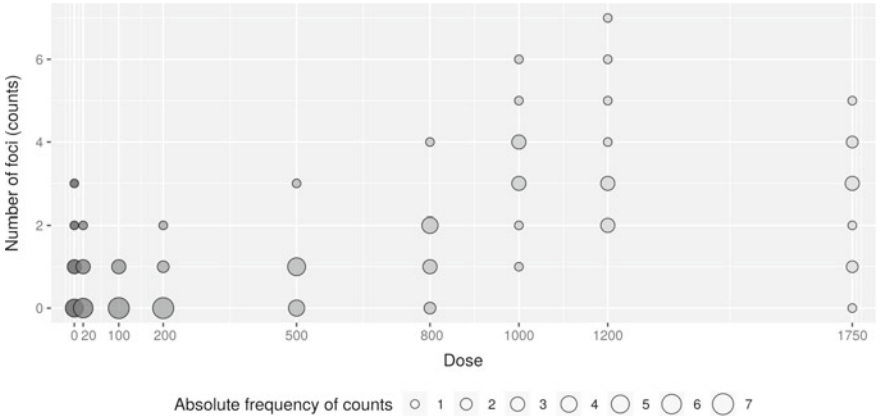


Fig. 1 Scatterplot of raw data: the size of points increases with the absolute frequency of counts

2.2 A CTA on O-Toluidine: A Case Study

The o-toluidine (CAS chemical registry number # 636-21-5) case study is briefly described below. Further toxicological details are available in published work [34].

A total of eight different doses plus the negative control are shown in Fig. 1, where diameter of circles represents the absolute frequency of counts. Doses of tested chemical are (μg/ml): 0 (negative control, $W = 1$), 20 ($W = 2$), 100 ($W = 3$), 200 ($W = 4$), 500 ($W = 5$), 800 ($W = 6$), 1000 ($W = 7$), 1200 ($W = 8$), 1750 ($W = 9$). A total of 90 Petri dishes containing BALB/c 3T3 cells sampled at the same passage from the original cell culture were treated after random assignment of each concentration to 10 dishes (replicates) for each concentration. All experimental units received protocol components taken from the same batch, including medium and serum. After 4 weeks from treatment, Petri dishes were visually scored under a light microscope and the number of Type III *foci* within each dish counted.

2.3 CTAs: Assumptions, Beliefs, Recommendations

Several features common to almost all CTAs make the quantitative analysis of counts not trivial. First, there is a high variability in the dose-response relationship, which is typically non-monotone and dependent on the considered chemical. Second, substantial differences of variance are often observed at different concentrations. Third, sample size is not less than (but almost never greater than) 10 observations for each concentration. Fourth, the variance may be null or smaller than the mean in samples at low concentrations of chemical. Last, replications of the same experiment on the same chemical in different laboratories quite often may provide inconclusive findings [7, p. 44].

The document elaborated by ECVAM's Expert Group [11, EEG] contains the most recent recommendations for the analysis of BALB/c 3T3 CTA experiments. For a detailed description of the BALB/c 3T3 CTA protocol see Sasaki et al. [30]. Here we maintain the experimental context and beliefs settled in Hoffmann et al. [11]:

1. the outcome is 'number of fully transformed *foci*', called Type III *foci*, on a Petri dish;
2. the number of Petri dishes at each concentration (dose) is typically 10 (never below 9);
3. five to ten concentrations of a test chemical are randomly assigned to experimental units (homogeneous Petri dishes);
4. *focus*-inducing chemicals are expected to show non-monotone dose-concentration relationships;
5. positive controls serve for quality assurance purposes only;
6. at small concentrations the empirical distribution of counts may be degenerate, typically at zero;
7. concentrations have to be considered as levels of a qualitative factor, although originally on a quantitative scale (e.g. μg/ml), given that EEG's recommendation discourages the use of more elaborated quantitative dose-response relationships;
8. the probability mass function of Type III *foci* counts is smooth at every concentration of a given chemical;
9. concentrations are chosen so that the lowest one is likely to behave like the vehiculus, the highest is above the 50% lethal dose (LD50);
10. high concentrations of a *focus*-inducing chemical bear cytotoxicity, an event that shrinks *foci* formation.

2.4 Potential Outcomes in a CTA

Let $Y_j^{<i>}$, $j = 1, 2, \ldots, n$ be random variables representing potential outcomes for the number of Type III *foci* within Petri dish $j = 1, \ldots, n$ under dose level (treatment) $i \in \Omega_D = \{1, \ldots, L\}$; the sample space of each count is $\Omega_Y = \{0, 1, \ldots, C\}$, whatever the dose of chemical and whether observed or counterfactual. In a given experiment the maximum value that can be observed is C and it may vary according to the size of Petri dishes between 20 and 50. In the case study presented below, Petri dishes are all of equal size and $C = 30$. The potential outcomes at dose level i define the vector $Y^{<i>} = (Y_1^{<i>}, \ldots, Y_n^{<i>})^T$ and the vector of indicators of treatment assignment is $W = (W_1, \ldots, W_n)^T$ with the cartesian product $\Omega_W = \{1, \ldots, L\}^n$ as sample space. The probability mass function of potential outcome $Y_j^{<i>}$ at concentration $i \in \Omega_D$ is:

$$p_{Y_j^{<i>}}(y \mid \pi_{i,0}, \pi_{i,1}, \ldots, \pi_{i,C}] = \sum_{k=0}^{C} \pi_{i,k} \; I_{\{k\}}(y) \tag{1}$$

where $(\pi_{i,0}, \pi_{i,1}, \ldots, \pi_{i,C})$ is the vector of probability values, thus $\sum_k \pi_{i,k} = 1$; $I_G(y)$ is the indicator function equal to 1 if $y \in G$ and zero otherwise.

Given that CTAs belong to the class of randomized experiments, the assignment mechanism is ignorable and characterized by unit-level probability of treatment assignment in the interval (0, 1). The probability mass function of vector W that represents the assignment mechanism [12, chap. 3] is based on the multinomial coefficient:

$$p(w \mid y^{<1>}, \ldots, y^{<L>}) = \binom{n}{n_1, n_2, \ldots, n_L}^{-1} \tag{2}$$

for all w satisfying $\sum_{j=1}^{n} I_{\{i\}}(w_j) = n_i$ at each i, with n_i the number of experimental units treated with concentration level i. Under row (unit) exchangeability of matrix $(Y^{<1>}, \ldots, Y^{<L>})$ the joint distribution of potential outcomes is:

$$p\left(y^{<1>}, \ldots, y^{<L>}\right) = \int \prod_{j=1}^{n} p(y_j^{<1>}, \ldots, y_j^{<L>} \mid \theta)\ p(\theta)\ d\theta \tag{3}$$

where θ is a vector of model parameters belonging to the parameter space Θ.

The elicitation of conditional distributions for potential outcomes given model parameters (Eq. 3), often called "the science", should take into account the main processes driving the emergence of *foci*. Even if it is not carcinogenic, a chemical may exert a toxic effect on cultured cells, thus causing a reduction in the final number of Type III *foci*. If a chemical is carcinogenic then it is expected to stimulate the emergence of *foci*, but this driving force also depends on concentration: too low doses are ineffective, too high doses are often cytotoxic. Despite that concentrations are selected to be within a convenient range, it is quite difficult to anticipate any correlation between potential outcomes. For these reasons, conditional independence among potential outcomes is assumed:

$$p\left(y_j^{<1>}, \ldots, y_j^{<L>}, \theta\right) = \prod_{i=1}^{L} p(y_j^{<i>} \mid \theta_i)\ p(\theta_i) \tag{4}$$

where the joint distribution of model parameters is factorized into marginally independent subvectors, $\theta = (\theta_1, \ldots, \theta_i, \ldots, \theta_L)$.

At the end of an experiment, the $n \times 1$ vector of observed potential outcomes $Y^{obs} = (Y_1^{obs}, \ldots, Y_n^{obs})^T$ has elements $Y_j^{obs} = \sum_{i=1}^{L} Y_j^{<i>} I_{\{i\}}(W_j)$, with $j = 1, \ldots, n$. The set of unobserved potential outcomes for the experimental unit j gathers all values of i out of the assigned W_j, that is $Y_j^{mis} = \left\{ Y_j^{mis,<i>} : i \in \{1, \ldots, L\} \setminus W_j \right\}$, therefore the collection of counterfactuals is $Y^{mis} = \{Y_1^{mis}, \ldots, Y_n^{mis}\}$.

The causal estimand considered in this work is the average difference of pairs of potential outcomes, where averaging is performed over a subset of the finite sample:

$$\tau_{fs}(i,r) = \frac{\sum_{j=1}^{n}(y_j^{obs} - Y_j^{mis,<r>})I_{\{i\}}(w_j) + \sum_{j=1}^{n}(Y_j^{mis,<i>} - y_j^{obs})I_{\{r\}}(w_j)}{\sum_{j=1}^{n} I_{\{i\}}(w_j) + \sum_{j=1}^{n} I_{\{r\}}(w_j)} \tag{5}$$

thus for $i = 2, \ldots, L$ the effect of a tested chemical with respect to the negative control (water) is $\tau_{fs}(i, 1)$, while values of $\tau_{fs}(i, i-1), i > 1$, quantify changes of causal effect due to the increase of concentration. In order to calculate tau using Eq. (5), the conditional predictive distribution of Y^{miss} given y^{obs} and w may be exploited to impute counterfactuals in Y^{mis}.

The likelihood function is built from the probability mass function in Eq. (1):

$$L(\pi_{1,0}, \pi_{1,1}, \ldots, \pi_{i,k}, \ldots \mid y^{obs}, w] = \prod_{j=1}^{n} \left(\sum_{k=0}^{C} \pi_{w_j,k}\ I_{\{k\}}\left(y_j^{obs}\right) \right) \tag{6}$$

where (y^{obs}, w) are vectors of observed counts and treatment assignments.

Note that, in some models recommended in the literature [11], the value of C is implicitly set to infinity, as it happens if the Poisson family of distributions is preferred in the elicitation instead of Eq. (1).

2.5 *Elicitation of Expert Beliefs*

Carcinogenicity is a multistep process characterized by inherent high heterogeneity of mechanisms and variability, thus no wonder that experts do not agree about how to allocate the probability mass over counts, despite that some shared beliefs exist. A plausible upper limit for size of counts is $C = 30$ in small Petri dishes, because the available physical space is limited: probability mass functions should not allocate a relevant portion of the distribution above C. Furthermore, small changes of probability values should occur along subsequent *foci* counts, even when sampled counts are all equal, for example, to 0; in this case Hoffmann et al. [11] suggested to artificially increase the sample of one (fake) Petri dish whose number of Type III *foci* is artificially set equal to one. The proposed manipulation is intended to provide a smooth estimate of probability values on *foci* counts, and in particular a point estimate of variance strictly greater than zero. Here, we preferred to explicitly address beliefs about smoothness by developing models based on the ratio of probability values in subsequent counts.

Two classes of models for potential outcomes are proposed in this section: (i—USMs) the class of unimodal smooth models for a strong degree of belief about the presence of just one mode; (ii—SMOs) the class of smooth models if only smoothness is highly plausible. Expert beliefs and numerical evidence may drive model selection, although model averaging is also an option (not considered here).

In the class of SMOs, the initial (prior) distribution on probability values for counts was elicited by defining logits between pairs of subsequent counts:

$$\psi_{i,k} = ln\left(\frac{\pi_{i,k}}{\pi_{i,k-1}}\right) \sim Normal(0, \tau_i); \quad k = 1, \ldots, C; \quad i = 1, \ldots, L \tag{7}$$

where the variance parameter τ_i regulates the amount of smoothness in the probability mass function defined by Eq. (6). By exploiting the probability simplex, we have:

$$\pi_{i,0} = \frac{1}{1 + \sum_{r=1}^{30} exp\left(\sum_{s=1}^{r} \psi_{i,s}\right)} \tag{8}$$

$$\pi_{i,k} = \frac{exp\left(\sum_{s=1}^{k} \psi_{i,s}\right)}{1 + \sum_{r=1}^{30} exp\left(\sum_{s=1}^{r} \psi_{i,s}\right)}. \tag{9}$$

Initial distributions of $\tau_i, i = 1, 2, \ldots, 9$ were elicited as marginally independent members of the Exponential family of distributions $\tau_i^{-1} \sim Exponential(3.5)$. The final expected value of each parameter $\pi_{i,k}$, $k = 0, 1, \ldots, 30$, was inspected after conditioning to a degenerate sample of counts on 0 (Fig. 2, top right panel). At the end of elicitation, the probability mass located outside the observed count was close to $1/11$ and concentrated on *foci* counts not greater than 3. A similar inspection was performed for an empirical distribution fully concentrated on 1, and also in this case smoothing is apparent (Fig. 2, bottom right panel). The distributions in Fig. 2, right panels, are close to what proposed in [11], but the inherent uncertainty here is not neglected. Note that negative controls, as well as dishes treated with low concentrations of tested compound, may also result in a sample where counts are all concentrated on 1 (one *focus* per dish), thus the artificial increase of sample size after adopting the suggestion in [11] could reach 20% of the original sample in this case.

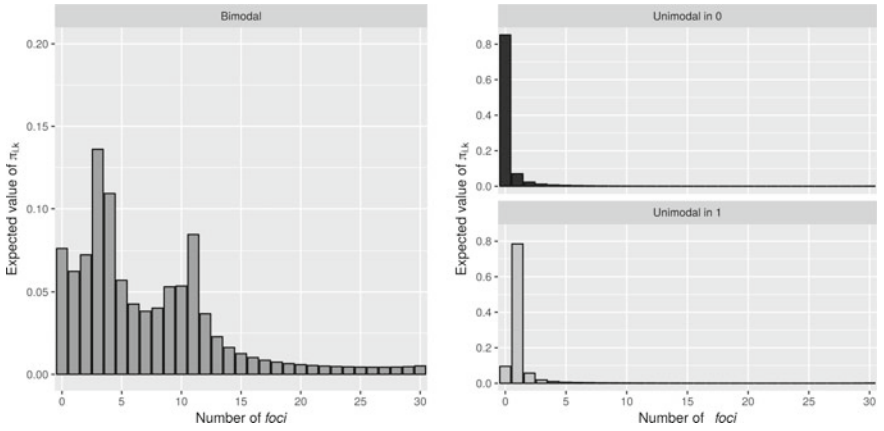


Fig. 2 Expected value of the probability of each count given three virtual samples of 10 observations each: on the left, a multimodal sample; top-right a sample with all counts equal to 0; bottom-right a sample with all counts equal to 1

In Fig. 2 (left panel), the expected value of the probability at each count is shown given a virtual sample with two local modes quite far apart. Expected values at each count smoothly changes along counts, in close accordance with prior beliefs.

In the class of USMs, the initial (prior) distribution on probability values for counts at dose i was elicited by selecting a value M_i for the mode and then by assigning a distribution to logits calculated between pairs of subsequent counts, given the location of M_i:

$$\psi_{i,k} = ln\left(\frac{\pi_{i,k}}{\pi_{i,k-1}}\right) \sim Beta(2,2), \quad k \leq M_i \tag{10}$$

$$\psi_{i,k}^{-1} \sim Beta(2,2), \quad k > M_i \tag{11}$$

where Beta stands for the Beta probability density function: hyper-parameters values were elicited plotting the expected value of parameters $\{\pi_{i,k}\}$ given a virtual sample concentrated on one count value.

The predictive performance of our models was compared against the Bayesian counterpart of typical models with a limited number of parameters. In the Poisson model (POI), parameters take value according to the probability mass function:

$$\pi_{i,k} = Poisson(k \mid \lambda_i), \quad k = 0, 1, \ldots, \infty.$$

with $i = 1, \ldots, L$ the considered dose levels. Canonical parameters were considered marginally independent in the initial distribution, with $\lambda_i \sim Gamma(0.2, 0.2)$, after considering beliefs about the expected value of lambdas and about intervals of highly plausible values. The Negative Binomial model (NBM) was recommended in [11]: the probability of a count equal to k at dose i is

$$\pi_{i,k} = NegativeBin(y_k \mid \mu_i, \phi_i) = \binom{y_k + \phi_i - 1}{y_k} \left(\frac{\mu_i}{\mu_i + \phi_i}\right)^{y_k} \left(\frac{\phi_i}{\mu_i + \phi_i}\right)^{\phi_i}$$

with μ_i the expected value of $Y_j^{<i>}$ and ϕ_i the scale parameter at dose i. Marginally independent components were elicited in the initial distribution of hyper-parameters: $\mu_i \sim Gamma(7, 1)$ and $\phi_i \sim Exponential(1.5)$, thus they are also independent on doses.

3 Results

Computations were performed in R[1] using the R packages[2] *rstan, loo, ggmcmc*. Samples from the posterior distributions of model parameters were obtained by Markov Chain Monte Carlo (MCMC) simulation after implementing all models in the Stan

[1] https://www.r-project.org/.

[2] https://cran.r-project.org/web/packages/.

programming language [4, 32]. After each MCMC run, graphical and numerical output diagnostics were calculated to check chain convergence (autocorrelations, cross-correlations, running means, traceplots, potential scale reduction factors, Geweke diagnostics) without finding relevant evidence of violations.

Model selection was performed using the LOO criterion [35] both to find the best model in the class of USMs and to compare predictive performances across different classes of models. As regards USMs, preliminary MCMC runs were devoted to find the best model with respect to modes m_i at each dose i. Within the unimodal class of models, a successful attempt was also performed to reduce the number of parameters by aggregating the first few doses, given that a CTA experiment is always planned to include one or more doses known to be ineffective [30, p. 33]. Values of LOO information criterion were (model, LOOIC): Negative Binomial (NBM, 317.7), Unimodal Smooth (USM, 287.6), Smooth (SMO, 256.6), Poisson (POI, 255.9), Restricted Unimodal (smallest four doses aggregated, RUSM, 254.0).

The Poisson model obtained the best LOO score in the o-toluidine case study. Nevertheless, we decided to work with the Unimodal Smooth Model for several reasons. First, standard errors of LOOIC values were all in the interval (14.7, 16.5), thus a pure criterion-based selection should not be considered conclusive anyway, with such small sample size. Second, the RUSM model performed better than POI after aggregation of the lowest dose levels, thus there is the potential of a good fit also in the RUSM class. Third, we judged model flexibility to be the most important feature to preserve because we can't assume that o-toluidine is representative of dozens thousands chemicals yet to be tested. Fourth, in the literature on CTAs the Poisson model has been criticized for the impossibility of describing CTAs in which the parameter variance is smaller than the mean. Further investigations are required with many other chemicals before considering the USM model as a reference model in CTAs.

Conditional distributions of causal estimands $\tau_{fs}(i, 1),\ i = 2, \ldots, 9$ obtained from the Unimodal Smooth Model are shown in Fig. 3, left panel. The causal effect is possibly null for the first four doses, given that each 95% highest posterior density (HPD) credibility interval contains the null value (solid vertical line between the two vertical dashes lines). At the sixth dose, the 95% HPD interval does not contain 0 thus we claim that a positive effect of the chemical on *foci* counts may be present. A positive causal effect (increase) in the number of *foci* remains likely present for the higher doses.

In Fig. 3, right panel, eight distributions for $\tau_{fs}(i, i-1),\ i = 2, \ldots, 9$ are shown. Changes of causal effects across subsequent doses are small, if not null, for the first four doses. The 95% highest posterior density (HPD) credibility intervals contain 0 thus no turning point is declared. At doses $i = 6$ and $i = 7$, upward turning points are claimed because the null value is outside on the left of both the HPD intervals. At $i = 8$ the increase of causal effect is likely to be null, while at $i = 9$ weak evidence of a downward turning point (decrease of causal effect) is present, given that the null value is close to the right boundary of the 95% HPD interval.

A remark is due about the choice of causal estimands. Finite sample quantities were selected because the purpose of CTA tests is to provide evidences that transfor-

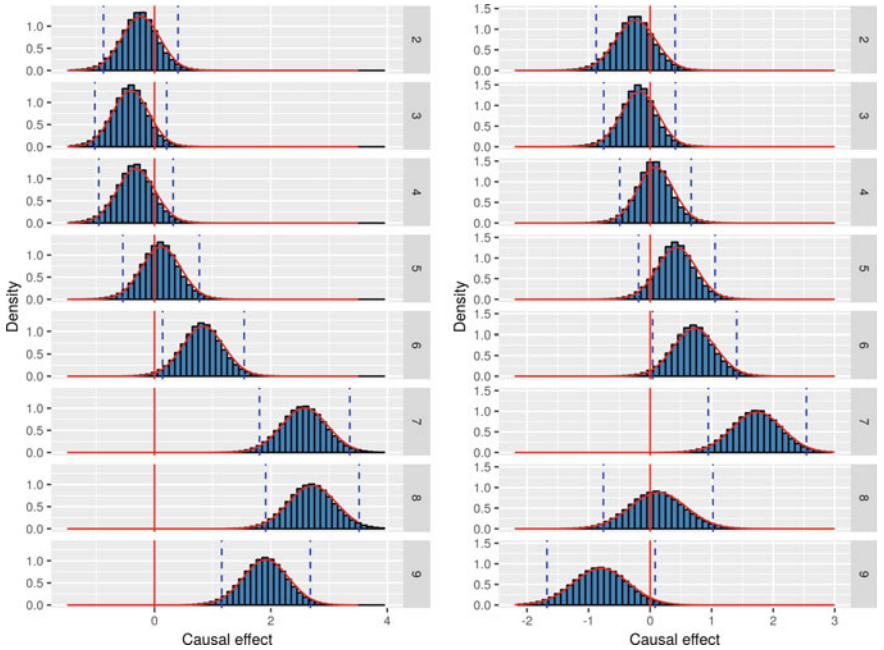


Fig. 3 Distribution of causal estimand $\tau_{fs}(i, 1)$ at dose i $(i = 2, \ldots, 9)$ given the vehicle $i = 1$ (left panel). Distribution of causal estimand $\tau_{fs}(i, i-1)$ at a dose i given dose $i-1$ (right panel). Vertical solid lines are located at zero. Vertical dashed lines define 95% credibility intervals obtained by Normal approximation (continuous line on top of histograms)

mation events increase in number due to triggering induced by a chemical substance. We foresee some difficulties in defining a reference population for the infinite collection of CTA experiments on a given chemical under test (see next section).

4 Discussion and Conclusions

The multistep nature of the carcinogenic process mimicked by a CTA may explain the need of flexible models in the analysis of CTA experiments. Large differences in the distribution of the outcome have been observed for the same chemical tested at different concentrations. Large variability is also present in the outcome of CTAs when the same chemical is tested in different laboratories at equal or close concentrations. The small sample size of CTA experiments performed in production does not typically suffice to definitely rule out competing classes of models.

Previous proposals discussed in the literature had the merit of clearly stating the inadequacy of Student-t distribution and of remaining accessible to researchers with basic training in statistics, even though several drawbacks were left unsolved. The

first class of models proposed in [11] is based on the Nishiyama transform [19] of original counts, $x_{i,j} = \sqrt{y_{i,j}} + \sqrt{y_{i,j} + 1}$, followed by a general linear model with diagonal matrix of variances-covariances. In a recent contribution [33], it has been shown that such class of models may be unsuited for the study at hand, for example because it provides plug-in estimates of the probability $P[X < 1 \mid \hat{\theta}]$ above 0.19 for several values of concentration, although this event is impossible like an original negative count. In [11], a second proposal based on the Negative Binomial family of models has been presented. Nevertheless, at low concentrations, it often happens that the empirical distribution of counts has a variance smaller than the mean, and it may be even degenerate into 0 or into 1. These features pose problems while fitting Negative Binomial models by maximum likelihood, thus an artificial increase of the sample size by one (virtual) unit has been recommended. Further alterations of the collected sample were suggested in [11] because a balanced design simplifies the statistical downturn test: the imputation of one missing values by the median of collected observations (10% of sample size increase) was adopted to recover full design balance. The statement "... bias expected is negligible ..." might be questioned, given that the median is calculated on 9 observations and that missing values could be generated not at random.

The "minor shortcomings" exploited in [11] have a Bayesian flavor, given that the above described data manipulations are rooted in the expert belief about what could have been found in larger, or complete, samples. We do not agree with the conclusion of authors in [11, right column, p. 40], who stated that the above alterations of collected data consist of "... negligible data manipulations", given that, at any rate, a relevant amount of uncertainty is neglected by considering actual data and virtual observations on equal footing. The proposed smooth models based on experts information offer a simple approach to estimate causal effects without introducing fake observations and without limiting the choice of families of distributions to those where the variance is equal or larger than the mean. This last feature is particularly important because predictive distributions of counts are exploited to impute counterfactuals, an essential ingredient in the estimate of causal effects. For these achievements being possible, the researcher must be prepared to state substantive beliefs by informative prior distributions, a practice requiring some training, if not the help of a facilitator [25].

The causal perspective on CTAs adopted in this work might throw some light on open issues, like the possibility of obtaining inconclusive evidence (IE) from multilaboratory replicates of the same CTA experiment [7]. The event IE is realized during multilaboratory replication if, at a given concentration of a chemical under test, estimated averages from several laboratories show very large differences, for example 3 or more times the standard error of the difference between means. Given that the same chemical is tested using the same reagents, such large discrepancies might surprise. Nevertheless, it is widely recognized that the number of passages, performed by sampled cells from the primary cell culture may exert a huge effect on the number of fully transformed *foci* at the end of an experiment. Similarly, different batches of serum may contain substantially different amounts of key substances, like growth factors, which are known to enhance cell replication and, possibly, transformation.

Thus, from a multilaboratory perspective, we provided the final distribution of causal estimands that are conditional to a given homogeneous batch of cells and reagents within an experiment.

A candidate approach to solve the occurrence of inconclusive evidence could be based on stratification. Nevertheless, explicit conditioning to batches of cells and reagents could be far from trivial. The toxicologist should not expect that sampled cells at the same number of passages from the primary culture necessarily behave the same way: the distribution over cells states should be considered in order to increase homogeneity of response during a CTA. Second, dosage (quantification) of growth factors and other important serum components is likely to be expensive and time consuming, thus at the end not accepted in production. Thus, further work could be directed towards the indirect assessment of growth factors, for example using positive controls, and towards the definition of the actual state-step of a batch of cells with respect to the response in a CTA.

Multilaboratory CTA experiments currently use different treatment versions with the aim of compensating differences among batches of reagents and cells. The recommended CTA protocol prescribes the selection of concentrations of a tested chemical after performing a preliminary test in which the dose resulting lethal for 50% of cells is discovered [7]. It follows that different laboratories may replicate the same experiment with different concentrations of the considered chemical to test, for example because cell survival differ in different laboratories. The current protocol runs in contrast with the EEG's recommendation to build models in which the concentration is considered as a qualitative factor. Therefore, it is of primary interest to investigate on model extensions that, besides exploiting prior information further, also consider the variable concentration on the original quantitative scale.

Acknowledgements We thank Chiara Urani and Raffaella Corvi for fruitful discussions about CTAs. We thank the anonymous reviewers for their careful reading of our manuscript and for comments and suggestions that remarkably helped to improve this work. This research was partially supported by University of Florence, frame 'Disegno e analisi di studi sperimentali e osservazionali per le decisioni'.

References

1. Barrett, J.C., Ts'o, P.O.: Evidence for the progressive nature of neoplastic transformation in vitro. Proc. Natl. Acad. Sci. U.S.A. **75**(8), 3761–3765 (1978)
2. Berwald, Y., Sachs, L..: In Vitro Cell Transformation with Chemical Carcinogens. Nature **200**, 1182–1184 (1963)
3. Bretz, F., Hothorn, L.A.: Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. ATLA-Altern. Lab. Anim. **31**(Suppl 1), 81–96 (2003)
4. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. J. Stat. Softw. **76**(1), 1–32 (2017)
5. Creton, S., Aardema, M.J., Carmichael, P.L., Harvey, J.S., Martin, F.L., Newbold, R.F., O'Donovan, M.R., Pant, K., Poth, A., Sakai, A., Sasaki, K., Scott, A.D., Schechtman, L.M.,

Shen, R.R., Tanaka, N., Yasaei, H.: Cell transformation assays for prediction of carcinogenic potential: state of the science and future research needs. Mutagenesis **27**, 93–101 (2012)
6. DiPaolo, J.A., Takano, K., Popescu, N.C.: Quantitation of chemically induced neoplastic transformation of BALB-3T3 cloned cell lines. Cancer Res. **32**(12), 2686–2695 (1972)
7. EURL-ECVAM Validation Management Team: BALB/c 3T3 Cell Transformation Assay Prevalidation study Report (2010)
8. EURL-ECVAM Recommendation on three Cell Transformation Assays (2012,2013)
9. EURL-ECVAM Recommendation Bhas-CTA (2013)
10. Hernán, M.A., Robins, J.M.: Causal Inference. Chapman & Hall/CRC, Boca Raton. forthcoming (2018), Downloaded on Dec. 2017 https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
11. Hoffmann, S., Hothorn, L.A., Edler, L., Kleensang, A., Suzuki, M., Phrakonkham, P., Gerhard, D.: Two new approaches to improve the analysis of BALB/c 3T3 cell transformation assay data. Mutat. Res.-Genet. Toxicol. Environ. **744**(1), 36–41 (2012)
12. Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences-An Introduction. Cambridge University Press, New York (2015)
13. Jacobs, M.N., Colacci, A., Louekari, K., Luijten, M., Hakkert, B.C., Paparella, M., Vasseur, P.: International regulatory needs for development of an IATA for non-genotoxic carcinogenic chemical substances. ALTEX-Altern. Anim. Exp. **33**(4), 359–392 (2016)
14. Kakunaga T.: A quantitative system for assay of malignant transformation by chemical carcinogens using a clone derived from BALB-3T3. Int. J. Cancer **12**(2), 463–473 (1973)
15. Landolph, J.R.: Chemical transformation in C3H 10T1/2 Cl 8 mouse embryo fibroblasts: historical background, assessment of the transformation assay, and evolution and optimization of the transformation assay protocol. IARC Sci. Publ. **67**, 185–203 (1985)
16. LeBoeuf, R.A., Kerckaert, K.A., Aardema, M.J., Isfort, R.J.: Use of Syrian hamster embryo and BALB/c 3T3 cell transformation for assessing the carcinogenic potential of chemicals. IARC Sci. Publ. **146**, 409–425 (1999)
17. Lilienblum, W., Dekant, W., Foth, H., Gebel, T., Hengstler, J.G., Kahl, R., Kramer, P.J., Schweinfurth, H., Wollin, K.M.: Alternative methods to safety studies in experimental animals: role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH). Arch. Toxicol. **82**(4), 211–236 (2008)
18. Maurici, M., Aardema, M., Corvi, R., Kleber, M., Krul, C., Laurent, C., Loprieno, N., Pasanen, M., Pfuhler, S., Phillips, B., Sabbioni, E., Sanner, T., Vanparys, P.: Genotoxicty and mutagenicity. ALTEX-Altern. Anim. Exp. **33**(Suppl 1), 117–130 (2005)
19. Nishiyama, H., Omori, T., Yoshimura, I.: A composite statistical procedure for evaluating genotoxicity using cell transformation assay data. Environmetrics **14**, 183–192 (2002)
20. OECD: Detailed Review Paper on Cell Transformation Assays for detection of chemical carcinogens. OECD Series on Testing and Assessment No. **31** (2007)
21. OECD: Test guideline 451: carcinogenicity studies. OECD guidelines for the Testing of Chemicals-Section 4 (2009)
22. OECD: Test guideline 453: combined chronic toxicity/carcinogenicity studies. OECD guidelines for the Testing of Chemicals-Section 4 (2009)
23. OECD: Guidance Document on the in vitro Syrian Hamster Embryo (SHE) Cell Transformation Assay - Series on Testing and Assessment No. **214** (2015)
24. OECD: Guidance Document on the in vitro Bhas 42 Cell Transformation Assay-Series on Testing and Assessment Number **231** (2016)
25. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T.: Uncertain Judgements: Eliciting Experts Probabilities. Wiley, New York (2006)
26. Ponti, J., Munaro, B., Fischbach, M., Hoffmann, S., Sabbioni, E.: An optimised data analysis for the BALB/c 3T3 cell transformation assay and its application to metal compounds. Int. J. Immunopathol. Pharmacol. **20**(4), 673–684 (2007)
27. Reznikoff, C.A., Brankow, D.W., Heidelberger, C.: Establishment and characterization of a cloned line of C3H mouse embryo cells sensitive to postconfluence inhibition of division. Cancer Res. **33**(12), 3231–3238 (1973)

28. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Edu. Psychol. **66**(5), 688–701 (1974)
29. Rubin, D.B.: Bayesian inference for causal effects: the role of randomization. Ann. Stat. **6**, 34–58 (1978)
30. Sasaki, K., Bohnenberger, S., Hayashi, K., Kunkelmann, T., Muramatsu, D., Phrakonkham, P., Poth, A., Sakai, A., Salovaara, S., Tanaka, N., Thomas, B.C., Umeda, M.: Recommended protocol for the BALB/c 3T3 cell transformation assay. Mutat. Res.-Genet. Toxicol. Environ. **744**, 30–35 (2012)
31. Sasaki, K., Umeda, M., Sakai, A., Yamazaki, S., Tanaka, N.: Transformation assay in Bhas 42 cells: a model using initiated cells to study mechanisms of carcinogenesis and predict carcinogenic potential of chemicals. J. Environ. Sci. Health, Part C **33**(1), 1–35 (2015)
32. Stan Development Team: Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0. (2017). http://mc-stan.org
33. Stefanini, F.M.: Causal analysis of Cell Transformation Assays. SIS 2017 Congress Acta, pp. 1–6 (2017)
34. Tanaka, N., Bohnenberger, S., Kunkelmann, T., Munaro, B., Ponti, J., Poth, A., Umeda, M.: Prevalidation study of the BALB/c 3T3 cell transformation assay for assessment of carcinogenic potential of chemicals. Mutat. Res.-Genet. Toxicol. Environ. **744**(1), 20–29 (2012)
35. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. **27**(5), 1413–1432 (2017)
36. Yamasaki, H.: Non-genotoxic mechanisms of carcinogenesis: studies of cell transformation and gap junctional intercellular communication. Toxicol. Lett. **77**(1–3), 55–61 (1995)

Performance Comparison of Heterogeneity Measures for Count Data Models in Bayesian Perspective

Meenakshi Sundaram Subbiah, Rajamani Renuka Devi, Michele Gallo and Mamandur Rangaswamy Srinivasan

Abstract Random effects model is one of the widely used statistical techniques in combining information from multiple independent studies and examine the heterogeneity. The present study has focussed on count data model which is comparatively uncommon in such research studies. Also the interest is to exploit the advantage of Bayesian modelling by incorporating plausible prior distributions on the parameter of interest. The study is illustrated with a data on rental bikes obtained from UC Irvine Machine Learning Repository. Results have indicated the impact of prior distributions and usage of heterogeneity estimators in count data models.

Keywords Meta analysis · Random effects model · Variance component · Poisson model

1 Introduction

The problem of understanding variability on subjects of interest and its association with other variables is more interesting and challenging in statistical inference. The problem continues to be the major objective even in the present era of massive data sets. The present work deals with one such fundamental inferential problem under

M. S. Subbiah
HCL Technologies Ltd, Chennai, India
e-mail: sisufive@gmail.com

R. Renuka Devi
Department of Statistics, Presidency College, Chennai, India
e-mail: redevir6@gmail.com

M. Gallo (✉)
Department of Human and Social Sciences, University of Naples - L'Orientale, Naples, Italy
e-mail: mgallo@unior.it

M. R. Srinivasan
Department of Statistics, University of Madras, Chennai, India
e-mail: mrsrin8@gmail.com

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_13

Bayesian perspective. The inherent advantage of handling Bayesian hierarchical modelling is exploited with the underlying nature as a Random Effects Model (REM). This model has been used extensively in many applications especially in medical, epidemiological, ecological, and in social sciences. The areas in medical studies can further be summarised as clinical trials, case-control studies, meta-analysis or systematic reviews.

The model is basically a hierarchical structure, mostly based on normal distribution in two stages. Though Non-normal models have been attempted in Bayesian literature [5, 21] majority of studies even today [11] follow normally distributed models using appropriate transformation of underlying parameters. REM has received active research attention for many decades and following is a partial list of significantly recent studies that have dealt with the measure of heterogeneity. These studies have included both frequentist and Bayesian procedures, but, most of them do not make any formal comparisons. However, the present work has been confined to Bayesian approaches and its implications.

Engels et al. [3] could be one of the earliest works in this millennium on meta-analysis. Higgins and Thompson [6], Warn [31] and Leonard and Duffy [12] have focussed on heterogeneity measure and Bayesian methods for REM. Subsequently, [2, 19, 20, 27, 28] have discussed variance estimators in REM. The list also include [17, 23, 25] for the impact on data sets with zero occurrences. Further, [9–11, 18, 24, 26] provide ample scope for comparison of heterogeneity measure and its persistent research attention.

Viechtbauer [29] is among the widely used sources for REM computation in R software [16]. Few notable applications include, [1, 13, 30] for transportation data analyses, Hillebrand [7] for ecological data analysis and Subbiah and Rajeshwaran [22] for sports data analysis. Further, the impact of a measure across different study effects are well studied in literature; for example, Hunter et al. [8] and Zwetsloot et al. [32] discussed the treatment of funnel plot (a graphical tool for publication bias) in meta-analyses of proportion and standardized mean difference. Nazarzadeh and Bidel [15] distinguished the choice between fixed and random effects model based on funnel plots using relative risk as an effect size measure.

The extensive review of similar studies has shown that the study effects of interest are basically odds ratio, relative risk or difference of proportions and mean difference. However, very limited studies have considered a count variable as study effect of interest in performing REM. On the other hand, count data which is quite pervasive in many applications found limited research attempts in regression model building with Poisson or Negative Binomial distributions or its variants. However, these studies are less successful in explaining the variability across the predictors.

The present study has attempted Bayesian REM frame-work for count data with four plausible prior models including a zero-inflated distribution to account the zero counts. Also, the choice of hyper parameters involved in these models is appropriately discussed. The entire attempt is illustrated with a motivating study in transportation data analysis.

The paper is organized as follows: Sect. 2 describes the interesting aspects of the motivating bike sharing data set; Sect. 3 outlines Bayesian REM. Details of data analysis are discussed in Sects. 4 and 5 has the concluding remarks.

2 Motivation and Data Description

The main objective of this study is to understand the variability of response variable across different categorical variables. One of the major advantages of Bayesian analysis is treating a parameter as random variable which has been used in formulating a heterogeneity measure. This would be a pragmatic way of understanding the variability of a quantity of interest across different study characteristics. This aspect is best illustrated by the research interest of Capital Bike Sharing (CBS) problem.

Bike sharing system is a new age practice of traditional bike rentals. The entire process is automated to handle the membership and rental data [14]. The data set is a 2-year (2011–2012) usage detail of Capital Bike Sharing (CBS) at Washington, D.C., USA. This data set has additional information on weather conditions and weekday/holiday details. Such a real-world application with its impact on traffic, environment and health aspects provide a great interest and motivation for the present study.

The CBS data set has 17,379 instances of bike sharing details from January 1, 2011 to December 31, 2012. Seventeen variables and corresponding description can be obtained from the data source [4]. There are eight categorical variables (number of levels); season (4), year (2), month (12), hour (24), holiday (2), weekday (2), working day (2), and weather situation (4). Temperature, feeling temperature (both measured in Celsius), humidity, and wind speed are the four metric variables in normalized form. There are three count variables namely casual, registered users per the recorded index and the total count. Two variables are the index and date of each recording of bike rentals.

This study has considered two response count variables cnt and casual; cnt has no zeros but casual has a notable proportion of zeros. Beyond this characteristic these two variables are found to be important in the context of CBS. Five categorical variables have been identified (season, month, hr, weekday and weathersit) as stratification variables. Number of categories (k) in each of these five classifying variables is respectively 4, 12, 24, 7, and 4. An initial attempt to understand the two response variables across the five classifying variables have shown few interesting observations. A comparative summary is illustrated (Table 1) for the two response variables according to the four seasons; however, analysis includes all five classification variables.

From Table 1, it may be noted that there exists wide variation between upper quartile and maximum in both cases. Moreover, such difference can be seen across all the levels of the classification variables for both count response variables; for example, in the case of casual variable, across twelve months upper quartile varies

Table 1 Summary statistics of the two response count variables classified by the four seasons

Statistic	Casual				Cnt			
	1	2	3	4	1	2	3	4
Minimum	0	0	0	0	1	1	1	1
Lower quartile	1	7	10	4	23	46	68	46
Mid-quartile	5	27	36	14	76	165	199	156
Upper quartile	15	61	72	36	158	311	345	295
Maximum	367	361	350	362	801	957	977	967

Table 2 Cross-classification of quintile-based distribution of the two response count variables by the four weather situations

Weather situation	Casual					Cnt				
	[0, 3]	(3, 10]	(10, 26]	(26, 59]	(59, 367]	[1, 27]	(27, 98]	(98, 189]	(189, 321]	(321, 977]
1	2441	1821	2045	2454	2652	2183	2036	2220	2375	2599
2	1127	797	1023	917	680	929	935	990	931	759
3	588	346	242	145	98	425	462	261	158	113
4	2	1	0	0	0	1	1	1	0	0
	$p < 0.0001^{\#}$					$p < 0.0001^{\#}$				

$^{\#}$P-value from Chi-square test of independence

from 10 to 73 whereas maximum ranges from 156 to 367. This illustrate the variability of two count variables across a stratifying variable.

Also, the two response variables are categorised according to their quintiles and cross-classified with the five categorical variables. Chi-square test has been used to test the independence hypothesis. Table 2 illustrates this for both casual and cnt variables classified by four weather situations; last row is the p-value associated with Chi-square test of independence.

Table 2 and similar cross-classification by other four categorical variables have shown the variability of casual and cnt. Fourth season (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog) has a very low bike-sharing which may be apparent due to non-favourable weather conditions for bike-riding; nevertheless, other seasons have shown a notable variability in the count of casual (also, cnt) bike users. Similarly, it has been noted a larger variability of casual users in 24 levels of hr, one of the classification variables. Also, lower p-value rejects the hypothesis of independence in all the ten (5 categorical Vs 2 response variables) cases.

These characteristics encourage the research to quantify the heterogeneity of the response variables (casual and cnt) across five variables of interest. The study has fostered the random effects model for the count data applications. The underlying Poisson distribution for the data model is used in the first stage of this work. Subse-

quently, a normal model is used in the second stage to study the effect of heterogeneity and a summary for the mean count.

3 Materials and Methods

In a Bayesian model if $Y_i \sim \text{Poisson}(\lambda)$ where $\lambda > 0$ then in general, prior distribution for λ is a conjugate gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. Present study has included two more prior schemes using normal distribution and truncated normal distribution. Three Bayesian two-stage models are schematically presented.

Stage 1
Scheme I:
$Y_i \sim \text{Poisson}(\lambda)$
$\lambda \sim \text{Gamma}(\alpha, \beta)$
Scheme II:
$Y_i \sim \text{Poisson}(\lambda_1)$
$\lambda_1 \sim \exp(\lambda)$
$\lambda \sim \text{Normal}(d, v_1^2)$
Scheme III:
$Y_i \sim \text{Poisson}(\lambda)$
$\lambda \sim \text{Truncated Normal}(m, v_2^2, a, b)$

The next stage analysis is based on $log\lambda$ and its variance (estimated from data in Stage 1).

Stage 2
For each of k studies
$log(\lambda) \sim \text{Normal}(\mu_j, \sigma_j^2)$
$\mu_j \sim \text{Normal}(\mu, \tau^2); -\infty < \mu < +\infty, \tau^2 > 0.$
σ_j^2 is within the variance and its estimate will be used in the analysis, τ^2 is the between variance, a measure of heterogeneity among the k studies. Appropriate distributions are assumed for the subsequent level of parameters which are main quantities of interest in the study. Further to model zero cases (casual), this study includes Zero-inflated Poisson distribution. A Bayesian hierarchical model provides a straight forward way to construct a ZIP-Bernoulli model.
That is, scheme IV for the first stage will be
$Y \sim \text{Poisson}(\lambda Z) \quad \lambda > 0$
$Z \sim \text{Bernoulli}(p) \quad 0 \leq p \leq 1$
$\lambda \sim \text{Gamma}(\alpha, \beta) \quad \alpha, \beta > 0$
$p \sim \text{Beta}(a, b) \quad a, b > 0$

This model is a direct representation of understanding a ZIP model (Y) as a product of Poisson(X) and Bernoulli(Z) distributions; Precisely, $Y = XZ$ where $X \sim$ Poisson(λ); $Z \sim$ Bernoulli(p) with X and Z as independent variables.

All hyper parameters (α, β, d, v_1^2, m, v_2^2, a, b, γ, δ,a_1 and a_2) are appropriately chosen and the choices are presented in Sect. 4. The entire procedure is implemented in R using WinBugs for MCMC.

4 Data Analysis

The identified Bayesian REM with prior schemes is carried out for two response variables (cnt and casual). Choices of hyper priors are based on non-informative priors and three comparative choices are made in each set of analysis with cnt and casual. Hyper priors for gamma distribution in Scheme I is based on Jeffreys non-informative prior; 0.5 for scale parameter and larger positive value for shape parameter. For comparative purpose, the choice of shape parameter is chosen exactly 0.5 and a random value closer to 0.5. Hence shape parameter is randomly drawn from a uniform distribution in (0.49, 0.51). For scale parameter, the choice is from a uniform distribution in three intervals (5, 50), (0, 1), and (0, 50).

For Scheme II, the choice for mean parameter is N(0, 10^5); N(0, 10) and N(0, 10^3) and for variance it is gamma distribution with shape and scale parameters (3, 1), (0.5, 0.5), and (5, 5). In the case truncated normal distribution (Scheme III), the hyper parameters are listed as $m = 10$; $t = 0.001$; $a = 0$; $b = 50000$, $m = 100$; $t = 0.001$; $a = 0$; $b = 5000$, and m is uniformly distributed in (0, 100); $t = 0.001$; $a = 0$; $b = 1000$. Additional scheme for ZIP has the same set of values and for hyper priors for ZIP models will be 0.5 and 1 for symmetry and 0.5, 5 for asymmetry priors on beta parameters; whereas hyper priors in gamma distribution is retained with 3 and 1.

A measure of heterogeneity (H) is computed using MCMC simulated output as the posterior probability that the parameter between-variance exceeds 0.5, i.e. H = p($\tau^2 > 0.5$). The idea is to use the posterior probability and the exceedance of the between variance from any specific cut-off value. The choice of 0.5 is based on the notion that any value of τ^2 more than 0.5 is presumed to have a notable positive between variance. However, the model allows to choose any reasonable positive constant.

Tables 3 and 5 are the posterior summaries of overall count of response variables across five classification variables. Tables 4 and 6 are the corresponding summaries on heterogeneity measure. From Table 3, it can be observed that a difference (though less) prevails between the estimates of conjugate and other normal distributions; especially in the case of hour and weather situations. Interval estimates also differ in the case of weather situations, conjugate prior provide wider intervals than the other two. This is apparent when the scale parameter of a gamma distribution is allowed to vary while the choice is a non-informative (Jefferys) prior and the intervals differ largely with non-conjugate priors.

Table 3 Point and 95% interval estimates (LL Lower limit, UL Upper limit) for the response variable cnt, across the five classification variables; k refers the number of levels. Summaries are in the log-scale

Overall summary										
Study Variable	Posterior Summary	Prior 1			Prior 2			Prior 3		
		G	N	TN	G	N	TN	G	N	T
Season $k = 4$	Mean	5.149	5.157	5.157	5.157	5.157	5.157	5.157	5.157	5.157
	LL	5.149	5.157	5.157	5.157	5.157	5.157	5.157	5.157	5.157
	UL	5.149	5.157	5.157	5.157	5.157	5.157	5.157	5.157	5.157
Month $k = 12$	Mean	5.177	5.193	5.193	5.192	5.193	5.193	5.178	5.193	5.193
	LL	4.970	4.985	4.985	4.984	4.985	4.985	4.972	4.985	4.985
	UL	5.375	5.391	5.391	5.391	5.391	5.391	5.374	5.392	5.391
Hour $k = 24$	Mean	4.723	4.759	4.759	4.758	4.759	4.759	4.723	4.759	4.759
	LL	4.198	4.235	4.235	4.234	4.235	4.235	4.202	4.235	4.235
	UL	5.230	5.265	5.265	5.264	5.265	5.265	5.226	5.265	5.265
Weekday $k = 7$	Mean	5.234	5.243	5.243	5.243	5.243	5.243	5.232	5.243	5.243
	LL	5.170	5.181	5.181	5.181	5.181	5.181	5.170	5.181	5.181
	UL	5.295	5.303	5.302	5.302	5.303	5.302	5.292	5.303	5.302
Weather $k = 4$	Mean	3.699	4.809	4.803	4.775	4.808	4.813	3.620	4.808	4.806
	LL	0.783	4.022	4.001	3.914	4.016	4.032	0.599	4.016	4.011
	UL	5.640	5.467	5.469	5.485	5.467	5.465	5.624	5.467	5.468

G = Gamma distribution; N = Normal distribution; TN = Truncated normal distribution

A similar effect can be observed for the overall estimate of casual (Table 5). Especially in the case of fifth classification (weather situation) difference between gamma and normal priors are quite apparent. Prior II behaves differently compared to other two choices of hyper parameter. Another interesting point is the difference between ZIP and other three prior schemes as observed in all the point and interval estimates of casual. However, ZIP is quite sensitive to the choice of its hyper parameters.

As far the between variance is considered for cnt (Table 4) point and interval estimates are quite similar whenever H is small as can be observed from all the classification except third and fifth. In $k = 24$ for third variable, the sensitivity is not so high but not in the case of fifth variable. Gamma priors (with hyper prior choices 1 and 3) show appreciable difference in the estimator of τ^2 (7.174 with H = 0.999) with wider intervals. Surprisingly the case differs when hyper prior is changed to U(0, 1) for the scale parameter. Similar observation can be made for casual from Table 6. Gamma prior in fifth classification is still very much distinctive. Also, it can be noted that ZIP estimates of H are quite different to that of other priors. However,

Table 4 Point and 95% interval estimates (LL Lower limit, UL Upper limit) for the between variance of the response variable cnt, across the five classification variables; k refers the number of levels

Between variance										
Study Variable	Posterior Summary	Prior 1			Prior 2			Prior 3		
		G	N	TN	G	N	TN	G	N	T
Season $k = 4$	Mean	0.461	0.456	0.456	0.456	0.456	0.456	0.455	0.456	0.456
	LL	0.040	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039
	UL	1.563	1.545	1.545	1.544	1.545	1.545	1.541	1.545	1.545
	H = p($\tau^2 > 0.5$)	0.122	0.121	0.121	0.121	0.121	0.121	0.121	0.121	0.121
Month $k = 12$	Mean	0.124	0.126	0.126	0.126	0.126	0.126	0.123	0.126	0.126
	LL	0.052	0.052	0.052	0.052	0.052	0.052	0.051	0.052	0.052
	UL	0.291	0.294	0.294	0.294	0.294	0.294	0.287	0.294	0.294
	H = p($\tau^2 > 0.5$)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Hour $k = 24$	Mean	1.591	1.587	1.587	1.588	1.587	1.587	1.568	1.587	1.587
	LL	0.866	0.864	0.864	0.864	0.864	0.864	0.854	0.864	0.864
	UL	2.823	2.815	2.815	2.817	2.816	2.815	2.782	2.816	2.815
	H = p($\tau^2 > 0.5$)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Weekday $k = 7$	Mean	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
	LL	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	UL	0.024	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023
	H = p($\tau^2 > 0.5$)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weather $k = 4$	Mean	7.174	0.773	0.796	0.888	0.779	0.762	7.802	0.779	0.785
	LL	0.876	0.068	0.070	0.080	0.068	0.067	0.959	0.068	0.069
	UL	35.962	2.842	2.968	3.378	2.884	2.785	38.871	2.884	2.921
	H = p($\tau^2 > 0.5$)	0.999	0.259	0.270	0.312	0.261	0.255	1.000	0.261	0.264

G = Gamma distribution; N = Normal distribution; TN = Truncated normal distribution

there is a direct relationship with lower the value of H lower is the difference in the estimator. This is evident when H = 0.059 for fourth classification and when H varies from 0.931 to 1 in fifth classification.

5 Conclusions

The abundant volume of data availability in various fields and computing power has enhanced the application of count data models with more relevant statistical techniques. This motivates to study the heterogeneity measure in REM using three prior schemes for no zero case and an additional prior for zero case counts. The study

Table 5 Point and 95% interval estimates (LL Lower limit, UL Upper limit) for the response variable casual, across the five classification variables; k refers the number of levels

Over all measure casual													
Study Variable	Posterior Summary	Prior 1				Prior 2				Prior 3			
		G	N	TN	ZIP	G	N	TN	ZIP	G	N	T	ZIP
Season $k = 4$	Mean	3.376	3.381	3.382	3.494	3.382	3.382	3.382	3.494	3.376	3.382	3.382	3.493
	LL	2.396	2.401	2.402	2.639	2.402	2.402	2.402	2.639	2.399	2.402	2.402	2.639
	UL	4.199	4.204	4.204	4.221	4.204	4.204	4.204	4.221	4.198	4.204	4.204	4.221
Month $k = 12$	Mean	3.377	3.395	3.395	3.496	3.395	3.395	3.395	3.496	3.379	3.395	3.395	3.494
	LL	2.952	2.970	2.970	3.115	2.969	2.970	2.970	3.115	2.952	2.970	2.970	3.114
	UL	3.781	3.799	3.799	3.858	3.798	3.799	3.799	3.858	3.785	3.799	3.799	3.857
Hour $k = 24$	Mean	2.941	2.975	2.975	3.079	2.975	2.975	2.976	3.079	2.936	2.975	2.975	3.076
	LL	2.365	2.401	2.401	2.564	2.400	2.401	2.401	2.564	2.360	2.401	2.401	2.561
	UL	3.501	3.533	3.533	3.579	3.532	3.533	3.533	3.579	3.495	3.533	3.533	3.576
Weekday $k = 7$	Mean	3.468	3.481	3.481	3.578	3.481	3.481	3.481	3.578	3.471	3.481	3.481	3.577
	LL	3.089	3.100	3.100	3.224	3.100	3.100	3.100	3.224	3.088	3.100	3.100	3.223
	UL	3.822	3.835	3.835	3.908	3.835	3.835	3.835	3.908	3.827	3.835	3.835	3.907
Weather $k = 4$	Mean	1.787	2.527	2.568	2.594	2.538	2.518	2.579	2.594	1.574	2.524	2.571	2.434
	LL	−1.18	0.573	0.690	0.541	0.610	0.542	0.718	0.539	−1.63	0.555	0.699	0.142
	UL	4.143	4.021	4.007	4.129	4.019	4.022	4.004	4.129	4.164	4.021	4.006	4.176

G = Gamma distribution; N = Normal distribution; TN = Truncated normal distribution

Table 6 Point and 95% interval estimates (LL Lower limit, UL Upper limit) for the between variance of the response variable casual, across the five classification variables; k refers the number of levels

Between variance casual													
Study Variable	Posterior Summary	Prior 1				Prior 2				Prior 3			
		G	N	TN	ZIP	G	N	TN	ZIP	G	N	T	ZIP
Season $k = 4$	Mean	0.977	0.977	0.975	0.782	0.975	0.975	0.975	0.782	0.973	0.975	0.975	0.781
	LL	0.110	0.110	0.110	0.086	0.110	0.110	0.110	0.086	0.110	0.110	0.110	0.086
	UL	4.264	4.261	4.254	3.285	4.254	4.255	4.255	3.285	4.244	4.255	4.255	3.282
	H = p($\tau^2 > 0.5$)	0.425	0.425	0.424	0.324	0.424	0.424	0.424	0.324	0.423	0.424	0.424	0.324
Month $k = 12$	Mean	0.523	0.523	0.523	0.420	0.523	0.523	0.523	0.420	0.529	0.523	0.523	0.420
	LL	0.217	0.216	0.217	0.174	0.217	0.217	0.217	0.174	0.219	0.217	0.216	0.174
	UL	1.221	1.219	1.220	0.980	1.220	1.220	1.220	0.980	1.233	1.220	1.219	0.980
	H = p($\tau^2 > 0.5$)	0.419	0.418	0.418	0.249	0.418	0.418	0.418	0.249	0.429	0.418	0.417	0.250
Hour $k = 24$	Mean	1.949	1.933	1.933	1.549	1.933	1.933	1.933	1.549	1.943	1.933	1.933	1.549
	LL	1.063	1.055	1.055	0.844	1.055	1.055	1.055	0.844	1.060	1.055	1.055	0.844
	UL	3.460	3.433	3.432	2.753	3.432	3.433	3.432	2.752	3.450	3.434	3.433	2.753
	H = p($\tau^2 > 0.5$)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Weekday $k = 7$	Mean	0.254	0.255	0.255	0.221	0.255	0.255	0.255	0.221	0.259	0.255	0.255	0.221
	LL	0.070	0.070	0.070	0.061	0.070	0.070	0.070	0.061	0.071	0.070	0.070	0.061
	UL	0.852	0.856	0.856	0.737	0.856	0.856	0.856	0.737	0.867	0.856	0.856	0.737
	H = p($\tau^2 > 0.5$)	0.081	0.082	0.082	0.059	0.082	0.082	0.082	0.059	0.084	0.082	0.082	0.059
Weather $k = 4$	Mean	10.181	3.446	3.178	3.708	3.372	3.508	3.104	3.711	12.858	3.470	3.155	4.892
	LL	1.423	0.383	0.349	0.435	0.374	0.390	0.340	0.436	1.870	0.385	0.347	0.608
	UL	48.086	16.631	15.242	18.064	16.301	16.864	14.941	18.093	60.139	16.731	15.172	23.430
	H = p($\tau^2 > 0.5$)	1.000	0.946	0.931	0.961	0.943	0.949	0.926	0.961	1.000	0.947	0.929	0.991

G = Gamma distribution; N = Normal distribution; TN = Truncated normal distribution

has identified Poisson distributed variables and the behaviour of between variance parameter when k (>1) studies are combined.

The results have shown the behaviour of Bayesian hierarchical model with respect to priors on Poisson parameter. The sensitive nature of estimates for overall mean is visible in non-zero cases under gamma and normal priors for transformed parameters. Number of strata is also observed as a possible factor for such behaviour in the estimates. In a similar note, the effect of heterogeneity relatively depends on the choice of priors under gamma or normal distributions. Under all the three priors the estimates are notably different for ZIP with prior on proportion of zeros when compared to other models. This study has involved only non-informative priors on hyper parameters in the Bayesian model. Nevertheless, this may provide a methodology for modelling count data with additional covariates and/or including more plausible priors.

Acknowledgements Funding for this project was provided by 2017 funds of the University of Naples - L'Orientale (I).

References

1. Caird, J.K., Johnston, K.A., Willness, C.R., Asbridge, M., Steel, P.: A meta-analysis of the effects of texting on driving. Accid. Anal. Prev. **71**, 311–318 (2014)
2. DerSimonian, R., Kacker, R.: Random-effects model for meta-analysis of clinical trials: an update. Contemp. Clin. Trials **28**(2), 105–114 (2007)
3. Engels, E.A., Schmid, C.H., Terrin, N., Olkin, I., Lau, J.: Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat. Med. **19**(13), 1707–1728 (2000)
4. Fanaee-T, H., Gama, J.: Event labeling combining ensemble detectors and background knowledge. Prog. Artif. Intell. **2**(2–3), 113–127 (2014). https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset
5. Higgins, J.P., Spiegelhalter, D.J.: Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. Int. J. Epidemiol. **31**(1), 96–104 (2002)
6. Higgins, J., Thompson, S.G.: Quantifying heterogeneity in a meta-analysis. Stat. Med. **21**(11), 1539–1558 (2002)
7. Hillebrand, H.: Meta-analysis in ecology. Encyclopedia of Life Sciences (ELS). Wiley, Chichester (2008)
8. Hunter, J.P., Saratzis, A., Sutton, A.J., Boucher, R.H., Sayers, R.D., Bown, M.J.: In meta-analyses of proportion studies, funnel plots were found to be an inaccurate method of assessing publication bias. J. Clin. Epidemiol. **67**(8), 897–903 (2014)
9. Jackson, D.: Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. Res. Synth. Methods **4**(3), 220–229 (2013)
10. Langan, D., Higgins, J., Simmonds, M.: An empirical comparison of heterogeneity variance estimators in 12 894 meta-analyses. Res. Synth. Methods **6**(2), 195–205 (2015)
11. Langan, D., Higgins, J., Simmonds, M.: Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. Res. Synth. Methods **8**(2), 181–198 (2017)
12. Leonard, T., Duffy, J.C.: A Bayesian fixed effects analysis of the Mantel-Haenszel model applied to meta-analysis. Stat. Med. **21**(16), 2295–2312 (2002)

13. Mannering, F.L., Shankar, V., Bhat, C.R.: Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Acc. Res. **11**, 1–16 (2016)
14. Mátrai, T., Tóth, J.: Comparative assessment of public bike sharing systems. Transp. Res. Procedia **14**, 2344–2351 (2016)
15. Nazarzadeh, M., Bidel, Z.: Meta-analysis of sleep duration and obesity in children: fixed effect model or random effect model? J. Paediatr. Child Health **53**(9), 923–924 (2017)
16. R Development CORE TEAM: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org (2010)
17. Rücker, G., Schwarzer, G., Carpenter, J., Olkin, I.: Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. Stat. Med. **28**(5), 721–738 (2009)
18. Rukhin, A.L.: Estimating heterogeneity variance in meta-analysis. J. R. Stat. Soc. Ser. B (Statistical Methodology) **75**(3), 451–469 (2013)
19. Sidik, K., Jonkman, J.N.: Simple heterogeneity variance estimation for meta-analysis. J. R. Stat. Soc. Ser. C (Applied Statistics) **54**(2), 367–384 (2005)
20. Sidik, K., Jonkman, J.N.: A comparison of heterogeneity variance estimators in combining results of studies. Stat. Med. **26**(9), 1964–1981 (2007)
21. Smith, T.C., Spiegelhalter, D.J., Thomas, A.: Bayesian approaches to random-effects meta-analysis: a comparative study. Stat. Med. **14**(24), 2685–2699 (1995)
22. Subbiah, M., Rajeswaran, V.: A random effect model for the evolution of international cricket test matches evidenced from 1870 to 2016. Stat. Appl. **14**(2) (2016)
23. Subbiah, M., Srinivasan, M.R.: Classification of 22 sparse data sets with zero cells. Stat. Probab. Lett. **78**(18), 3212–3215 (2008)
24. Thorlund, K., Wetterslev, J., Awad, T., Thabane, L., Gluud, C.: Comparison of statistical inferences from the DerSimonian-Laird and alternative random-effects model meta-analyses-an empirical assessment of 920 Cochrane primary outcome meta-analyses. Res. Synth. Methods **2**(4), 238–253 (2011)
25. Tian, L., Cai, T., Pfeffer, M.A., Piankov, N., Cremieux, P.Y., Wei, L.J.: Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2×2 tables with all available data but without artificial continuity correction. Biostatistics **10**(2), 275–281 (2008)
26. Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Salanti, G.: Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res. Synth. Methods **7**(1), 55–79 (2016)
27. Viechtbauer, W.: Bias and efficiency of meta-analytic variance estimators in the random-effects model. J. Educ. Behav. Stat. **30**(3), 261–293 (2005)
28. Viechtbauer, W.: Confidence intervals for the amount of heterogeneity in meta-analysis. Stat. Med. **26**(1), 37–52 (2007)
29. Viechtbauer, W.: Conducting meta-analyses in R with the meta for package. J. Stat. Softw. **36**(3), 1–48 (2010)
30. Vienneau, D., Schindler, C., Perez, L., Probst-Hensch, N., Röösli, M.: The relationship between transportation noise exposure and ischemic heart disease: a meta-analysis. Environ. Res. **138**, 372–380 (2015)
31. Warn, D.E., Thompson, S.G., Spiegelhalter, D.J.: Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. Stat. Med. **21**(11), 1601–1623 (2002)
32. Zwetsloot, P.P., Van Der Naald, M., Sena, E.S., Howells, D.W., IntHout, J., De Groot, J.A., Wever, K.E.: Standardized mean differences cause funnel plot distortion in publication bias assessments. ELife **6** (2017)

Sampling Techniques for Big Data Exploration

Sampling and Modelling Issues Using Big Data in Now-Casting

Maria Simona Andreano, Roberto Benedetti, Federica Piersimoni, Paolo Postiglione and Giovanni Savio

Abstract The use of Big Data and, more specifically, Google Trends data in now- and forecasting, has become common practice nowadays, even by Institutes and Organizations producing official statistics worldwide. However, the use of Big Data has many neglected implications in terms of model estimation, testing and forecasting, with a significant impact on final results and their interpretation. Using a MIDAS model with Google Trends covariates, we analyse sampling error issues and time-domain effects triggered by these digital economy new data sources.

Keywords Google trend · Mixed frequency · Representativeness

M. Simona Andreano (✉)
Faculty of Economy, Universitas Mercatorum, Piazza Mattei 10, 00186 Rome, Italy
e-mail: s.andreano@unimercatorum.it

R. Benedetti · P. Postiglione
Department of Economic Studies (DEC), "G. d'Annunzio" University,
Viale Pindaro 42, 65127 Pescara, Italy
e-mail: benedett@unich.it

P. Postiglione
e-mail: postigli@unich.it

F. Piersimoni
Istat, Directorate for Methodology and Statistical Process Design,
Via Cesare Balbo 16, 00184 Rome, Italy
e-mail: piersimo@istat.it

G. Savio
UN-ECLAC, United Nations Economic Commission for Latin America and the Caribbean,
Av. Dag Hammarskjöld 3477, Vitacura, Santiago de Chile, Chile
e-mail: giovanni.savio@un.org

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_14

1 Introduction

The emergence of Big Data, and their capacity to help in now-casting, forecasting, disaggregating, and filling in gaps of conventional and official statistics data sources, is now history [28]. Nowadays, data are automatically and continuously generated in digital form in many different respects, and from different data sources.

Although Big Data represent a data source of great potential interest for official statistics, extracting relevant and reliable information from them for now-casting and forecasting is not an easy task, and many challenges and open questions still remain, including:

- the representativeness and selectivity of Internet data sources;
- the synthesis of the information contained in the data;
- the identification of sampling errors in Internet data;
- the estimation methods and modelling for disaggregated (in space and/or time) data;
- the now- or forecasting model evaluation.

While there is no an agreed–upon definition of Big Data, literature often refers to the characteristics that data-sets should have, namely the three Vs: volume, velocity, and variety [20]. Volume denotes that Big Data are massive datasets, with a large number of records stored. Velocity refers to the speed between the occurrence of an event and the time it is stored. Variety denotes the high heterogeneity of data sources and formats, which is closely related to an increase in complexity structure.

A great part of literature on Big Data largely addresses IT implications from their use, focusing on software, format, and dimensionality challenges. Purely statistics implications from their use have often been undervalued, and important statistics questions such as representativeness, coverage and sampling errors, have not been addressed by literature. As clearly noticed by [6], unlike traditional data collection mechanisms, such as those used for regular business surveys and business registers, Big Data are generated by processes not primarily aimed at data collection. This implies significant differences in the statistics characteristics of the data, as underlined in Table 1 [6].

Table 1 Characteristics of data sources

Data source	Sample survey	Register	Big Data
Volume	Small	Large	Big
Velocity	Slow	Slow	Fast
Variety	Low	Low	High
Records	Units	Units	Units/events
Unit selection	Probabilistic	Administrative	Non probabilistic
Reference population	Total	Total	Partial

Table 1, which includes different coverage of the data source with respect to the population of interest, shows that one of the main differences between registers and Big Data is that the former often has nearly complete coverage of the population, while the latter generally does not.

Another important distinction highlighted in the previous table, relates to the errors, as for Big Data there is not yet a framework to assess errors and quality aspects.

This paper will try to fill in this gap, by dealing with some of these issues. The paper mainly focuses on Google Trends data and their use in now- and forecasting. Buono et al. [7] present a detailed review of the various types of Big Data that can be useful in macroeconomic now-casting, providing many real applications.

In the next section we introduce Google Trends data and highlight the main sample and modelling problems from their use in empirical applications. Section 3 focuses on the representativeness and selectivity problems of internet data, and suggests also a procedure in the time domain. Final remarks conclude the paper.

2 Google Trends Data as Covariates in Now-Casting

The use of Google Trends data started with the papers of [10, 11], who showed their relevance in predicting consumer behaviour and initial unemployment claims for the US. From then on, forecasters have been looking at Google search data as an information source that could improve their predictions. Apart from being free, the speed at which these data are available makes them very attractive for studying economic dynamics. The volume of queries made by users about products via the search engine provided by Google could reflect the potential volume of sales of these products. These data could therefore be considered as indicators or proxies of consumer's purchase intentions.

Most of the applications on Google Trends data concern unemployment rate predictions [2, 12, 15, 26]. Schmidt and Vosen [24] used Google Trends data to predict US private consumption and estimate a monthly indicator for private consumption in Germany. Bangwayo-Skeete and Skeete [3] introduced a new indicator for tourism demand forecasting from Google Trends search query time series data. Bontempi et al. [5] proposed a new uncertainty indicator based on internet search to anticipate changes in economic cycle. For a more extended review see [7].

The Google Trends website provides access to a weekly average query index:

$$S_{\omega_i r} = \frac{1}{7} \sum_{d \in \omega_i} \frac{V_{d,r}}{T_{d,r}}, \tag{1}$$

which indicates how often in day d a specific keyword is searched (V) relative to the total search volume (T), over a certain week ω_i in a given geographical region (r).

The index is then normalized with respect to the maximum observed over the whole analyzed period $[0, t]$:

$$GI = \frac{100}{\max\limits_{\omega_i \in [0,t]} \left(S_{\omega_i r}\right)} S_{\omega_i r}, \tag{2}$$

Google time series only go back to 2004.

The Google Trends data are instantaneous and released as soon as possible, therefore can be helpful not only in predicting future, but also in actual official statistics, which are usually defined at a monthly (or lower frequency) time intervals. Predicting the present, in the sense described above, is a form of contemporaneous forecasting or now-casting. However, typical time series regression models use data sampled at the same frequency. The idea to build regression models that combine data with different sampling frequencies is relatively new. Indeed, Mixed Data Sampling (MIDAS) models specify conditional expectations as a distributed lag of regressors at some higher sampling frequencies, and the lowest frequency series is regressed on the higher frequency one [17].

The basic MIDAS model with a single explanatory variable and h-step ahead forecasting, with $h = h_m/m$, is given by:

$$y_{t+mh} = y_{t_m+h_m} = \beta_0 + \beta_1 B\left(L^{\frac{1}{m}}; \theta\right) x^{(m)}_{t_m-\gamma} + \epsilon^{(m)}_{t_m+h_m}, \tag{3}$$

where $B\left(L^{\frac{1}{m}}; \theta\right) = \sum_{k=0}^{K} B\left(k; \theta\right) L_m^k$ denotes a weighting function of the L fractional lag operator, t indexes the basic time unit, m is the frequency mixture and γ is the number of values of the indicators that are available earlier than the lower-frequency variable to be estimated. There are several possible finite and infinite polynomials $B\left(k; \theta\right)$ specification. Between t and $t-1$, the higher-frequency variable is observed m times and after one week it is possible to forecast y_t one month ahead.

In recent literature we can find many attempts to use Google Trends data to now-cast, i.e. [8, 9, 14, 23].

In a recent paper [1] used weekly Google Trends search in car sales to forecast through a MIDAS model, 6-month ahead, car registrations in Italy. However, before using Google Trends data for now-casting, some previous check of the data should be made. In fact, Google currently calculates the GI index based on a random sample whose design is completely unknown and this will result in a sampling error. Moreover, [13] point out that the indices can vary depending on the IP address and it is unknown how Google applies its algorithm to millions of queries to form the indices in Google Trends. Therefore, before including a Google Trends variable as a covariate in the model, it is necessary to evaluate the magnitude of the sampling error.

In the present paper, we take up the Google Trends series of car sales used in [1], and perform some in-depth analysis in order to measure the magnitude of the sampling error and assess its effect on forecasting performance. We downloaded 30

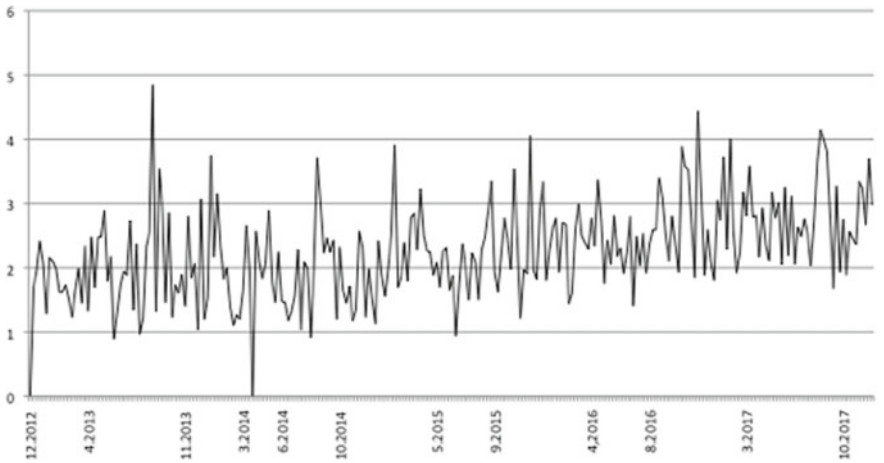


Fig. 1 Car sales Google Trends series: standard deviation

series from different IP address, with different Gmail accounts, but over the same time interval, geographical area and search query. Although the correlation between these series is high, ranging from 0.75 to 0.99, the series are different. In Fig. 1 we show the standard deviation of the 30 series.

Looking at the graph, we see that the standard deviation does not remain constant over the analyzed time interval, confirming the idea that the between-day variations is caused by the dynamic algorithm applied by Google [19].

The data downloaded over a shorter time vary much less than those downloaded over a longer period. Therefore, if the sampling error is large, one should use the sample mean of the multiple downloads into the forecasting model, instead of a single download.

A second, and maybe more relevant, problem with Google Trends data, refers to the updating of the series. In Fig. 2 we report the same Google Trends search series, downloaded over an iterated time interval, augmented one week each time.

The graph clearly shows that the iterated downloaded series changes each time. The difference can be very high, especially if a new maximum is reached in the updated interval. The correlation between two consecutive updated series can even go down at 0.3.

If the MIDAS model is estimated on a given Google Trends data and new internet downloaded observations are used to now- and forecasting, particular attention should be paid to the coherence of the model over the new data. If the revision of the new series is high, the estimated model cannot be applied for now-casting using the new data, because the model is not robust to different vintages. This problem will be more and more evident with increasing h–ahead time forecast intervals.

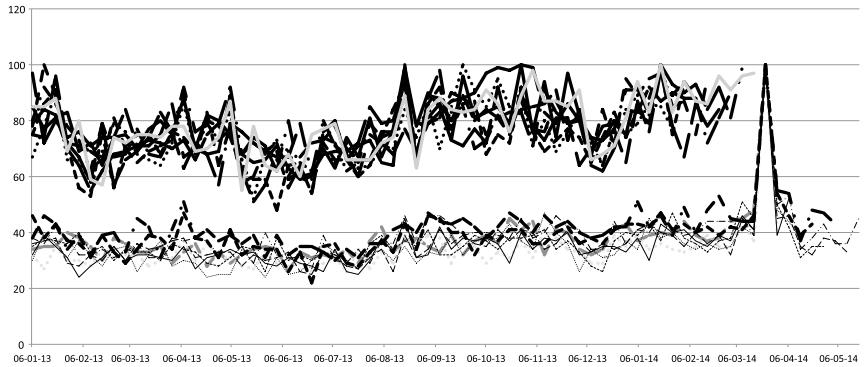


Fig. 2 Iterated car sales Google Trends data: weekly updates

From the previous two exercises, it is clear that the sampling error may vary, and the forecast performance may be affected by sampling error. Given the sampling approach of Google, downloading the series from multiple IP addresses over a short time period, and getting the average, seems a preferable solution.

3 Representativeness and Selectivity

One of the main open questions regarding the use of Big Data in official statistics is representativeness and selectivity of the reference population. Standard inference techniques are based on sample surveys, with estimation theory able to quantify the error of unknown population parameters. However, Big Data are not created by statisticians or for statistical purposes. They represent a self-selected (non probabilistic) sample, with generating mechanisms often unknown. Therefore, there is no guarantee that the data are representative, unless they cover the full population of interest, as it is the case for satellite sensing [18].

Big Data can be used in several ways to improve the information of standard statistics and, depending on the nature of application, different issues arise. In this paper, we focus on the use of Big Data (more specifically Google Trends data) as auxiliary variable X in a time series model (MIDAS) to obtain more timely and better forecasts of a target variable Y. Therefore, our reference scheme is that reported in Fig. 3, where the target variable Y_U is observed over a population U, and the covariate X_B comes from a Big Data population U_B, $U_B \subset U$. However, standard inference could be applied, providing unbiased estimations only if representativeness of the sub-population is satisfied. A subset of a finite population is said to be representative of that population with respect to a target variable, if the distribution of that variable within the subset is the same as in the population [6]. Unfortunately, there

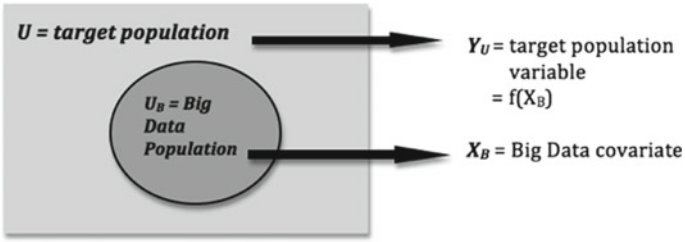


Fig. 3 Representativeness of internet data

are few references on this topic, analysing the issue from a theoretical and empirical perspective [4, 6, 22].

Following [27], we denote with I the sampling process, explaining the selection of U_B from U. In most situations with Big Data cases, I is unknown. Moreover, R denotes the censoring (missing data) process, which makes the observation of the covariate X incomplete and limited only to X_B, on the sub-population U_B. The problem is to calculate the posterior distribution:

$$f\left(Y_U \mid X_B, I, R\right) \equiv \left[Y_U \mid X_B, I, R\right], \tag{4}$$

If ignorability conditions for sampling and censoring are satisfied (see [21], p. 13), then:

$$\left[Y_U \mid X_B, I, R\right] \equiv \left[Y_U \mid X_B\right], \tag{5}$$

If we assume that the probability density function of Y_U is known with parameter φ, the estimation of Y_U through the covariate X_B can then be performed by disregarding the sampling and censoring processes:

$$\left[\varphi \mid X_B, I, R\right] \equiv \left[\varphi \mid X_B\right], \tag{6}$$

Unfortunately, when dealing with Google Trends data, the ignorability condition is not ensured. Google Trends data are a self-selected sample, where some population subgroups may be under-represented, causing biased estimates [25].

In literature, we can find different solutions to overcome this problem. A weaker condition of ignorability can be obtained, if Big Data can be matched to other variables Z_B, for which ignorability holds. This is the case when Big Data contain records at unit level that can be identified through, for example, registers or sample surveys [4].

Marchetti et al. [21] overcome the problem of unit level matching by applying area-level models with area-level auxiliary variables, and aggregate Big Data in the domains/areas of interest.

These solutions are not appropriate in our MIDAS model with Google Trends covariate for different reasons. First, Google Trends data are not observed and

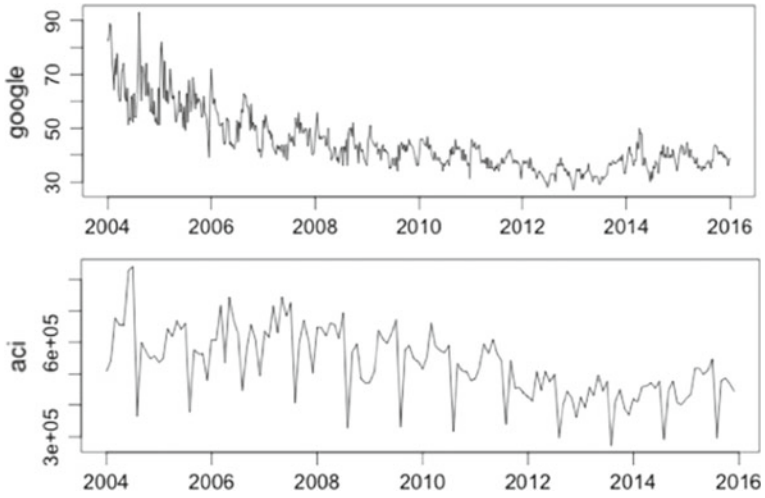


Fig. 4 Google trends search and car registration, Italy (2004–2016)

measured at unit level, essentially for privacy and confidentiality reasons. Second, the aim of the mixed frequency models is to explain a target variable with respect to covariates observed at a finer time frequency level, avoiding aggregating data, contrary to the case of area-level models. Finally, Google Trends data are classical examples of Big Data without coverage of all population, and therefore, with self-selection problem. Therefore, a different solution is needed.

A solution is to derive background characteristics from within the Big Data variable X_B that match those of the target variable Y_U. In the time domain, as in our case, one should apply the cointegration analysis on the observed series. This avoids the risk of discovering spurious or false correlations between two—or more—time series and allows verifying if the time series share common trends.

Figure 4 shows the two time series used in our MIDAS model: monthly cars registrations (ACI) and weekly Google Trends search in car sales.

Although the MIDAS model is estimated on stationary time series, a preliminarily long-run analysis between the variables is needed to solve the representativeness issue of Google Trends data. However, the application of cointegration analysis in a mixed frequency context is not immediate. The effects of aggregation on the size of standard cointegration tests may be severe. Different procedures are proposed by literature to perform a cointegration analysis on time series observed at different frequencies [16].

The presence of common trends between target and explanatory variables is a necessary condition for representativeness, however more analysis should be made to verify its sufficiency.

Finally, few words should be spent regarding the use of remote sensing and satellite data as additional data source. Digital technologies provide more and more potential new sources of data, offering more information to official statistician. Satellite sens-

ing data are available once every fortnight, with a clearly advantage over annual or sub-annual official data. These data are perfectly identifiable (through pixel longitudes and latitudes), with Big Data population U_B overlapping the target population U. Therefore, the following holds:

$$\left[X_B\right] \equiv \left[X_U\right], \tag{7}$$

The requirements of ignorability condition (6) are fully satisfied. Attention should only be paid when there are missing data, and the process R needs to be carefully checked. Satellite images and geographic information systems (GIS) are widely used in agriculture, urbanization and population studies [22]. Henderson et al. [18] applied satellite National Oceanic and Atmospheric Administration night-lights data to improve on official income growth measures and to obtain small area estimation of GDP growth.

4 Concluding Remarks

The use of Big Data as additional source in now-casting is a great opportunity for official statisticians to obtain more efficient, effective and timely information, but it also defines a new paradigm for data and models. We overview in the paper several unique features brought by Big Data, and more in detail by online search data as Google Trends. We focused, more specifically, on their use in nowcasting as covariates in mixed frequency models, highlighting a set of key challenges that, at present, hinder and restrict the accuracy and effectiveness of forecasting with Big Data. Our empirical results showed that a preliminary evaluation of the sampling errors and magnitude of the standard deviation is needed, before using such data for forecasting, because these can significantly affect the estimation model output. Moreover, the online downloaded series should be carefully updated because revisions may be high, compromising the robustness of the estimated model.

Moreover, representativeness and selectivity of Big Data remain important issues that are worth considering in future research if one wants to reduce the risk of misleading forecasting results. In the present paper, we suggest to apply the cointegration analysis between target and explanatory variables of mixed frequency model, as a tool to verify representativeness.

Although Big Data represents a data source of great potential interest for official statistics, many challenges still remain, and more in-depth analysis is needed to explore their statistical quality and characteristics.

References

1. Andreano, M.S., Benedetti, R., Postiglione, P.: Forecasting with mixed data sampling models (MIDAS) and Google trends data — the case of car sales in Italy. In: Proceedings of the 48th SIS Scientific Meeting, Salerno (2016)
2. Askitas, N., Zimmermann, K.: Google econometrics and unemployment forecasting. Appl. Econ. Quart. **55**, 107–120 (2009)
3. Bangwayo-Skeete, P.F., Skeete, R.W.: Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. Tour. Manag. **46**, 454–464 (2015)
4. Barcaroli, G., Golini, N., Nurra, A., Righi, R., Piersimoni, F., Salamone, S., Scarnò, M.: Joint use of sampling data and big data: the experience with the Istat survey on the use of ICT by enterprises. In: ITACOSM Bologna (2017)
5. Bontempi, M.E., Golinelli, R., Squadrani, M.: A new index of uncertainty based on internet search: a friend or a foe of the indicators? Working Papers 1062. Dipartimento Scienze Economiche, Università di Bologna (2016)
6. Buelens, B., Burger, J., Daas, P., Puts, M., van den Brakel, J.: Selectivity of big data. Discussion Paper 11, Statistics Netherlands (2014)
7. Buono, D., Mazzi, G.L., Kapetanios, G., Marcellino, M., Papailias, F.: Big data types for macroeconomic nowcasting. EURONA Eurostat Rev. Natl. Acc. Macroecon. Ind. **1** (2017)
8. Carriere-Swallow, Y., Labbe, F.: Nowcasting with Google trends in an emerging market. J. Forecast. **32**, 289–298 (2013)
9. Chamberlain, G.: Googling the present. Econ. Labour Market Rev. **4**, 59–95 (2010). Office for National Statistics
10. Choi, H., Varian, H.: Predicting initial claims for unemployment benefits. http://research.google.com/archive/papers/initialclaimsUS.pdf (2009)
11. Choi, H., Varian, H.: Predicting the present with Google trends. Econ. Rec. **88**, 2–9 (2012)
12. D'Amuri, F.: Predicting unemployment in short samples with internet job search query data. MPRA Working Paper 18403 (2009)
13. D'Amuri, F., Marcucci, J.: The predictive power of Google search in forecasting unemployment. Int. J. Forecast. **33**, 801–816 (2015)
14. Ferreira, P.: Improving prediction of unemployment statistics with Google trends: part 2. Eurostat website. https://ec.europa.eu/eurostat/ (2015)
15. Fondeur, Y., Karamè, F.: Can Google data help predict French youth unemployment? Econ. Model. **30**, 117–125 (2013)
16. Ghysels, E., Miller, J.I.: Testing for cointegration with temporally aggregated and mixed-frequency time series. J. Time Ser. Anal. **36**, 797–816 (2015)
17. Ghysels, E., Santa-Clara, P., Valkanov, R.: MIDAS regressions: further results and new directions. Econom. Rev. **26**, 53–90 (2006)
18. Henderson, J.V., Storeygard, A., Weil, D.N.: Measuring economic growth from outer space. Am. Econ. Rev. **102**, 994–1028 (2012)
19. Li, X.: Nowcasting with big data: is Google useful in presence of other information? London Business School, Unpublished manuscript (2016)
20. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute (2011)
21. Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., Gabrielli, L.: Small area model-based estimators using big data sources. J. Off. Stat. **31**, 263–281 (2015)
22. Pratesi, M., Petrucci, A.: Spatial disaggregation and small-area estimation methods for agricultural surveys: solutions and perspectives. Technical Report Series GO-07-2015 (2015)
23. Ross, A.: Nowcasting with Google trends: a keyword selection methods. Fraser Allander Econ. Comment. **37**, 54–64 (2013)
24. Schmidt, T., Vosen, S.: Forecasting private consumption: survey-based indicators vs Google trends. J. Forecast. **30**, 565–578 (2011)

25. Smith, T.M.F.: On the validity of inferences from non-random samples. J. R. Stat. Soc. A **146**, 394–403 (1983)
26. Suhoy, T.: Query indices and a 2008 downturn: Israeli data. Bank of Israel Discussion Paper 06 (2009)
27. Tam, S., Clarke, F.: Big data, statistical inference and official statistics. Research Paper, Australian Bureau of Statistics 1351.0.55.054 (2015)
28. United Nations: Big data and modernization of statistical systems. Statistical Commission, Forty-fifth session E/CN.3/2014/11 (2013)

Sample Design for the Integration of Population Census and Social Surveys Il Disegno Campionario per L'integrazione Del Censimento Della Popolazione e delle Indagini Sociali

D'Alò Michele, Falorsi Stefano, Fasulo Andrea and Solari Fabrizio

Abstract Starting from 2018, the Italian National Statistical Institute launched a new census system, named Permanent Census, which, integrating administrative data and data coming surveys, will be carried out every year. This put an end to the era of traditional decennial censuses. The census survey sample is aimed at updating the data contained in the integrated system of registers. Furthermore, the new census will be integrated with the main social surveys. The aim of this work is to compare two sampling strategies for the census survey sample. The first comprises pooling together the samples of the main social surveys, while the second consists of an ad hoc sampling design. Different estimation procedures are taken into account in order to compare the two sampling strategies.

Keywords Population census · Social surveys · Projection estimator

1 Introduction

In 2012, the so-called Permanent Census of Population and Housing was introduced in Italian legislation (Article 3 of Legislative Decree 179/2012, converted with amendments into Law 221/2012). The goal of the Permanent Census is to produce annual data, replacing the previous decennial cycle, using information from administrative sources integrated with an ad hoc sample survey, called Master Sample. The Permanent Census is embedded within the Italian National Institute of Statistics

D. Michele · F. Stefano · F. Andrea (✉) · S. Fabrizio
ISTAT, Rome, Italy
e-mail: fasulo@istat.it

D. Michele
e-mail: dalo@istat.it

F. Stefano
e-mail: stfalors@istat.it

S. Fabrizio
e-mail: solari@istat.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_15

(Istat) modernization program, whose focus is to boost the use of administrative data within the statistical production process.

The census information will be assured by integrating survey data with administrative data stored in the integrated statistical system of registers, which is a set of interlinked base and thematic registers. The new census strategy will allow a significant reduction of census costs, of respondents burden, and of the organizational impact on municipalities (traditionally responsible for the census field work).

This document presents two alternative sampling strategies for the Permanent Census sample: an ad hoc sampling design and the sampling strategy deriving from pooling together the samples of the main social surveys. Section 2 describes the two scenarios. The sampling designs and the estimation methods are presented in Sect. 3. Section 4 reports the results of the simulation study. Finally, last is devoted to the conclusions.

2 Pooled Sample and Census and Social Survey Integrated System Scenarios

Two different sampling strategies have been taken into account for the Permanent Census.

The first scenario consists of pooling together the samples of the main social surveys carried out by Istat, namely Labour Force Survey (LFS), Living Conditions Survey (LCS), Aspects of Daily Life Survey (ADLS), Consumer Expenditure Survey (CES). From now on, we will refer to these pooling samples as Pooled Sample (PS).

The second scenario aims at fully integrating socio-economic information from administrative data and sample surveys. It is based on the Master Sample (MS), which is selected to run the new Permanent Census. The MS sampling design is composed by two components: a list and an areal component. The first is selected from a population frame and it is based on a two stage sampling design. Municipalities and households are the primary and the secondary sampling units respectively. Municipalities are divided into self-representative and non self-representative units. The former are included in the sample every year, while the latter are surveyed only once in 2018–2021 according to rotation scheme. The list component allows the estimation of the census tables that cannot be computed using only the information already available from registers. Furthermore, the Master Sample can be considered as a first phase sample for the main social surveys samples, which can then be viewed as a second phase sample. Indeed, from the Master Sample a set of negatively coordinated samples of households can be selected for the second phase surveys. This strategy is named Census and Social Survey Integrated System (CSSIS). It aims at the harmonization of social surveys and at ensuring a maximum integration with the system of registers. The second component is based on an areal sampling design in which enumeration areas and addresses are the sampling units. It is aimed at estimating under and over coverage rates of the population register.

For an optimal CSSIS design the classification of the survey variables as fully or partially substitutable and not substitutable is worthwhile. The first set of variables are those for which the administrative sources provide the corresponding information. These variables are considered complete since they are available for all population units, and accurate because of their good level of coverage and quality. Administrative sources may provide just a proxy information for a set of target variables. In this case, the variables are classified as partially replaceable, since they are considered complete and accurate only for a subset of the target population, while for the remaining subset of the population they are unknown or are not considered so reliable. Finally, for not replaceable variables no information coming from administrative registers is available. Therefore, the target parameters can be estimated only by means of sample survey data. In this case, the administrative data stored in the population register can be used only as auxiliary information.

In short, the CSSIS is designed to fill up the information gap of the population register by efficiently estimating social and economic parameters of interest that are partially replaceable or not replaceable. This strategy should be able to produce more effective direct estimates than those computed by means of separate survey strategies, even when pooling samples are involved in estimating harmonized common variables.

The Master Sample is a two phase sample. The first phase survey aims to:

- to collect information on partially replaceable and on not replaceable core variables useful for integrating the structural information stored in the population register;
- to set a first contact with the sample households, from which a subsample will be re-interviewed in the second phase module the following year. The first contact could reduce potential second phase non-response obtaining updated contact information on telephone numbers and email addresses. This contact information, which is not available on the sampling frame, may allow to carry out less expensive interview techniques (CAWI or CATI) in the second phase.

The second phase sample is a negatively coordinated sample of households drawn from the MS for the main social surveys and it is aimed at the following:

- to provide information on harmonized and specific socio-economic variables currently observed by LFS, LCS, ADLS, and CES;
- to confirm the common structural variables already surveyed in first phase interview.

The first phase sample is based on a yearly sample size of about 2800 municipalities out of 8,000 and around 1,400,000 households. The first phase sample size should be at least large enough to cover the 140,000 households sample size needed for the second phase. For an overview of CSSIS, based on the two phase MS design, see Fig. 1.

Referring to similar international experiences, analogous modular approaches have been proposed by Eurostat for the design of integrated social surveys. Furthermore, the ABS is designing an integrated system of investigations very similar to

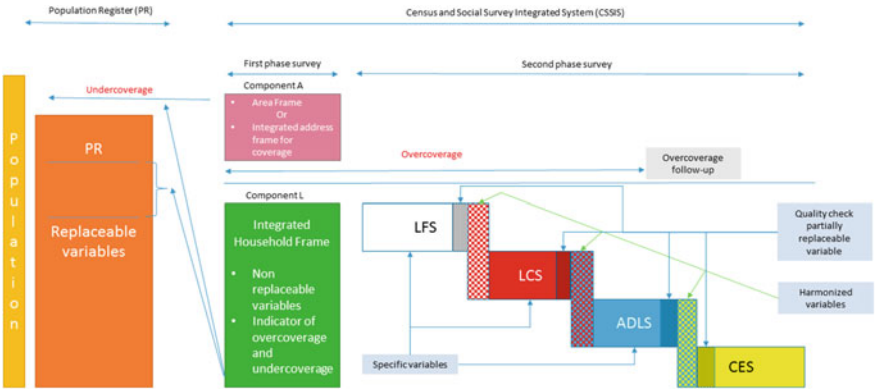


Fig. 1 An overview of the CSSIS

what described here and called Australian Population Survey, does not replace the census. The design with two components supporting the register census is similar to what ONS has been studying for the register-based census supposed to start in 2023 after the 2021 census run [7]. In particular, in 2021 the ONS will conduct a traditional census and, at the same time, will carry on a parallel census run based on the construction of an integrated population register using several administrative sources and a survey sample, similarly to the Italian MS. It is worthwhile to mention that every year since 2015 and until 2023 the ONS will produce an assessment to evaluate how much they are away from the model to be. Another international experience showing similarities with what is planned in Italy is the Israelian rolling integrated census. They use an integrated register which is adjusted by means of weights computed by means of an EDSE [8].

3 Estimation Methods

This section is devoted to the description of the estimation methods used to compare the properties of the two alternative sampling design strategies: the Pooled Sample strategy (scenario 1) and the Master Sample strategy (scenario 2). Both design and model based estimation methods are considered.

The estimators taken into consideration are:

- ratio estimator, which is applicable only for in-sample domains;
- design-based projection estimator from Master/Pooled Sample to register;
- model-based projection estimator from Master/Pooled Sample to register.

The projection estimator, proposed by [5], is an asymptotically unbiased model-assisted estimator that combines information from different sources, using common unit-level auxiliary information. A working model is fitted to the units of a smaller

sample, and synthetic values are then obtained for the units of a larger sample or, if available, for the units of the register. Denoting with k and d the generic sampling unit and the generic domain respectively, the following linear model is considered:

$$\boldsymbol{y}_{kd} = \boldsymbol{x}_{kd}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{kd}, k = 1, \ldots, n_d, d = 1, \ldots, D,$$

$$\mathrm{E}(\boldsymbol{\varepsilon}_{kd}) = 0, \mathrm{Var}(\boldsymbol{\varepsilon}_{kd}) = \sigma^2, k = 1, \ldots, n_d, d = 1, \ldots, D, .$$

The sampling weights w_k are used for the estimation of the model parameters:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{kd} \boldsymbol{x}'_{kd}\boldsymbol{x}_{kd}w_{kd}\right)^{-1}\left(\sum_{kd} \boldsymbol{x}'_{kd}y_{kd}w_{kd}\right).$$

Fitting the model using MS or PS data, it is possible to project synthetic values $\hat{y}_{kd} = \boldsymbol{x}_{kd}\hat{\boldsymbol{\beta}}$ of the variable of interest on the register. This method requires a high level of quality of the auxiliary variables and a high goodness of fit of working models.

The model based estimator considered in the experimental study is the model based counterpart of the projection estimator. In this case a linear mixed model is considered as in [1]. D'Alò and Solari [2] showed that the model-based projection estimator is equivalent to the EBLUP related to the unit level linear mixed model. Denoting with k the generic sampling unit and with d the generic domain, the linear mixed model is specified as:

$$y_{kd} = \boldsymbol{x}_{kd}\boldsymbol{\beta} + v_d + \varepsilon_{kd}, k = 1, \ldots, n_d, d = 1, \ldots, D,$$

$$\mathrm{E}(\varepsilon_{kd}) = 0, \mathrm{Var}(\varepsilon_{kd}) = \sigma^2, k = 1, \ldots, n_d, d = 1, \ldots, D,$$

$$\mathrm{E}(v_d) = 0, \mathrm{Var}(v_d) = \sigma_v^2, d = 1, \ldots, D,$$

where v_d is an area random effect included in the model to take into account the between area variability. Estimation of $\boldsymbol{\beta}$ and of the variance components σ^2 and σ_v^2 can be obtained using Maximum Likelihood or Restricted Maximum Likelihood iterative procedures (see [6]). Analogously to the design based case, synthetic values, defined as $\hat{y}_{kd} = \boldsymbol{x}_{kd}\hat{\boldsymbol{\beta}} + \hat{v}_d$, can be projected from the sample to the register.

4 Simulation Study

In this section a case study to compare the MS and PS designs using the set of estimation methods illustrated in the previous section is described. To this aim four

sub-regional domains are considered: provinces, aggregation of Labour Market Areas (macro-LMAs), Labour Market Areas (LMAs), municipalities. In order to evaluate the empirical properties of the estimates in terms of bias and mean square error, a Monte Carlo simulation study has been carried out using the 2011 Population Census data. Two hundreds samples have been drawn from the 2011 Italian Population Census, for two Italian regions, Trentino-Alto-Adige and Marche, using both MS and PS sampling designs. The sample size of each simulated sample is equal to 140,000 households, that is the sampling size needed for the second phase, that is selecting the samples for the four social surveys.

The target variables taken into account are the employed and unemployed counts in the two regions.

The linear model adopted for the design-based projection estimator uses an intercept at province level. The auxiliary information considered in the model specification is: marital status, citizenship, 28 age-gender classes, and an additional variable taken from an administrative register built in Istat (see [4]). The last variable is an individual indicator of the presence of signals in at least one administrative source related to the employment market. The same covariates are used in the fixed part of linear mixed model, while the area random effects are defined at provincial level.

All the estimators are compared by means of standard indicators of accuracy of prediction: the Average Absolute Relative Bias (AARB) and Average Relative Root Mean Squared Error (ARRMSE). Expression of the evaluation indicators are given by:

$$\text{AARB} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{200}\sum_{r=1}^{200}\left|\hat{Y}_{rd} - Y_d\right|,$$

$$\text{ARRMSE} = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{200}\sum_{r=1}^{200}\left(\hat{Y}_{rd} - Y_d\right)^2},$$

where $\hat{\boldsymbol{Y}}_{rd}$ and $\boldsymbol{Y}_d$ are, respectively, the predicted value in the r-th simulated sample and the correspondent true value of the target variable y in the domain d.

Tables 1 and 2 report the results for the variable employment status for MS and PS, respectively. Tables 3 and 4 display the analogous outputs for the unemployed counts. AARB and ARRMSE indicators are computed for the four types of domain considered in the case study.

As can be seen from the tables, it results that the design-based projection estimator often outperforms the other two types of estimator. When the target variables is the employment status, the ratio estimator and the design-based projection estimator show good performances in terms of AARB. However, the ratio estimator is always worse than the design based projection estimator in terms of ARRMSE. Slightly worse performances are displayed for all type of estimators when moving from the largest domains (provinces) to the smallest ones (municipalities). Furthermore, the Master Sample strategy leads to better results than the Pooled Sample strategy, especially in terms of ARRMSE.

Table 1 Pooled sample—AARMB and ARRMSE for the variable employed

Average relative root mean squared error			
	Ratio estimator	Design-based projection	Model-based projection
Provinces (7)	0.3	0.5	–
Macro LMA (20)	0.4	0.1	2.4
LMA (54)	0.8	1.1	2.7
100% in-sample LMA (26)	0.6	0.4	2.5
90% in-sample LMA (41)	0.7	0.6	2.4
50% in-sample LMA (50)	0.7	0.8	2.6
Municipalities (572)	1.8	2.0	3.5
100 % in-sample municipalities (27)	1.1	0.6	2.1
90% in-sample municipalities (32)	1.4	0.7	2.1
50% in-sample municipalities (113)	1.5	0.9	2.8
Average relative root mean squared error			
	Ratio estimator	Design-based projection	Model-based projection
Provinces (7)	4.1	2.3	–
Macro LMA (14)	2.7	1.3	3.7
LMA (54)	7.0	1.9	4.0
100% in-sample LMA (26)	5.0	1.3	3.8
90% in-sample LMA (41)	5.3	1.4	3.8
50% in-sample LMA (50)	6.1	1.6	3.9
Municipalities (572)	11.1	1.3	4.2
100% in-sample municipalities (27)	9.2	1.5	3.7
90% in-sample municipalities (32)	9.5	1.5	3.6
50% in-sample municipalities (113)	10.5	1.6	4.0

In regard to the estimation of unemployment counts, the values reported in Tables 3 and 4 displays that the ratio estimator outperforms the other two estimators in terms of AARB. The design based projection estimator is, instead, the best choice in terms of ARRMSE. The bad performance in terms of bias of the design based projection

Table 2 Master sample—AARMB and ARRMSE for the variable employed

Average absolute relative bias			
	Ratio estimator	Design-based projection	Model-based projection
Provinces (7)	0.3	0.1	–
Macro LMA (20)	0.5	0.5	2.3
LMA (54)	0.7	1.2	2.9
100% in-sample LMA (30)	0.5	0.6	2.6
90% in-sample LMA (38)	0.6	0.6	2.5
50% in-sample LMA (50)	0.6	0.9	2.7
Municipalities (572)	1.6	2.0	3.6
100 % in-sample municipalities (49)	1.0	0.9	3.0
90% in-sample municipalities (50)	1.3	0.9	3.0
50% in-sample municipalities (99)	1.4	1.0	3.1
Average relative root mean squared error			
	Ratio estimator	Design-based projection	Model-based Projection
Provinces (7)	2.6	1.0	–
Macro LMA (20)	3.7	1.1	3.5
LMA (54)	6.6	1.7	4.0
100% in-sample LMA (30)	4.0	1.1	3.8
90% in-sample LMA (38)	5.5	1.1	3.7
50% in-sample LMA (50)	5.7	1.4	3.9
Municipalities (572)	8.9	2.3	4.6
100% in-sample municipalities (49)	4.7	1.3	4.2
90% in-sample municipalities (50)	4.8	1.3	4.1
50% in-sample municipalities (99)	5.5	1.4	4.2

estimator for the unemployment status is likely due to the fact that the model is significantly less predictive than employment status case.

The MS strategy seems to outperform the PS strategy for LMA and municipalities and to be less efficient for the largest domains, that is provinces and Macro-LMAs. The ARRMSE values related to the 100, 90 and 50% in-sample municipalities for the

Table 3 Pooled sample—AARMB and ARRMSE for the variable unemployed

Average relative root mean squared error			
	Ratio estimator	Design-based projection	Model-based projection
Provinces (7)	0.4	0.5	–
Macro LMA (20)	1.0	1.1	34.4
LMA (54)	3.2	12.3	48.8
100% in-sample LMA (26)	1.5	5.2	35.8
90% in-sample LMA (41)	1.7	9.7	40.0
50% in-sample LMA (50)	1.8	11.3	48.0
Municipalities (572)	10.9	33.4	73.2
100 % in-sample municipalities (27)	2.0	9.9	30.2
90% in-sample municipalities (32)	2.2	8.8	31.1
50% in-sample municipalities (113)	4.1	17.1	48.3
Average relative root mean squared error			
	Ratio estimator	Design-based projection	Model-based projection
Provinces (7)	10.4	9.4	–
Macro LMA (14)	15.2	14.3	42.2
LMA (54)	42.6	21.7	57.3
100% in-sample LMA (26)	24.1	15.2	44.5
90% in-sample LMA (41)	27.1	19.2	54.6
50% in-sample LMA (50)	29.2	20.9	56.7
Municipalities (572)	75.4	40.9	83.0
100% in-sample municipalities (27)	29.3	17.7	39.4
90% in-sample municipalities (32)	34.1	16.9	39.8
50% in-sample municipalities (113)	45.7	25.2	58.0

projection estimators in the Pooled Sample case are smaller than the corresponding values for the Master Sample case. This is simply due to the fact that the number of in-sample LMAs or municipalities in PS is smaller than the analogous size in MS. For instance, in the Pooled Sample 27 municipalities are always included in the samples, while for the Master Sample the corresponding value is 49. This means

Table 4 Master sample—AARMB and ARRMSE for the variable unemployed

Average absolute relative bias			
	Ratio estimator	Design-based projection	Model-based projection
Provinces (7)	0.3	0.5	–
Macro LMA (20)	1.1	1.3	37.0
LMA (54)	2.9	15.9	45.7
100% in-sample LMA (30)	1.3	10.3	36.3
90% in-sample LMA (38)	1.5	13.7	42.5
50% in-sample LMA (50)	1.9	15.1	44.9
Municipalities (572)	8.8	34.8	69.5
100 % in-sample municipalities (49)	2.2	16.1	38.7
90% in-sample municipalities (50)	2.6	15.8	38.7
50% in-sample municipalities (99)	4.2	22.1	49.1
Average relative root mean squared error			
	Ratio estimator	Design-based projection	Model-based projection
Provinces (7)	11.2	10.1	–
Macro LMA (20)	20.6	16.0	44.9
LMA (54)	39.5	20.3	53.8
100% in-sample LMA (30)	22.5	15.9	44.7
90% in-sample LMA (38)	24.9	18.3	50.7
50% in-sample LMA (50)	27.4	19.7	53.0
Municipalities (572)	61.0	38.8	78.7
100% in-sample municipalities (49)	28.5	20.4	48.7
90% in-sample municipalities (50)	29.6	20.3	48.4
50% in-sample municipalities (99)	32.4	26.3	58.9

that the sampling units of the Pooled Sample are spread over a smaller number of municipalities than the sampling units of the Master Sample. This implies that, in case of pooling strategy, the in-sample municipalities usually have larger sample sizes than the corresponding in-samples municipalities drawn with the MS sampling strategy. Only the results for the 50% in-sample LMA are directly comparable for MS

and PS. In this case, the MS strategy outperforms the PS strategy for the estimation of both employment and unemployment counts.

5 Conclusions

In this work two different sampling strategies for carrying out the new Italian census are compared: a strategy involving the pooling of social surveys sample, named Pooled Sample strategy, and an ad hoc census strategy, named Master Sample strategy.

The Master Sample strategy results to outperform the Pooled Sample strategy in many combinations of domain and estimator taken into account. Furthermore, in the Master Sample case the sampling units are spread over a larger number of domains than what happen in the Pooled Sample case. This could result particularly useful when adopting, for instance, small area estimation methods.

Besides, the areal component of the Master Sample design allows to estimate the coverage of the population registers, which is not possible to estimate when adopting the pooling strategy.

The second phase of the Master Sample permits a complete integration and harmonization between the statistical system of register and the main social surveys carried out by Istat. This aspect is extremely important since it follows the framework proposed by [3] on the harmonization of social surveys.

Appendix

See Tables 1, 2, 3 and 4

References

1. Battese, G.E., Harter, R.M., Fuller, W.A.: An error component model for prediction of county crop areas using survey and satellite data. J. Am. Stat. Assoc. **83**(401), 28–36 (1988)
2. D'Alò M., Solari, F.: Integrated estimation combining different sources of information, in Technical Report on the Integrated Survey Framework, FAO Technical Report Series GO–02–2014, pp. 214–266 (2014)
3. EUROSTAT: Roadmap for the integration of European social surveys (2013). http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D12_Roadmap.pdf
4. Garofalo, G.: Il progetto ARCHIMEDE: obiettivi e risultati sperimentali. Istat Working Papers, 9/2014 (2014). https://www.istat.it/it/files//2014/11/IWP-n.-9-2014.pdf
5. Kim, J.K., Rao, J.N.K.: Combining data from two independent surveys: a model-assisted approach. Biometrika **991**, 85–100 (2012)
6. Molina, I., Rao, J.N.K.: Small Area Estimation, 2nd edn. Wiley, New York (2015)

7. ONS: Annual assessment of ONS's progress towards an administrative data census post-2021 (2016). https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments
8. Pfeffermann, D.: Methodological issues and challenges in the production of official statistics. J. Surv. Stat. Methodol. **3**, 425–483 (2015)

Sampling Schemes Using Scanner Data for the Consumer Price Index

Claudia De Vitiis, Alessio Guandalini, Francesca Inglese and Marco Dionisio Terribili

Abstract The Italian National Institute of Statistics (ISTAT) is carrying out a redesign of Consumer Price Survey (CPS). The availability of Scanner Data (SD) from retail modern distribution, provided to ISTAT by Nielsen for a large number of stores selling food and grocery, is the starting point of this innovation. Indeed, SD represent a big opportunity for improving the computation of Consumer Price Index (CPI). This work aims to study the properties of alternative aggregation formulas of the elementary price index in different sampling schemes implemented on SD. Bias and efficiency of the estimated indices are evaluated through a Monte Carlo simulation.

Keywords Consumer price index · Scanner data · Sampling · Fixed and dynamic approach

1 Introduction

The Italian National Institute of Statistics (ISTAT) is carrying out a redesign of the Consumer Price Survey (CPS). The main aim of the project is to modernise the survey, improving and unburdening the data collection phase, together with the progressive introduction of more rigorous sampling procedures, probabilistic where possible, for the selection of outlets and products (items) [1, 6].

C. De Vitiis (✉) · A. Guandalini · F. Inglese · M. D. Terribili
ISTAT, Rome, Italy
e-mail: devitiis@istat.it

A. Guandalini
e-mail: alessio.guandalini@istat.it

F. Inglese
e-mail: fringles@istat.it

M. D. Terribili
e-mail: terribili@istat.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_16

The current strategy of the CPS, carried out at territorial level, is based on three purposive sampling stages. The sampling units are respectively the municipalities, the outlets and the elementary items for which the prices are collected. At each stage the selection criterion is based on the concept of most representative units: the first stage units are the chief town of provinces; the outlet sample is chosen to be representative of the consumer behavior in the municipality; for each product of the basket the most sold item is selected and the prices of these items are collected throughout the year. At municipality level the elementary price indices are currently obtained by unweighted geometric mean. The general price index is calculated using a Laspeyres type index, through subsequent aggregation of elementary indices: at different levels weights are used based on population proportions and national account data on consumer expenditure.

The availability of Scanner Data (SD) from the retail modern distribution (food and grocery) are the starting point for the implementation of the innovation in the survey (CPS). At present food and grocery sector cover 11% of the total and modern distribution the 55% of the total. Another aspect involved in the review of the survey is the use of web-scraping for collecting prices referred to online market (tourism, mobile phone, etc.).

Scanner data files contain elementary information (turnover and quantities) referred to the items which are sold weekly in a specific outlet: each item is uniquely identified by its barcode (GTIN—Global Trade Item Number, or EAN—European Article Number). Information available on turnover and quantities sold in a week do not provide the "shelf price" of the references (or series individuated by EAN and outlet codes) but allows to define a unit value or average weekly price. Due to operational constraints of the productive process, a restriction is introduced regarding the *observable* weeks: only the relevant weeks are considered, defined as the first three full weeks (composed of seven days) in each month.

SD introduce important advantages compared to data collected through traditional survey. In particular, the availability of turnover and quantity data at item level offer a real possibility of calculating more accurate indices: it is possible, in fact, to include in the calculus the expenditure share of each product sold. SD also contain descriptive information about items characteristics useful to treat quality change, to identify relaunches of existing products or new products, etc. [2, 7].

On the other hand, the use of SD in the compilation of Consumer Price Index (CPI) must take into account some important drawbacks, as attrition of products, temporary missing products, entry of new products and volatility of the prices and quantities due mainly to sales. These are aspects that need to be addressed from both a theoretical and a practical point of view [5].

To maximize the potential offered by SD it would be necessary to go beyond those methods of price index compilation which do not exploit all the information provided by the data and do not take into account the population dynamics [2]. Weighted and chained indices should be considered to incorporate the overall price trend over a given time, including the prices of new products. Furthermore, the problem of shrinkage over time due to the attrition of a fixed basket of products is solved automatically using chain indices. However, even though in a dynamic approach it is necessary

to construct series of chained indices, high-frequency chaining of weighted indices (also superlative Fisher and Törnqvist indices) are affected by *chain drift*, due to non-symmetric effects on quantities sold and expenditure share of goods before and after sale (Ivancic et al. 2011; [4]).

In recent years, an important debate has taken place among the researchers dealing with the estimate of the consumer price index starting from SD. The focus, above all, has been on the transition from a static population approach (fixed basket) to a dynamic population approach (flexible basket) and it is based on the study of alternative price index formulas based on matched-model methods (matching of products sold during two months in a row) or other methods that are transitive and, therefore, free from chain drift [5].

Other aspects discussed are the quality of SD (completeness and correctness) and the definition of methods to treat appearing and disappearing products, temporary missing products, relaunches, quality change, etc. [15, 16].

The aim of this paper is to present the ISTAT SD experimental framework in which, firstly, probability and nonprobability selection schemes of series (references individuated by EAN and outlet codes) are compared and then different probability sampling designs are examined.

As the main objective of this experimental phase has been to study the robustness of several price index estimators under different selection schemes, the sampling frame has been defined by considering a *panel* data set that contains *permanent* series.

Moreover, through a further experiment, the differences between a fixed and a dynamic population approach in the construction of the elementary price indices are highlighted. In this case the purpose is trying to measure the magnitude of sampling error and non-sampling error for different price index formulas in both approaches. When a static population is assumed, non-sampling errors are generated by disappearing products, ignoring entries of new products and temporary missing products. In a dynamic population context, non-sampling errors are generated by the variability in the number of matched-items between the months and the use of weighted index formulas (chain drift).

An important issue, which is out of the purpose of this paper but is crucial for the ISTAT CPI, is the need for combining estimates derived from SD with those that will be still produced through current on field survey for the traditional retail distribution.

Currently ISTAT is performing a transition phase in which SD are used for the production of CPI but following a static approach not far from the traditional survey. However, the dynamic approach is tested, too. The indices derived from SD, both following fixed or dynamic approach, are combined with indices deriving from the traditional survey using weights of modern distribution and traditional distribution estimated by the consumer expenditure survey.

The paper is organized as follows: Sect. 2 provide a brief overview of the use of SD for the CPI compilation in some countries; Sect. 3 presents the context and the methodological approach of experiments used to compare different sampling designs from SD—analysis on available SD, description of different selection schemes of series; Sect. 4 shows the main results regarding the accuracy of price indices estimates. Finally, in Sect. 5 some conclusions and future developments are exposed.

2 Use of Scanner Data for CPI

At the end of 2014 the Italian National Institute of Statistics (ISTAT), through a contract with Nielsen and an agreement with the six main retail chains operating in Italy, started receiving SD referred to food and grocery markets and processing them for experimenting the calculus of CPI. This data acquisition places Italy among countries using or testing the use of this source of data for compiling CPI.

Moreover, Nielsen provide the dictionary for the classification of EAN codes to GS1-ECR-Indicod product classification. ISTAT ensures internally the translation from ECR to COICOP, the classification of products used for the CPI. Consumption segments, which are not coded in the EU-COICOP, are the most detailed domain of estimate for the Italian CPI and are constituted by groups of homogeneous products; those defined for the food and grocery are 126 out of a total of 324.

As noted some years ago in the ILO CPI Manual [9], "*Scanner data constitute a rapidly expanding source of data with considerable potential for CPI purposes*" (p. 54); "*Scanner data obtained from electronic points of sale include quantities sold and the corresponding value aggregates on a very detailed level*" (p. 92); "*Scanner data are up to date and comprehensive*" (p. 478).

Scanner data have been used in the compilation of the CPI for some years in four countries, i.e., Switzerland, Norway, Netherlands and Sweden, while Belgium and Denmark started only from 2016. SD can be exploited in different ways. The simplest way is using SD as an alternative source for price collection, replacing collection within the stores, without changing the traditional principles of computing the price indices. This method is currently applied by the Swiss Federal Statistical Office [16]. Alternatively, as in Norway and Sweden, SD can be used as universe from which samples of references can be selected following different methods [11, 12]. Finally, all (or almost all) SD can be used to compile price indices, without a strict sample selection, but with consequences on the theoretical definition of the index. In the Netherlands, the computation methods is different and the data are used in a more extensive way to calculate price indices [15]. The method used assumes a dynamic population approach: elementary price indices of homogeneous items are calculated by monthly chained unweighted geometric index (Jevons); no explicit weighting is applied and expenditure information is used just to select a cut-off sample of matched items during two months in a row.

In a study perspective, moreover, SD from retail stores allow researchers to evaluate how different price index formulas perform at the elementary level. In fact, official CPI are usually constructed in two broad steps. First, elementary price indices are calculated for narrowly defined and relatively homogeneous products, known as elementary aggregates. In a second step, these elementary indices are aggregated into a single consumer price index using expenditure weights. Elementary indices, named also higher level elementary indices, are therefore the building blocks of price index numbers.

While the aggregation at higher level is carried out using generally Laspeyres type formulas with weights deriving from national account or expenditure survey data,

official practices in elementary price index construction are still not uniform across countries, deserving further investigation in the consequences of different choices [8].

3 Context and Methods for Sampling Scanner Data Series

3.1 Outline

The experiments carried out so far in ISTAT on the SD aimed to evaluate the properties of the weighted and unweighted elementary price indices in different selection schemes of *series*, under a fixed population approach. In this experiments the implications of life-cycle of series, seasonality issues and missing data are not taken into account and a simplification is used: only panel series are considered as universe for sampling and price index evaluation.

In this context the definition of panel data is based on the *permanent series* concept which refers to those series with not-null turnover for at least one relevant week (the first three full weeks) in each month of the considered year, starting from the December of previous year. A sample of series is selected at the beginning of the reference period and it is followed during the whole year, without considering either new entries nor discontinuities. For each selection scheme, starting from the monthly price ratios with fixed base (December 2013) available for 2014, the elementary price indices is calculated using three classic aggregation formulas: Jevons (unweighted), Fisher (ideal) and Lowe (weights from quantities of previous year). The choice of these indices is made on the basis of theoretical and empirical considerations: Fisher ideal index is thus preferred by economic theory, as it uses quantities in different times and allows for substitution effects.

The experimental study is developed in two phases: firstly, probability and non-probability selection schemes of series are compared; then, several sampling designs are considered, each of them characterized by the use of different criteria of sample allocation, both for outlets and elementary items (EANs), and by different selection methods of the sampling units. The comparison among the alternative selection schemes is made, for each price index formula, taking the corresponding true value of the index computed on the whole universe as a benchmark. Indices performance is evaluated in terms of bias for all selection schemes. For probability selection schemes, accuracy (bias and sampling variance) of the price indices are studied through a Monte Carlo simulation: 500 samples are selected, for each different sampling design. Variability and bias are computed on the estimated indices in the replicated samples. The sample selection and the weighting of price indices is based on the total annual turnover of 2013.

Explorative analyses have been conducted beforehand on SD relative to the six retail chains (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) available in 2014 for five Italian provinces. The Turin province and some consumption segments (Coicop 6 digits) has been chosen for the experiments.

3.2 Analysis of Scanner Data

Scanner data acquired by ISTAT in the first phase and analyzed herein, cover five Italian provinces (Ancona, Cagliari, Palermo, Piacenza, Torino) and six chains of modern distribution (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) for years 2013 and 2014, for hypermarkets and supermarkets. Afterward data from other chains have been acquired up to now, from sixteen chains. The Italian modern distribution sector is, in fact very complex and heterogeneous, especially with respect to the territory: different chains cover different regions in a very jeopardized way and with high variability.

In the following tables some aspects of the chains and outlet type distributions in the five provinces are highlighted. The analysis has been carried out on the whole of the 289 outlets of the 5 provinces for the 52 weeks of the year 2014. Table 1 contains the whole turnover and the number of outlets by chain and province and the percent coverage of the six chains with respect to the total turnover of modern distribution for food and grocery at province level. The table shows a heterogeneous situation both among the chains and the provinces: Turin province represent more than 50% of turnover involved; this fact could be influenced by the high number of outlets observed in this province, 130 on a whole set of 289. In the other provinces

Table 1 Total turnover (in EUR'000) and number of outlets by chain and province (2014)

Chain		Province					
		Ancona	Cagliari	Palermo	Piacenza	Turin	Total
A	Turnover	14'287	63'553	56'233	38'883	47'924	220'882
	Outlet	3	10	15	7	13	48
B	Turnover	79'793	–	35'976	32'823	272'443	421'035
	Outlet	10	–	2	4	26	42
C	Turnover	–	–	–	58'936	108'144	167'080
	Outlet	–	–	–	2	3	5
D	Turnover	106'364	54'248	64'424	4'891	157'247	387'174
	Outlet	11	2	7	1	8	29
E	Turnover	11'094	38'062	64'896	2'659	424'278	540'990
	Outlet	1	2	21	1	40	65
F	Turnover	74'711	77'909	–	28'197	140'879	321'696
	Outlet	34	21	–	4	39	98
Total	Turnover	286'249	233'773	222'098	166'390	1'152'109	2'058'857
	Outlet	59	35	46	19	130	289
Coverage turnover		87.00	74.28	69.95	73.68	72.65	73.96

Table 2 Total turnover (in EUR'000), number of outlets and number of items by province and outlet type (2014)

Province		Outlet type		Total
		Hypermarket	Supermarket	
Ancona	Turnover	136'000	150'249	286'249
	Outlet	10	49	59
	Item	63'112	43'000	106'112
Cagliari	Turnover	126'002	107'771	233'773
	Outlet	7	28	35
	Item	53'053	26'418	79'471
Palermo	Turnover	81'250	140'129	222'097
	Outlet	4	41	46
	Item	40'532	33'349	73'881
Piacenza	Turnover	83'803	82'586	166'390
	Outlet	4	15	19
	Item	44'341	50'466	94'807
Turin	Turnover	738'419	412'833	1'152'109
	Outlet	28	101	130
	Item	83'342	54'420	137'762

the number of outlets varies from 19 (Piacenza) to 59 (Ancona), with a turnover between 166 and 286 million euro. The last row shows the high level of coverage of the six chains, although with a certain heterogeneity among provinces: the coverage is generally close to 72%, with a maximum level of 87% assessed in Ancona province.

In the table above, turnover and number of outlets are reported considering the six chains and two outlet types. This variable is particularly important because it is connected both with the chain (which can have a higher/lower propensity to set up a hypermarket/supermarket) and with the number of elementary items sold, much higher in the hypermarkets than in the supermarkets. Table 2 shows that a hypermarkets have turnover higher than supermarkets, although the number of outlets is reasonably lower (53 hypermarkets and 234 supermarkets, for a total of 289 outlets). Moreover, hypermarkets are more in the Turin province (28 outlets) than in Palermo and Piacenza provinces (4 outlets). This evidence has to be linked to the local chains distribution policies, in fact some chains have a greater number of hypermarkets than the other chains.

Information on the observed series (EAN + outlet code) belonging to the relevant weeks of each month and on permanent series (panel series SD), as defined above, is reported in Table 3. It shows the whole turnover observed for the five provinces, respectively on the whole set of series (A), on the relevant week series (B) and on the panel series (C).

Table 3 Total turnover (in EUR'000) for all series, relevant week series and panel series, and number of panel series (in thousands) by province (2014)

Province	Turnover			% Coverage		No. of panel series
	All weeks all series (A)	Relevant weeks all series (B)	Relevant weeks panel series (C)	B/A	C/B	
Ancona	286'249	199'337	134'533	69.94	67.49	2'331
Cagliari	233'773	162'008	109'160	69.30	67.38	1'583
Palermo	222'097	153'925	91'513	69.31	59.45	1'389
Piacenza	166'390	115'388	82'736	69.35	71.70	1'123
Turin	1'152'109	793'434	562'758	68.87	70.93	7'185

Observing only relevant weeks allows to take into account about 70% of turnover of all weeks (52 weeks of the year 2014), without important local differences. Then, looking at the coverage turnover of the panel series with respect to relevant weeks series, it ranges from 59.45% (Palermo) to 71.70% (Piacenza).

3.3 Selection Schemes

In the first phase of the experiment, a nonprobability sampling scheme is defined by selecting series on the basis of cut-off thresholds of covered turnover in previous year, 2013: two samples of series covering respectively 60 and 80 percent of the total turnover in each of the considered consumption segment (coffee, pasta, mineral water) are considered. Moreover, considering the fixed basket approach currently used, a reference selection scheme is defined selecting the most sold EANs for each representative product in each outlet.

These nonprobability selection schemes are compared with a two-stage probability sampling design, where primary units (PSU) and secondary units (SSU) are respectively outlets and EANs. The size of the sample of outlets has been fixed at a number of 30 out of 121 outlets available for the Turin province. The sample size for SSU is fixed by a sampling rate of 5 percent of the number of EANs in each consumption segment in the sampled outlets. This choice has been made for computational reasons, linked to the processing capacities. Outlets are stratified by chain and outlet type (hypermarket and supermarket). In each stratum, the sample has been allocated proportionally to the turnover. The selection of outlets is carried out in each stratum by simple random sampling (SRS), while EANs are selected with probability proportional to size (PPS), on the basis of the total turnover of previous year, by adopting Sampford sampling [13, 14].

In the second phase of the experiments the following sampling designs are compared: (1) one stage stratified sample of EANs; (2) cluster sample of outlets (all EANs); (3) two-stage sampling with stratification of PSU (outlet) and SSU (EAN).

For each sampling design the size of the final sample of EANs is fixed in average at 7'400 to compare the different sampling strategies on equal computational effort. Moreover, different criteria of sample allocation, both for outlets and EANs, and different selection methods of the units are considered.

The first sampling design is carried out stratifying the EANs by market (ECR group) in each consumption segment (considering coffee, pasta, mineral water, olive oil, spumante and ice cream). Sample size is allocated among the strata through a Neyman formula, taking into account the variability of prices relatives in the markets observed in the reference year 2013. Two selection schemes are considered, SRS and PPS.

In the second design, cluster sampling, a sample of outlets (14 out of 121 outlets) is selected. Outlets are stratified by chain and type. In each stratum, two different allocation of outlets are tested: proportional to the strata turnover and optimal allocation (Neyman). Outlets are selected with both SRS and PPS methods. All the EANs in the selected outlets are included in the sample.

Finally, two-stage sampling design is characterized by a stratification of both PSU and SSU. The stratifications adopted for the PSU and the SSU are the same of the two schemes described above. The size of the sample of outlets is fixed at a number of 30 out of 121 outlets. For both outlets and EANs, sample allocation in the strata is proportional to the strata turnover. PSU are selected with a PPS method, while SSU are selected both with SRS and PPS methods.

3.4 Parameters and Unbiased Estimators

The parameters of interest are monthly Jevons, Fisher and Lowe indices. Jevons index is an unweighted CPI that uses price information only (it assumes that expenditure shares remain constant), while Fisher and Lowe use also quantity information. Fisher and Lowe indices consider turnover shares at different time periods as weights [8]. Indicating by the subscript t the current month (12 months in year 2014), t_0 the reference month (December 2013), l the previous year (2013), c ($c = 1,\ldots,C$) the generic homogeneous products group and m ($m = 1,\ldots,M_c$) the series, unbiased sampling estimators [3] of population parameters (elementary price indices aggregation) can be expressed as follows.

$$JEVONS^{\bullet}_{ct} = \prod_{m}^{M_c} \left(\frac{p_{cmt}}{p_{cmt_0}} \right)^{w_{cml} / \sum_m w_{cml}}$$

$$LASP^{\bullet}_{ct} = \sum_{m}^{M_c} \left(\frac{p_{cmt}}{p_{cmt_0}} \right) * \left(\frac{p_{cmt_0} * q_{cmt_0} * w_{cml}}{\sum_m^{M_c} p_{cmt_0} * q_{cmt_0} * w_{cml}} \right)$$

$$PAAS^{\bullet}_{ct} = \sum_{m}^{M_c} \left(\frac{p_{cmt_0}}{p_{cmt}} \right) * \left(\frac{p_{cmt} * q_{cmt} * w_{cml}}{\sum_m^{M_c} p_{cmt} * q_{cmt} * w_{cml}} \right)$$

$$FISH^{\bullet}_{ct} = \sqrt{LASP^{\bullet}_{ct} * PAAS^{\bullet}_{ct}}$$

$$LOWE^{\bullet}_{ct} = \sum_{m}^{M_i} \left(\frac{p_{cmt}}{p_{cmt_0}} \right) * \left(\frac{p_{cmt_0} * q^{z}_{cml} * w_{cml}}{\sum_{m}^{M_c} p_{cmt_0} * q^{z}_{cml} * w_{cml}} \right)$$

with $q^{z}_{cml} = \sum_{a=0}^{11} q_{cm(t_0-a)}$

The measure q^{z}_{cml} refers to the mth quantity series in the previous year l (2013), while the weight w_{cm} is obtained as the inverse of the inclusion probability of the sampling unit deriving from the sampling design.

3.5 Accuracy of Price Index

The accuracy of the estimated price indices (Lowe, Fisher and Jevons) is evaluated, for the probability sampling designs, on the replicated samples obtained through a Monte Carlo simulation.

Bias and relative sampling error formulas shown below are expressed for a generic parameter (price index) and with reference to the simulation context.

For a generic estimated index in the cth generic homogeneous products group, $\hat{\theta}_c$, bias and absolute relative bias (ARB) can be expressed as

$$B\left(\hat{\theta}_c\right) = E\left[\hat{\theta}_c\right] - \theta_c, \quad ARB = B\left(\hat{\theta}_c\right)/\theta_c$$

In the formulas, $E\left[\hat{\theta}_c\right]$ is the expected value of the estimated index $\hat{\theta}_c$ in the cth generic homogeneous products group, obtained from replicated samples, and θ_c is the corresponding index value computed on the reference universe (panel series SD). The Monte Carlo simulation regards the probabilistic sample schemes, while for the non-probabilistic design only one estimate is provided.

The relative sampling error of a generic estimated index $\hat{\theta}_c$ in the cth generic homogeneous products group can be expressed by

$$RE\left(\hat{\theta}_c\right) = \frac{\sqrt{Var\left(\hat{\theta}_c\right)}}{\hat{\bar{\theta}}_c},$$

in which mean and variance of $\hat{\theta}_c$ are calculated on the estimates generated from the selection of the simulated samples in the cth generic homogeneous products group.

4 Main Results

The most meaningful results of the two experimental phases are shown in the following tables and figures. In the tables the accuracy of the estimated price indices is shown for the probability samples evaluated through the Monte Carlo simulations. In the figures the analysis focuses on the comparison of the levels of the estimates obtained through the considered sampling schemes for the first experiment and on the comparison of precisions for the second experiment.

For each estimated price index, in the tables below, the values assumed by the absolute relative bias (ARB) and relative sampling error (RSE) distributions of the 12 monthly indices are exposed for each consumption segment. The generally low levels of ARB confirm empirically the unbiasedness of the estimators; moreover itmakes the chosen number of 500 replicates of the Monte Carlo simulation sufficient.

Tables 4 and 5 show that in the probability sample of the first experiment, Lowe and Fisher indices present the lowest levels of bias in each consumption segment, while opposite behavior can be seen for the Jevons index; besides, slightly higher relative sampling errors are found for Fischer index and lower for Lowe and Jevons indices in all consumption segments.

Table 4 First experimental phase: Absolute Relative Bias and Relative Sampling Error distributions of monthly Lowe, Fisher and Jevons indices for coffee consumption segment (Sampford sampling)

Index		Coffee consumption segment				
		Min	Q1	Me	Q3	Max
Lowe	ARB	– 0.0005	– 0.0014	– 0.0013	– 0.0008	– 0.0017
	RSE	1.32	1.38	1.41	1.43	1.52
Fisher	ARB	0.0005	0.0015	0.0021	0.0030	0.0039
	RSE	1.14	1.36	1.53	1.73	1.98
Jevons	ARB	– 0.0199	– 0.0485	– 0.0400	– 0.0353	– 0.0609
	RSE	1.05	1.09	1.14	1.20	1.24

Table 5 First experimental phase: Absolute Relative Bias and Relative Sampling Error distributions of monthly Lowe, Fisher and Jevons indices for pasta consumption segment (Sampford sampling)

Index		Pasta consumption segment				
		Min	Q1	Me	Q3	Max
Lowe	ARB	0.0008	– 0.0003	– 0.0001	0.0005	– 0.0007
	RSE	1.13	1.20	1.28	1.32	1.38
Fisher	ARB	– 0.0009	– 0.0043	– 0.0034	– 0.0025	– 0.0051
	RSE	1.21	1.61	1.67	1.83	2.06
Jevons	ARB	– 0.0017	0.0031	0.0105	0.0209	0.0314
	RSE	0.85	0.92	1.04	1.10	1.22

Table 6 Second experimental phase: Absolute Relative Bias distribution of monthly Lowe, Fisher and Jevons indices for coffee consumption segment (two stage sampling and one stage-cluster sampling design)

Index	Sampling	Coffee consumption segment				
		Min	Q1	Me	Q3	Max
Lowe	1stage	– 0.0015	– 0.0005	0.0002	0.0011	0.0029
	2stage	– 0.0010	– 0.0007	– 0.0001	0.0000	0.0007
Fisher	1stage	– 0.0011	– 0.0004	0.0004	0.0014	0.0034
	2stage	0.0004	– 0.0010	– 0.0006	– 0.0001	– 0.0014
Jevons	1stage	– 0.0002	– 0.0000	0.0001	0.0002	0.0010
	2stage	0.0003	– 0.0004	– 0.0002	– 0.0000	– 0.0010

Table 7 Second experimental phase: Relative Sampling Error distribution of monthly Lowe, Fisher and Jevons indices for coffee consumption segment (two stage sampling and one stage-cluster sampling design)

Index	Sampling	Coffee consumption segment				
		Min	Q1	Me	Q3	Max
Lowe	1stage	0.94	1.17	1.46	1.61	2.03
	2stage	1.16	1.28	1.40	1.58	1.69
Fisher	1stage	0.76	1.00	1.09	1.23	1.59
	2stage	1.15	1.33	1.42	1.55	1.71
Jevons	1stage	0.47	0.54	0.57	0.61	1.04
	2stage	0.50	0.58	0.61	0.67	0.79

Tables 6 and 7, referred to the second phase, show that all indices present low levels of bias, but higher relative sampling errors for Fischer and Lowe indices respect to Jevons index under both sampling designs. However, the performance of price indices is different in the two sampling designs both in terms of bias and relative sampling error.

Figure 1, from the first experimental phase, shows the level estimates of the monthly Jevons, Lowe and Fisher indices computed on probability and nonprobability samples and the true value (universe panel series SD) of the corresponding index for two consumption segments (coffee and pasta in Turin province). The number of panel series considered are 23'636 for pasta segment and 9'608 for coffee segment, with a coverage turnover respectively equal to 76.2 and 79.8 percent.

The comparison between probability and nonprobability selection schemes shows a common evidence for both products: pps sample estimates of weighted index, Fisher and Lowe, results quite overlapped to the "true" value U; cut-off estimates over-estimate, but follow the trend for coffee, while for pasta are quite overlapped to true value U. Most sold item estimates under-estimate and alter trend for coffee with weighted indices but not for Jevons, while for pasta they show different trends for the three indices. The mean of sample estimates of Jevons index strongly over-

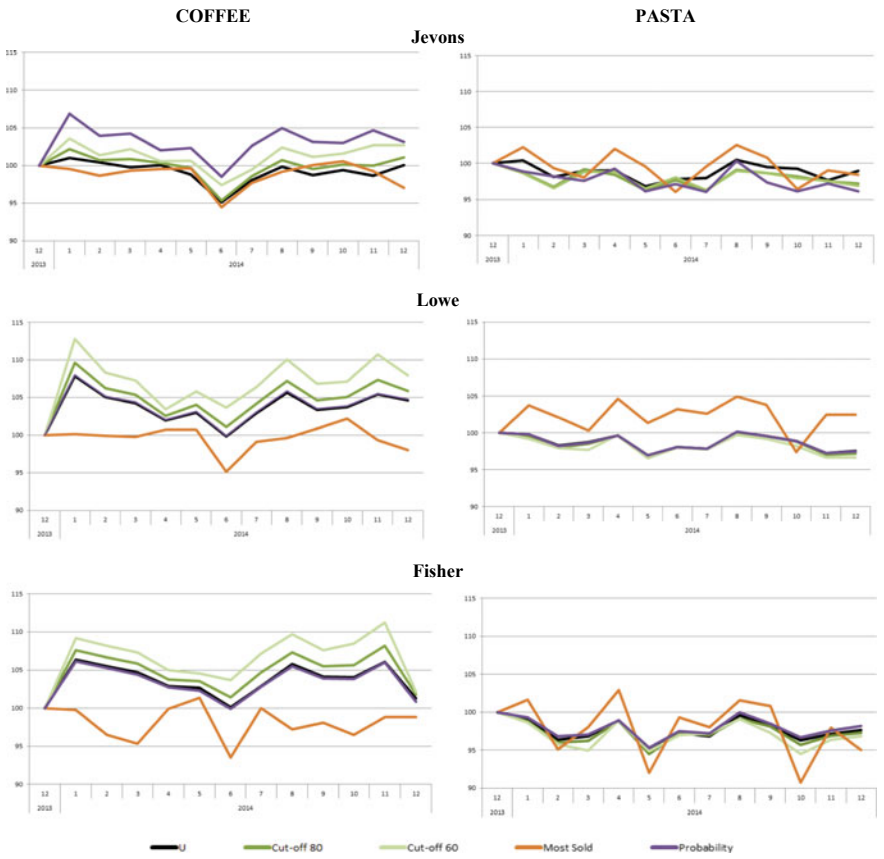


Fig. 1 Jevons, Lowe and Fisher indices computed with different selection schemes of series for coffee and pasta segments, Turin province, year 2014

estimates the "true" value U for coffee but not for pasta. These opposite performance for the two products can be explained by the different number of items and turnover distributions. In general, also from other evidences not shown for sake of brevity, (i) probability sampling always produces more accurate estimates than nonprobability selection scheme; (ii) sampling scheme is not neutral with respect to the choice of aggregation formulas; (iii) sampling error varies among consumption segments.

Figure 2, from the second experimental phase, illustrates the difference among the three indices estimated under two different sampling designs: cluster sample of outlets (with proportional allocation and PPS selection) versus two stage sample (proportional allocation and PPS selection of outlets and Neyman allocation and PPS selection of EANs).

The comparison between two probability selection schemes highlights that all the estimates seems to catch properly the level and the trend of the related true index. The estimator of Lowe and Fisher indices have in both cases wider confidence intervals

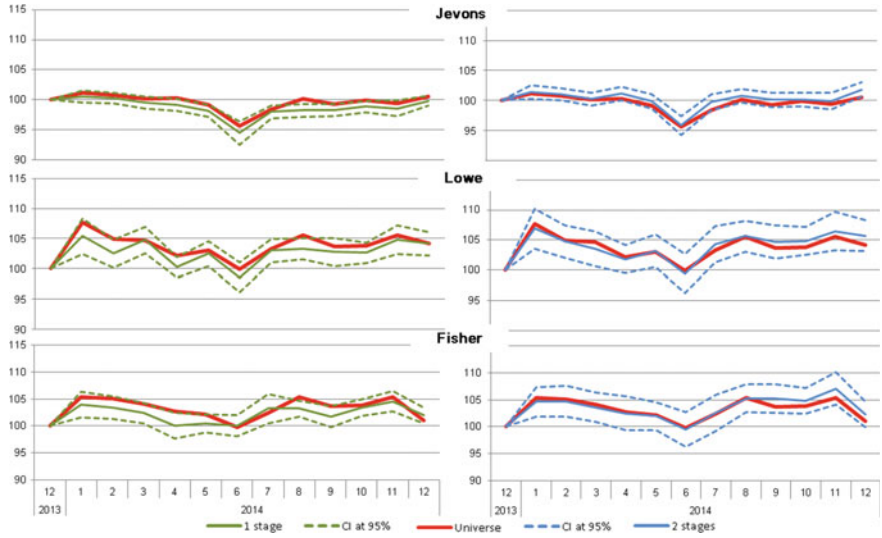


Fig. 2 Jevons, Lowe and Fisher indices for coffee segment estimated on one sample, confidence interval (CI) of estimates at 95% and true value (computed on the universe of SD). Turin, year 2014

(CI) with respect to the Jevons index, due to the variability if quantities involved in the weights. In general the width of CIs are greater under the two stage sampling than under cluster sampling design (one stage), even if the difference does not seem so large.

5 Concluding Remarks and Future Developments

The two experimental phases produced interesting results regarding the performance of sampling schemes and index formulas in a closed population context and fixed approach. They lead to the conclusion that probability sampling is the better choice in this context.

The successive phase, currently in progress, regards the comparison between a fixed and a dynamic approach, the latter consisting in considering all series of an open population [10]. The elementary price indices are computed considering both closed and open population. When assuming a closed population, direct indices are built on a fixed basket of products defined at reference time, ignoring new products (fixed approach). In this context the indices are affected by shrinkage over time due to the attrition of products during the year. However, in reality many products disappear and new products enter continuously. On the other hand, using chain indices the life cycle of products is taken into account as the basket of products changes months by months: the flexible basket is constituted by the matching products sold

during two months in a row (dynamic approach). In order to evaluate the impact of the life cycle of products, direct and chain price indices are compared. For this purpose, an artificial population is generated, with appearing and disappearing products (momentarily and permanently). Starting from a panel of products, new products are introduced considering the monthly birth rates and old products have been removed in accordance to a survival function (both monthly birth and survival rates have been estimated on the real open population).

The outlined new experimental phase will provide evidences on the pros and cons of the two approaches, highlighting in particular empirical and theoretical drawbacks of the dynamic approach which is the one that ISTAT is oriented to choose for the future.

References

1. Bernardini, A., De Vitiis, C., Guandalini, A., Inglese, F., Terribili, M.D.: Measuring inflation through different sampling designs implemented on scanner data. Paper presented at the UNECE meeting of the group of experts, Geneve 2–4 May (2016)
2. Chessa, A.G., Verburg, J., Willenborg, L.: A Comparison of Price Index Methods for Scanner Data (2017)
3. de Haan, J., Opperdoes, E., Schut, C.M.: Item selection in the Consumer Price Index: Cut-off versus probability sampling. Surv. Methodol. **25**(1), 31–41 (1999)
4. de Haan, J., van der Grient, H.A.: Eliminating chain drift in price indexes based on scanner data. J. Econ. **161**, 36–46 (2011)
5. de Haan, J., Willemborg, L. and Chessa, A. G. An overview of price index methods for scanner data (2016)
6. De Vitiis, C., Casciano, M.C., Guandalini, A., Inglese, F., Seri, G., Terribili, M.D., Tiero, F.: Sampling design issues in the first Italian experience on scanner data. Paper presented at the Scanner Data Workshop. Rome 1–2 October (2015)
7. Feldmann, B.: Scanner-data-current-practice (2015). http://www.istat.it/en/files/2015/09/5-WS-Scanner-data-Rome-1-2-Oct_Feldmann-Scanner-data-current-pratice.pdf
8. Gábor, E., Vermeulen, P.: New evidence in elementary index bias. Working Paper, Working Paper, European Central Bank (2014)
9. ILO, IMF, OECD, Eurostat, United Nations, World Bank Consumer Price Index Manual: Theory and Practice. ILO Publications, Geneva (2004)
10. Ivancic, L., Diewert, W.E., Fox, K.J.: Scanner data, time aggregation and the construction of price indexes. J. Econ. **161**(1), 24–35 (2011)
11. Nygaard, R.: Chain drift in a monthly chained superlative price index. Workshop on Scanner Data, Geneva, 10 May (2010)
12. Norberg, A.: Sampling of scanner data products offers in the Swedish CPI. Draft version 8 – Statistics. Sweden (2014)
13. Rosén, B.: On sampling with probability proportional to size. J. Stat. Plan. Inference **62**(2), 159–191 (1997)
14. Sampford, M.R.: On sampling without replacement with unequal probabilities of selection. Biometrika **54**(3–4), 499–513 (1967)
15. van der Grient, H.A., de Haan, J.: The Use of Supermarket Scanner Data in the Dutch CPI. Paper presented at the Joint ECE/ILO Workshop on Scanner Data, Eurostat (2010)
16. Vermeulen, B.C., Herren, H.M.: Rents in Switzerland: Sampling and Quality Adjustment. 11th Meeting - Ottawa Group - Neuchâtel 27–29 May (2006)

An Investigation of Hierarchical and Empirical Bayesian Small Area Predictors Under Measurement Error

Silvia Polettini and Serena Arima

Abstract In this paper we focus on small area models with measurement error in covariates. Based on data from the Measuring Morality study, a nationally representative survey of United States residents, that contains a validated behavioural measure of generosity (the dictator game) along with the household income of respondents, we define a measurement error model suitable to obtain area-level estimates of generosity at the district level. We investigate the effect of introducing the measurement error in this model, focusing on fully Bayesian as well as Empirical Bayesian (EB) estimation proposed in the recent literature. We discuss the characteristics of each of the two approaches and analyze the impact of the measurement error on the resulting estimates based on real data and a simulation study.

Keywords Small area estimation · Measurement error · Misclassification · Empirical Bayesian estimators · Hierarchial Bayesian estimators

1 Introduction

Small area estimation has emerged in recent years as an important area of statistics as private and public agencies try to extract the maximum information from sample survey data. Sample surveys are generally designed to provide estimates of totals and means of variables of interest for large subpopulations or domains. However, governments are more and more interested in obtaining statistical summaries for smaller domains such as states, provinces, or different racial and/or ethnic subgroups. These domains are called small areas. In recent years, demand for reliable estimates

S. Polettini (✉)
Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma, P.le Aldo Moro, 5, 00185 Rome, Italy
e-mail: silvia.polettini@uniroma1.it

S. Arima
Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, Via del Castro Laurenziano, 9, 00185 Rome, Italy
e-mail: serena.arima@uniroma1.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_17

for small area means has greatly increased due to their growing use in formulating policies and programmes, allocating government funds, regional planning and other uses.

Small area estimation tackles the problem of providing reliable estimates of one or several variables of interest in areas where the information available on those variables is, on its own, not sufficient to provide accurate *direct estimates* using the domain-specific sample data [10, 11].

Indirect estimators are often employed in order to increase the effective domain sample size by borrowing strength from the related areas using linking models, census, administrative data and other auxiliary variables associated with the small areas. Estimates for all areas are produced using the sample and the auxiliary information which should be available for all small areas. A comprehensive account of model-based small area estimation is given in [11].

Area level models relate the small area means to area-specific auxiliary variables and become essential if unit level data are not available. When unit-level auxiliary information is available, nested error linear regression models are often used to obtain efficient model-based estimators of small area means. However, it is often the case that the auxiliary information is subject to measurement error. In such circumstances, it is natural to consider the small area estimation problem under a measurement error approach. This topic has been widely documented in the literature: see among the others [3, 9, 10, 12, 14] and references therein. In particular, [3, 9, 10, 12] consider the unit level regression model for small area estimation when the area level covariate is subject to functional measurement error. We consider the same problem, focusing on both Empirical Bayesian (EB) and fully Bayesian (Hierarchical Bayesian, HB) small area estimators that have been recently proposed. After briefly reviewing the literature in Sect. 2, we propose a small area model accounting for measurement error in both continuous and discrete covariates (Sect. 3), testing it on a real data application, namely the generosity data. Such dataset has been obtained in [6] by coupling Gini inequality indices derived from the American Community Survey with data from the Measuring Morality study. The latter is a nationally representative survey of United States residents, that contains a validated behavioural measure of generosity (the dictator game) along with the household income of respondents. The data were collected under the supervision of Stephen Vaisey, Duke University and can be downloaded from the Association of Religion Data Archives, www.TheARDA.com. Based on these data, as well as a follow-up experiment, [6] test the relationship between economic inequality, income, and generosity. The authors identify a previously undocumented effect of economic inequality, namely that higher income individuals in the US tend to be less generous than poorer individuals, *but only in contexts where macro-level economic inequality is high*, or is perceived as high. Noticing that both income and the Gini index are subject to measurement error (indeed income is self reported and the Gini index is estimated from another survey), we allow for measurement error in the covariates, fitting a model that generalizes the one used by the previously mentioned Authors. We do not perform a point-by-point comparison with [6] since the models differ in several respects: besides the inferential approach, the income variable, treated in [6] as continuous covariate, is

coherently modelled as categorical since in the survey it has been recorded as a 19 classes variable. As stressed in the literature, ignoring the measurement error in the covariates may lead to inconsistent estimates and can severely invalidate inferences, see e.g. [5]. Our aim is to investigate the latter aspect in the same application described in [6].

In Sect. 4 we focus on estimation of the regression model parameters; we assess the performance of the proposed measurement error model in our application and compare the estimates obtained accounting for measurement error and ignoring it.

In Sect. 5 we focus on different small area mean estimators proposed in literature, namely the direct estimator, the empirical Bayes estimator and the posterior means. We compute such estimators for the generosity data and notice that, although the resulting estimates are very similar, some theoretical considerations lead us to investigate their performance in a controlled setting. In Sect. 6 we perform a simulation study and discuss potentialities and limitations of the aforementioned small area estimators in different scenarios. Simulation results show that the posterior means are flexible and efficient tools for small area estimators. We conclude with a brief discussion in Sect. 7.

2 Measurement Error in Small Area Model

We consider a Bayesian formulation of the unit-level nested error linear regression model where the measurement error in auxiliary variables is explicitly modelled. Our approach is analogous to [9, 10], who were the first to consider the problem of measurement error in small area models for unit-level data. We rely on a superpopulation approach and specify a Bayesian hierarchical model, that is assumed to hold for the whole population as well as for the sample data, e.g. under the hypothesis of no selection bias. In the above mentioned papers, a single continuous area-level covariate, subject to measurement error, is introduced.

Suppose there are m areas and let N_i be the known population size of area i. We denote by Y_{ij}^P the response in the population of the jth unit in the ith area ($i = 1, \ldots, m$; $j = 1, \ldots, N_i$). A random sample of size n_i is drawn from the ith area. The goal is to predict the small area means $\Gamma_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}^P$, $i = 1, \ldots, m$, based on the available data. Adopting a superpopulation approach to finite population sampling, [9, 10] model the response variable Y as

$$Y_{ij}^P = \alpha + \beta x_i + u_i + \epsilon_{ij} \qquad i = 1, \ldots, m; \quad j = 1, \ldots, N_i \tag{1}$$

where x_i is an auxiliary variable observed for each area. ϵ_{ij} and u_i are assumed independent, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$ and $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$. To measure the true area-level covariate it is assumed that there are N_i units in the ith small area and that a random sample of size n_i is taken from the ith area, resulting in data X_{il} ($l = 1, \ldots, r_i$; $i = 1, \ldots, m$). For the sample, the measurement error model

$$X_{il} = x_i + \eta_{il}, \qquad \eta_{ij} \overset{iid}{\sim} N(0, \sigma^2_\eta) \qquad i = 1, \ldots, m; \qquad l = 1, \ldots, n_i \tag{2}$$

is assumed. Furthermore, ϵ_{ij}, u_i and η_{ij} are taken mutually independent. Reference [10] also assumed that $x_i \overset{iid}{\sim} N(\mu_x, \sigma^2_x)$, defining the structural measurement error model.

The aforementioned literature considers the case in which the measurement error only affects continuous variables, according to the measurement error model of Eq. (1). For discrete covariates, measurement error means misclassification. To allow for auxiliary discrete covariates measured with error, as in [2] we model the misclassification mechanism through an unknown transition matrix P and estimate all the unknown parameters in a fully Bayesian framework. The details of the model are described in the next section.

3 A Measurement Error Small Area Model for Both Continuous and Discrete Covariates

Following [2], for each unit in each area, we consider the following covariates: T_{ij}—the vector of p continuous or discrete covariates measured without error, w_i and x_{ij}—respectively, a vector of q continuous covariates and h discrete variables (with a total of K categories), both measured with error. Denote by S_{ij} and Z_{ij} the observed values of the latent w_i and x_{ij}, respectively. Without loss of generality, in what follows we assume $h = 1$.

Following the notation in [10], the proposed measurement error model can be written in the usual multi-stage way: for $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ and for $k, k' = 1, \ldots, K$

Stage 1. $Y_{ij} = \theta_{ij} + e_{ij} \quad e_{ij} \overset{iid}{\sim} N(0, \sigma^2_e)$

Stage 2. $\theta_{ij} = T'_{ij}\delta + w'_i\gamma + \sum_{k=1}^{K} I(x_{ij} = k)\beta_k + u_i \quad u_i \overset{iid}{\sim} N(0, \sigma^2_u)$

Stage 3. $S_{ij}|w_i \overset{iid}{\sim} N(w_i, \Sigma_s = diag(\sigma^2_{s_1}, \ldots, \sigma^2_{s_q}))$
$w_i \overset{iid}{\sim} N(\mu_w, \Sigma_w = diag(\sigma^2_{w_1}, \ldots, \sigma^2_{w_q}))$
$Pr(Z_{ij} = k|x_{ij} = k') = p_{k'k}$
$p_{k'\cdot} = (p_{k'1}, \ldots, p_{k'K}) \sim Dir(\alpha_{k',1}, \ldots, \alpha_{k',K})$
$Pr(x_{ij} = k') = \frac{1}{K}$

Stage 4. $\beta, \delta, \gamma, \sigma^2_e, \sigma^2_u, \sigma^2_{s_1}, \ldots, \sigma^2_{s_q}, \mu_w, \sigma^2_{w_1}, \ldots, \sigma^2_{w_q}$ are, loosely speaking, a-priori mutually independent.

Stage 3 defines the measurement error model for both continuous and discrete covariates. For the discrete covariates, the misclassification mechanism is specified according to the $K \times K$ matrix P, whose (k', k) element, $p_{k'k}$, denotes the probability that the observable variable Z_{ij} takes the kth category when the true unobservable variable x_{ij} takes the k'th category. We also assume that the misclas-

sification probabilities are the same across subjects and that all the categories have the same prior probability $\frac{1}{K}$ to occur. Over each row of P , we place a Dirichlet $Dir(\alpha_{k',1}, \ldots, \alpha_{k',K})$ prior distribution, with known $\alpha_{k',1}, \ldots, \alpha_{k',K}$. In Stage 4 we assume Normal priors for β, δ, γ and μ_w and inverse gamma distributions for all variances. Depending on the specific application and on the availability of prior information, one may fix some of these parameters (see the application). Hyperparameters have been chosen to have flat priors whose sensitivity has been widely investigated in [10]. For the Dirichlet distribution we specified a Perks' prior discussed in [1] as a default noninformative prior; robustness of model estimates with respect to such specification has been investigated in [2].

According to the above assumptions, we can estimate all model parameters including the transition matrix P. As the posterior distribution cannot be derived analytically in closed form, we obtain samples from the posterior distribution using Gibbs sampling.

4 Inference on Regression Coefficients Under Measurement Error: Application to the Generosity Data

In this section we fit the unit level small area model with measurement error in covariates defined in the previous section to the generosity data. First we consider the performance of the proposed model in estimating the regression parameters; indeed the mixed effects model also allows us to evaluate the relationship between economic inequality, income and generosity. In the next section we also consider estimation of the mean generosity score at the area level.

There is an increasing interest in understanding the implications of income for behaviour, in particular generosity toward others. Well grounded literature on this topic has portrayed a picture of higher-income individuals as consistently more selfish than poorer individuals [13]. A different perspective is reported in a recent paper [6], where the relationship between economic inequality, income, and generosity is tested on the generosity data. Fitting a linear mixed effects model, the authors identify a previously undocumented effect of economic inequality, namely that higher income individuals in the US tend to be less generous than poorer individuals, *but only in contexts where macro-level economic inequality is high*, or is perceived as high. The Authors comment that the results obtained challenge the prevailing view in the literature that higher income individuals are necessarily less generous and conclude that "inequitable resource distributions undermine collective welfare" and that redistributive policies may "attenuate, or even reverse, the negative relationship between income and generosity, in turn increasing the generosity of those individuals who have the most to give".

Data from the Measuring Morality study comprise of 1498 respondents in the US. For each respondent, income and some personal and demographic variables (such as age, gender, education, …) have been collected. Respondents completed a validated

behavioural measure of generosity: the dictator game [8]. Respondents learned that they had been randomly assigned the role of *decider* and had received 10 tickets, each worth one entry in a raffle to win a monetary prize of either 10 or 500. They could transfer any number of tickets to the next participant, a *receiver* who did not have any tickets. By giving tickets, respondents could benefit another person at a cost to themselves in a zero-sum opportunity to win money. This measure of generosity was administered to individuals with different incomes residing in areas (US states plus the District of Columbia) that vary in levels of inequality, measured according to the Gini's coefficient. The number of respondents in each area ($m = 9$ divisions) ranges from 72 to 286. In the proposed model we take generosity as the response variable and income, standardized Gini coefficients and their interaction as auxiliary variables. According to the survey design, household income was collected as a 19-classes variable; for ease of interpretation in the application we recoded it into five classes (C_1 : less than 12500; C_2 : [12500, 30000), C_3 : (30000, 60000], C_4 : (60000, 125000], C_5 : over 125000). Since income is self reported and the Gini index is estimated using data from the 2012 American Community survey, we can suspect that both auxiliary variables are subject to measurement error. In order to evaluate the impact of accounting for this source of error, we fit both the standard model that ignores the measurement error and the model proposed in Sect. 3. Figure 1 shows the posterior distribution of the model parameters. Since the American Community Survey (ACS) provides reliable estimates of the sampling error of the Gini index, we fix σ_s^2 equal to 0.01.

The left panel reports the posterior distribution of the regression parameters under the proposed measurement error model: income is the only factor that significantly impacts on the response variable, since for all the other parameters the 95% credible intervals contain the zero value (CI_{Gini} : $[-0.207, 0.349]$, $CI_{C1*Gini}$: $[-0.632, 0.241]$, $CI_{C2*Gini}$: $[-0.542, 0.217]$, $CI_{C3*Gini}$: $[-0.533, 0.189]$), with

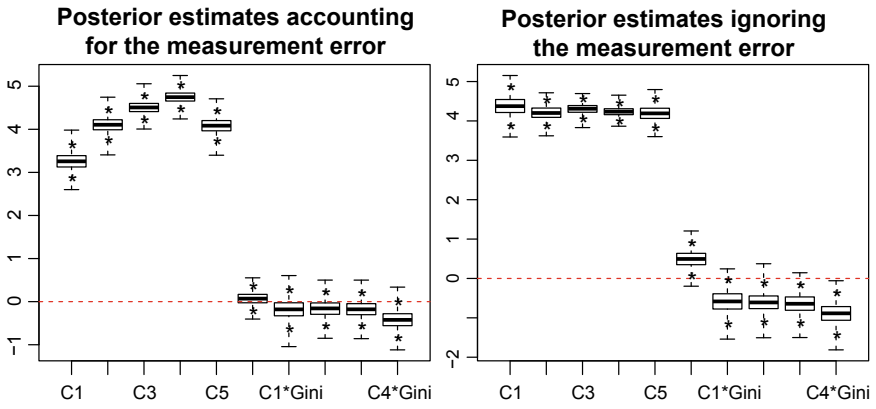


Fig. 1 Posterior distribution of the model parameters. **Left panel**: posterior distributions obtained from the proposed model. **Right panel**: posterior distributions from the model that ignores the measurement error. Stars denote the bounds of the corresponding 95% credible intervals

the exception of the negative interaction between the Gini index and the fourth income class ($CI_{C4*Gini} : [-0.827, -0.028]$). However, it is worth nothing that the size of the effect for this last interaction is very small. At the same time the comparison with the no measurement error model shows a different situation: all interactions are significant and comparable in size, implying a shrinkage effect of the predicted generosity towards the mean level (see Fig. 1).

With respect to the income, it is apparent that generosity increases with income, with the exception of the last class, in which the effect on generosity is comparable to that of the second one. This actually means that the richest are less generous with respect to the others, which is line with findings in the mainstream literature on the subject. On the other hand, when one ignores the measurement error, all the covariates and their interactions seem to be significant (Fig. 1, right panel). In particular, income exhibits a positive effect on generosity, with no distinctions between income classes, which contradicts the economic theories; moreover, an unexpectedly positive effect of inequality is found. With respect to the measurement error for income, the posterior distribution of the original covariate given the data, $P(x = k|Z = 1, data),\ k = 1, \ldots, K$ is about 0.5 for $k = 1$ and almost uniformly distributed over the remaining categories. This is an empirical evidence that income is often underreported by the respondents. Focusing on the other categories of Z, the posterior distributions $P(x = k|Z = j, data),\ k = 1, \ldots, K, j = 2, \ldots, K$ are concentrated at the corresponding categories of X, with $P(x = j|Z = j, data) \approx 0.9$ for $j = 2, \ldots, K$, the corresponding credibile intervals not containing 1. This is an indication that measurement error has a significant impact on income.

5 Small Area Means Estimators

In small area problems, the ultimate goal of the analysis is to predict the small area means

$$\Gamma_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij} \tag{3}$$

given the available information. Under the model described in Sect. 3,

$$Y_{ij}^P = \sum_{k=1}^{K} \beta_k I(x_{ij} = k) + \gamma w_i + \delta T_{ij} + u_i + \epsilon_{ij},$$

so that the expected values of the variable of interest given $\Phi = (\beta, \gamma, \delta, u_i)$ and the variables (x_{ij}, w_i, T_{ij}) for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$ can be written as

$$\theta_{ij} = \sum_{k=1}^{K} \beta_k I(x_{ij} = k) + \gamma w_i + \delta T_{ij} + u_i.$$

Under the hypothesis of no selection bias and assuming that the auxiliary information is available for each area, prediction of the small area means Γ_i can be based on prediction of the mixed effects. Indeed

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\theta_{ij} = \sum_{k=1}^{K}\beta_k F_{ik} + \gamma w_i + \delta \bar{T}_i + u_i \tag{4}$$

where $F_{ik} = {N_i}^{-1}\sum_{j=1}^{N_i} I(x_{ij} = k)$ are the relative frequencies of the kth category of the variable x for the ith area in the population, and $\bar{T}_i = {N_i}^{-1}\sum_{j=1}^{N_i} T_{ij}$ are the means of the auxiliary variable T for the ith area in the population.

As underlined e.g. in [11] (Sect. 7.1.1, p. 174) and [7], (4) can be used to predict the small area means Γ_i given the available information. Although prediction of Γ_i does not exactly correspond to predicting μ_i, as in fact $\Gamma_i = \mu_i + \bar{e}_i$ with $\bar{e}_i = {N_i}^{-1}\sum_{j=1}^{N_i} e_{ij}$, when N_i is large, the predictor of the mixed effects μ_i can be considered an appropriate predictor of Γ_i in (3).

Reference [10] considered both an Empirical Bayesian and a Hierarchical Bayesian approach to derive predictors of small area means, assuming the small area measurement error model encompasses a single covariate, measured with error as detailed in (1) and (2). Under their empirical Bayes approach, they first derived a predictor for the vector of $N_i - n_i$ units, conditional on the unknown parameters and the observed sample, denoted as Y. In particular, adopting the notation in [10], for any unsampled Y_{ij}^{US}, $j = n_i + 1, \ldots, N_i$, they obtained

$$E[Y_{ij}^{US}|Y, \beta, \alpha, \sigma_e^2, \sigma_u^2, \mu_x, \sigma_x^2, \sigma_\eta^2] = (1 - f_i B_i)\bar{Y}_i + f_i B_i(\alpha + \beta\mu_x) \tag{5}$$

where $B_i = \sigma_e^2/[\sigma_e^2 + n_i(\sigma_u^2 + \beta^2\sigma_x^2)]$ and $f_i = (N_i - n_i)/N_i$ is the finite population correction fraction. The empirical Bayes predictor is obtained by replacing the unknown model parameters with their estimators.

Reference [12] extended the approach in [10] including sample information on the covariate values and derive an empirical Bayes predictor, showing its asymptotical optimality.

Reference [10] also proposed a fully Bayesian approach; they define a hierarchical model based on Eqs. (1) and (2), specify vague prior distributions for all the model parameters, and estimate posterior distributions via Gibbs sampling. References [3, 4] extended the above approach, proposing to use the Jeffreys' prior on the model parameters.

Expression (5), upon which the EB estimator proposed in [10] is based, shows the usual structure of a convex combination between the direct estimator (sample mean) and a synthetic estimator based on the regression model. To account for the measurement error, the weights in this combination also depend on the measurement error variance. More specifically, B_i is inversely related to $n_i(\frac{\sigma_u^2}{\sigma_e^2} + \beta^2\frac{\sigma_x^2}{\sigma_e^2})$, which makes B_i very sensitive to the area size, to β and to the measurement error. As a consequence, unless the latter quantities are all very small, the estimator proposed by Ghosh et al.

[10] quickly goes to the direct estimator, which in practice makes the measurement error approach not worthwhile under EB prediction (or vice-versa). Although the measurement error model has been proven to provide a better fit to the data and improved model parameter estimates, when considering the small area predictions, a severe measurement error leads the EB predictor to abandon the (corrected) model in favour of the sample mean.

In our application, we consider a model with both continuous and discrete covariates measured with error. We extend the EB predictor proposed in [10] by including the categorical missclassified covariates in the expression (5). Also, we obtain fully Bayesian, HB, small area predictors by integrating the distribution of μ_i with respect to the posterior predictive distribution of $\Phi = (\beta, \gamma, \delta, u_i, x, w)$ given the sample data, the other model parameters and the population means of the auxiliary variables measured without error. This is easily accomplished under MCMC simulation schemes. In fact, we use the measurement error model to predict the distribution of the covariates x and w.

Table 1 reports the small area estimates produced under the model with and without measurement error for the generosity data, along with the direct estimator. As mentioned above, we consider both the extension of the EB estimator proposed in [10] and the HB estimator obtained under a fully Bayesian analysis of our measurement error model. From the table we can see that allowing for measurement error in both continuous and categorical covariates impacts on estimation of the small area means more in terms of uncertainty than in point estimates However, due to the large area sizes in this applications, it is an expected result that there is little difference between the fully Bayesian HB estimator, the EB estimator just mentioned, and the area means.

In our application we have therefore found a strong impact of measurement error on regression coefficients, but not such a large impact on small area predictions; this can be ascribed to the large area sizes in the real data problem, which prevents us from making a fair comparison between estimators. To be able to test the effect

Table 1 Small area estimates: posterior means of the small area means obtained with the model that does not account for the measurement error ($\hat{\theta}^{HB}_{NoErr}$) and the model that accounts for it ($\hat{\theta}^{HB}_{Err}$, second row). Direct estimator $\hat{\theta}^{D}$ and the empirical Bayes estimator in [10] ($\hat{\theta}^{EB}$) are reported in the third and last row. Standard deviations in brackets

Division	1	2	3	4	5	6	7	8	9
$\hat{\theta}^{HB}_{NoErr}$	4.17 (0.27)	4.11 (0.33)	4.25 (0.18)	4.44 (0.20)	4.19 (0.24)	4.28 (0.10)	4.25 (0.14)	4.37 (0.16)	4.22 (0.23)
$\hat{\theta}^{HB}_{Err}$	4.27 (0.36)	4.09 (0.41)	4.26 (0.38)	4.43 (0.37)	4.17 (0.40)	4.30 (0.33)	4.25 (0.34)	4.38 (0.32)	4.23 (0.40)
$\hat{\theta}^{D}$	4.23 (2.29)	4.05 (2.41)	4.08 (2.37)	4.52 (2.63)	4.10 (2.69)	4.61 (2.54)	4.40 (2.38)	4.40 (2.28)	4.28 (2.54)
$\hat{\theta}^{EB}$	4.23 (1.32)	4.05 (0.81)	4.08 (0.94)	4.52 (0.68)	4.11 (0.45)	4.61 (0.79)	4.40 (0.84)	4.40 (0.35)	4.28 (0.76)

of the measurement error in small area prediction in a controlled setting, we next consider a simulation study in which the area sizes are smaller, as typical in small area problems. We compare the EB and HB predictors in this framework to better understand their behaviour and each other's merits.

6 Simulation Study

In order to compare the performance of different small area mean estimators we perform a simulation study. We compare the direct estimator with the empirical Bayes estimator in [10] and the posterior means obtained under the measurement error model. To this end, we create a finite population of size 140000 spread across 12 strata of sizes 5000 25000 5000 10000 20000 15000 5000 15000 10000 15000 10000 and 5000. The responses Y_{ij} are generated under a superpopulation model with two continuous covariates measured with error. In particular, we set $\delta = 100$, $\gamma_1 = 2$, $\gamma_2 = 5$, $\sigma_e^2 = 100$, $\sigma_u^2 = 16$, $\sigma_{w_1}^2 = \sigma_{w_2}^2 = 25$, $\mu_{w_1} = 194$, $\mu_{w_2} = 120$ and $\sigma_{s_1}^2 = \sigma_{s_2}^2 = 2.7$. The settings are similar to the ones adopted in [10]. In order to investigate the behaviour of the aforementioned estimators with respect to the small area sample size, we select 0.5 and 0.05% simple random samples from each stratum. Accordingly, the sample size of the first random sample ranges from 25 to 125, while the sample size of the second sample ranges from 2 to 12. We draw 100 independent samples for each simulation scenario. To obtain the Hierarchical Bayesian estimators, we ran a Gibbs chain of size 15000 with a burn-in of the first 5000. The hyperparameters have been specified as discussed in Sect. 3. We then compute the direct estimators of the small area means, the empirical Bayesian estimators in [10] and the small area posterior means. As in the real data application, the estimator in [10] has been obtained by filling the posterior means of the model parameters in formula (5).

Table 2 shows the root mean squared errors (RMSEs) of the estimators we aim to compare in the two simulation scenarios. Notice that, when the sample size is quite large, although it is a small percentage with respect to the real population size, the three estimators perform very similarly. This simulation scenario somehow resembles the real data application in which, although the area sizes are very small compared to the population size, they are large enough to produce very similar estimators with all methods. Moreover, a different behaviour can be grasped when the sample size for each area decreases with a significant reduction of the RMSE of the posterior means.

Figure 2 shows the distribution of the empirical Bayes estimator and of the posterior means in the simulated datasets in the two scenarios: in the upper panel, we show the estimates when the small area size is about 0.5% of the population while in the lower one we show the estimates when the small area sample size is about 0.05% of the population size. Stars in the graphs denote the true small area means: as expected, the estimators are in agreement with themselves and with the true values

Table 2 Root mean squared errors (RMSE) of the direct estimator, the estimator in Ghosh et al. [10] and the posterior mean estimator for different sample sizes. Results are averaged over 100 simulated datasets

Area	n	Sample fraction: 0.5%			n	Sample fraction: 0.05%		
		$\hat{\theta}^D$	$\hat{\theta}^{EB}$	$\hat{\theta}^{HB}$		$\hat{\theta}^D$	$\hat{\theta}^{EB}$	$\hat{\theta}^{HB}$
1	25	1.76	1.72	1.72	2	7.98	7.41	7.34
2	125	0.86	0.85	0.85	12	2.93	2.95	2.86
3	25	2.02	2.00	2.00	2	7.13	6.35	6.96
4	50	1.42	1.40	1.41	5	4.90	4.62	4.64
5	100	0.97	0.96	0.96	10	3.14	2.98	2.98
6	75	1.24	1.24	1.23	8	3.41	3.29	3.14
7	25	1.85	1.82	1.84	2	7.29	6.49	6.49
8	75	1.08	1.08	1.07	8	3.36	3.23	3.22
9	50	1.46	1.45	1.45	5	4.01	3.83	3.75
10	75	1.26	1.25	1.25	8	3.69	3.57	3.47
11	50	1.31	1.31	1.30	5	3.73	3.60	3.46
12	25	1.93	1.89	1.89	2	5.96	5.05	5.26

when the sample size is large. When it decreases, the EB predictors perform worse than the posterior means, that are almost always centered around the true value.

As expected the variability of both estimators increases when the sample size decreases but the increase of variability of the empirical Bayes estimator does not always assure the coverage of the true value. Moreover, when the sample size decreases the estimator in [10] tends to be much closer to the direct estimator than the posterior means, as shown in Fig. 3.

7 Conclusions

Small area models are widely used in order to obtain estimates of unplanned domains from sample survey data. Model based small area estimation relies on the use of auxiliary variables: these variables are involved as covariates in regression models and allow to borrow strength from the related areas. The availability of such auxiliary variables is a key point of small area models. In this paper, we focus on small area models when auxiliary variables, both continuous and discrete, are measured with error. We show that the regression parameter estimates are biased if one does not account for measurement error and, indeed, inferences can be misleading. Moreover, we focus on small area mean predictors: an empirical Bayes small area mean estimator has been proposed in [10] and we compare it with the posterior means. Simulation studies and examples reported in the literature illustrate the performance of these estimators when just one covariate is affected by measurement error. In our

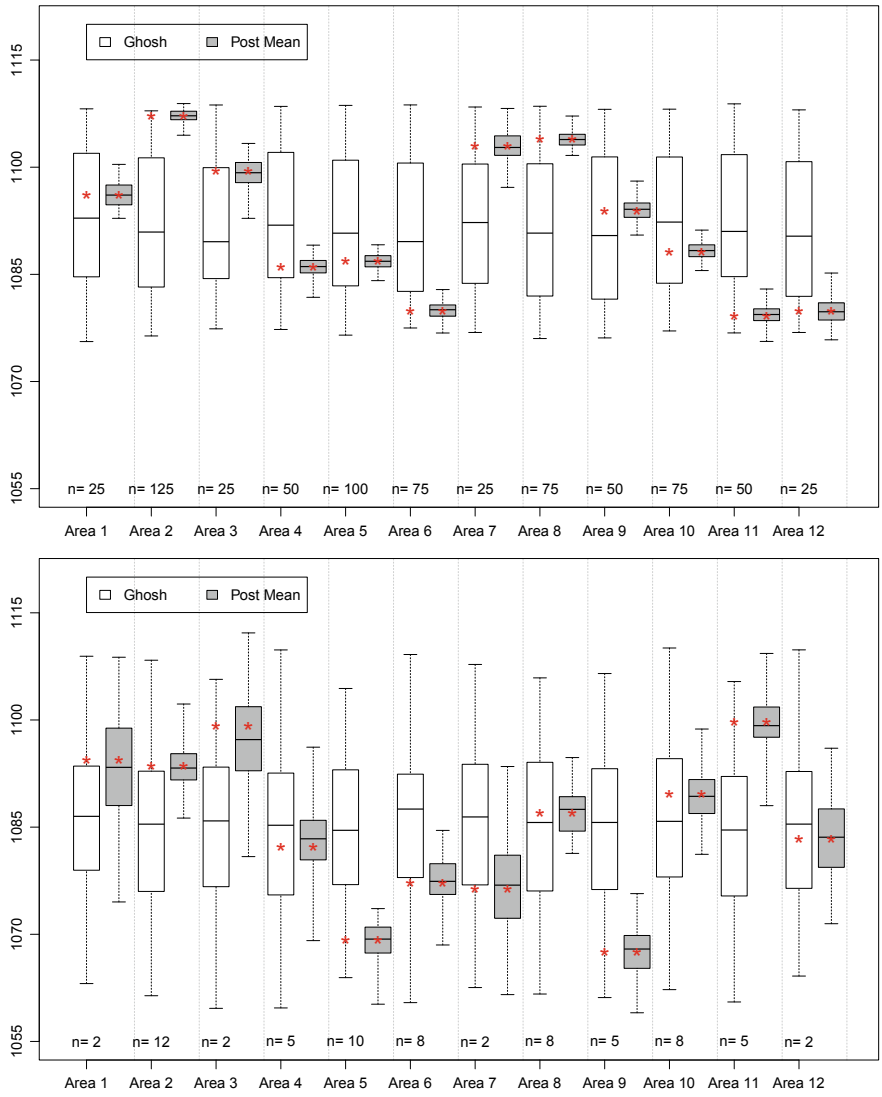


Fig. 2 Distribution of the Ghosh et al. [10] estimator and the posterior mean estimator of 100 datasets. **Upper panel**: estimates obtained when the small area sample sizes is 0.5% of the population. **Lower panel**: estimates obtained when the small area sample sizes is 0.05% of the population. Stars denote the true small area mean

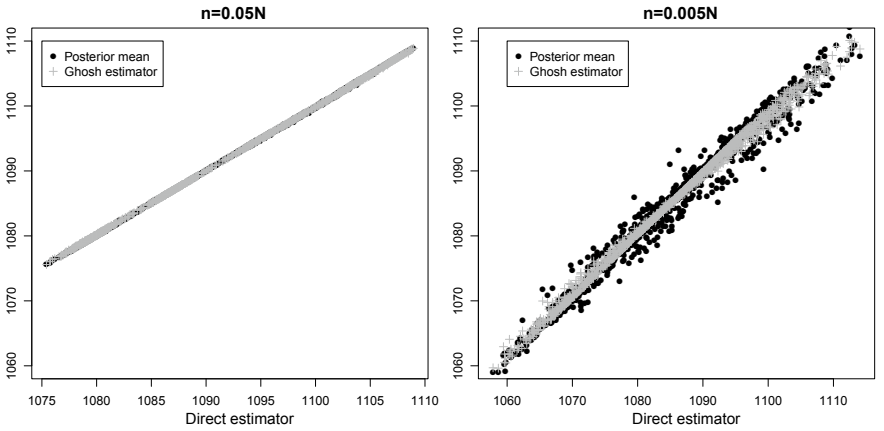


Fig. 3 Ghosh et al. [10] estimators and posterior mean estimators versus direct estimators of 100 datasets. **Left panel**: estimates obtained when the small area sample sizes is 0.5% of the population. **Right panel**: estimates obtained when the small area sample sizes is 0.05% of the population. Stars denote the true small area means

application, we consider two auxiliary variables, one of which is discrete. While the parameter estimates dramatically change when ignoring the measurement error, the small area predictions seem to be more robust. We argue that this can be ascribed to the large area sizes. Driven by analytical considerations, we empirically show through a simulation study that the EB and the HB predictors perform very similarly and also very similarly to the direct estimates, when the number of observations is large enough. However, the situation changes when the sample size decreases: the EB predictor is more similar to the direct one with a consequent increase in its variability. On the other hand, the posterior mean estimator seems to be more robust and in agreement with the true values.

References

1. Alvares, D., Armero, C., Forte, A.: What does objective mean in a Dirichlet multinomial process? Int. Stat. Rev. 106–118 (2017)
2. Arima, S., Polettini, S.: A unit-level small area model with misclassified covariates. J. R. Stat. Soc. A. (2019). https://doi.org/10.1111/rssa.12468
3. Arima, S., Datta, G.S., Liseo, B.: Objective Bayesian analysis of a measurement error small area model. Bayesian Anal. **72**(2), 363–384 (2012)
4. Arima, S., Datta, G.S., Liseo, B.: Models in small area estimation when covariates are measured with error. In: Pratesi, M. (ed.) Analysis of Poverty Data by Small Area Estimation, pp. 151–170. Wiley, New York (2016)
5. Carroll, R.J., Ruppert, D., Stefanski, L., Crainiceanu, C.: Measurement Error in Nonlinear Models: A Modern Perspective, 2nd edn. Chapman & Hall, CRC, Boca Raton (2006)
6. Côtè, S., House, J., Willer, R.: High economic inequality leads higher-income individuals to be less generous. Ann. PNAS **112**(52), 15838–15843 (2015)

7. Datta, G.S., Day, B., Maiti, T.: Multivariate Bayesian small area estimation: an application to survey and satellite data. Sankhyā: Indian J. Stat. Ser. A **60**(3), 344–362 (1998)
8. Engel, C.: Dictator games: a meta study. Exp. Econ. **14**(4), 583–610 (2011)
9. Ghosh, M., Sinha, K.: Empirical Bayes estimation in finite population sampling under functional measurement error models. J. Stat. Plan. Inference **137**, 2759–2773 (2007)
10. Ghosh, M., Sinha, K., Kim, D.: Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error model. Scand. J. Stat. **33**(3) (2006)
11. Rao, J.N.K., Molina, I.: Small Area Estimation, 2nd edn. Wiley, Hoboken (2015)
12. Torabi, M., Datta, G.S., Rao, J.N.K.: Empirical Bayes estimation of small area means under nested error linear regression model with measurement error in the covariates. Scand. J. Stat. **36**, 355–368 (2009)
13. Trautmann, S.T., van de Kuilen, G., Zeckhauser, R.J.: Social class and (un)ethical behavior: a framework, with evidence from a large population sample. Perspect. Psychol. Sci. **8**(5), 487–497 (2013)
14. Ybarra, L.M.R., Lohr, S.L.: Small area estimation when auxiliary information is measured with error. Biometrika **95**(4), 919–931 (2008)

Indicators for Monitoring the Survey Data Quality When Non-response or a Convenience Sample Occurs

Emilia Rocco

Abstract Non-response bias has long been a concern for surveys, even more so over the past decades with the increasing decline of the response rates. A similar problem concerns the surveys based on non-representative samples, the convenience and cost-effectiveness of which has increased with the recent technological innovations that allow for collecting large numbers of highly non-representative samples via online surveys. In both cases it must be assumed that the bias is the result of a self-selection process and, for both, quality indicators are needed to measure the impact of this process. The goal of this research is to show the opportunity in each survey of monitoring the risk of self-selection bias at two different level: at the level of the whole survey and at the level of each statistic of interest. The combined use of two indicators is suggested and empirically evaluated under various scenarios.

Keywords Auxiliary variables · Non-probability sampling · Non-response adjustment · Representativeness · Self-selection bias

1 Introduction

As response rates have declined over the past decades, the statistical benefits of probabilistic sampling have diminished. Assuming that a representative sample is initially selected, low response rates mean that those who ultimately supply the target data might not be representative. Moreover, with recent technological innovations, it is increasingly convenient and cost-effective to collect large numbers of highly non-representative samples via online surveys.

In the literature, there are many different interpretation of the 'representativeness' concept. See [6] for a thorough investigation of the statistical literature. Here we relate the concept of 'representativeness' to the possibility of obtaining, from the

E. Rocco (✉)
Dipartimento di Statistica, Informatica, applicazioni "G. Parenti",
Università degli Studi di Firenze, Viale Morgagni, 59-50134 Florence, Italy
e-mail: emilia.rocco@unifi.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_18

sample, results that tell us more or less what we would have found by measuring the whole population from which the sample has been selected. Of course this possibility implies the absence in the sampling process of unknown selective forces for whose some groups in the population are over or under represented, and these groups behave differently with respect to the survey variables. Although this definition is appealing, the validity of it can never be tested in practice since results for the whole population are unknown. Moreover as stated by [3] (on p. 286), there are various ways of selecting a sample, but only with random (probability) sampling it is possible to know how representative the sample results are likely to be. A weaker definition of the representativeness concept that can be tested in practise, whatever is the selection process of those who ultimately supply the target data, is that of 'representativeness with respect to a set of auxiliary variables'. A representative sample with respect to one or more auxiliary variables is a sample in which the distribution of these variables is the same as in the population from which the sample is selected. In this paper, when we refer to this last concept of representativeness, we explicitly declare it.

The main problem caused by non-representative survey data is that estimators of population characteristics must be assumed to be biased unless convincing evidence to the contrary is provided. This problem influences the data coming from a probability sample affected by non-response and the data obtained with a convenience sample in the same way. Hence, in both the cases, the same quality indicators may be used in order to evaluate the impact of non-representativeness and the same post-survey adjustment methods may be used to deal with it.

In the remainder of this paper we just consider non-response but the points made for it also apply in general to all generation processes of non-representative survey data.

It is well known that non-response bias is the product of non-response rates and differences between respondents and non respondents on the statistic of interest. Of course previous to the survey the statistic of interest is unknown and when non-response occurs its value can be estimated only for respondents. Therefore the non-response bias cannot be assessed except through indirect measures based on more or less reasonable assumptions and on the use of data external to the survey.

Despite its incomplete nature, the response rate has long been used as a key measure of the risk of non response bias. But, nowadays, it is well-known finding in survey methodology that it is a poor indicator of non-response bias, see, among others, [2, 4, 11].

Consequently, in recent literature, various alternative indicators for monitoring the risk of non-response bias in surveys have been proposed. Wagner [13] provides a taxonomy of such measures based on the types of data used to estimate each one. More in detail, in order to explicitly differentiate the response rate from the other indicators, Wagner [13] describes the following three types of alternative indicators:

1. indicators involving only the response indicator that is a binary variable that indicates if a sampled unit responds or not;

2. indicators involving the response indicator and auxiliary data that are known for all sample units and may stem from sampling frame data, administrative data and data about the data collection process;
3. indicators involving the response indicator, auxiliary data and survey data (i.e. the data for respondents).

It is evident that the only indicator of the first type is the response rate.

Indicators of the second type use auxiliary data for predicting the response indicator and provide a single measure of the risk of non-response bias for the whole survey, relying on the implicit assumption that the auxiliary variables used to create them are correlated with all the survey estimates. The fact of providing a single measure for the whole survey is a strength of such indicators since it allows them to be used as tools for comparing different surveys and surveys over time, and for a comparison of different data collection strategies and modes. However, it is also a weakness, because, a single measure of the risk of non-response bias for the whole survey could lead to incorrect conclusions for the survey statistics for which the implicit assumption of correlation with the auxiliary data used to create such risk measure is not likely to be true.

Indicators of the third type, which, in addition to the response indicator and the auxiliary variables, use the observed survey data are defined at a statistic level. Since non-response bias depend on the difference between the statistic of interest and its estimate based on the observed (respondents) data, an indicator that uses these data, if the model assumption on which it relies is good, allows for directly achieving information about the bias. However the definition of such indicators at statistic level is also a weakness of them. Given that most surveys have multiple objectives, there would be more indicators that makes the computation process more complex than for the other two types of indicators and could lead to potentially different conclusion about data-collection strategy.

In this paper we suggest the combined use of two indicators. The first, is a prominent indicator of the second type, the 'R-Indicator' suggested by [11]. It assumes the availability of auxiliary variables at the sample level and employs them for estimating the response probability. Then, it judges the level of representativeness of the respondents with respect to these auxiliary variables, by measuring the extent to which the estimated response probabilities vary. In fact, according to the concept of representativeness with respect to a set of auxiliary variables, if the set of respondents is representative then the estimated response probabilities are the same for all units in the population. The R-Indicator, by judging the composition of respondents by a pre-defined set of variables that are observed outside of the survey, provides a single measure of the response quality for the survey as whole. It does not give any direct information about the bias of a single survey statistic. Assuming that the model adopted for estimating the response probabilities is valid, in order to have direct information about the bias of a single survey statistic we may investigate the relation between the specific survey variable and the estimated response probabilities. It is obvious that when the estimated response probabilities are the same for all the population units, the respondents are representative with respect to any variable and

therefore the risk of non-response bias is negligible for all the statistics of interest. When response probabilities vary the risk of bias is not the same for all the possible statistics of interest since it depends not only on the variability of the response probabilities but also on the relation between them and the specific survey variable: if the survey variable is on average the same for different values of the response probabilities the risk of bias is low, the more it varies as the response probabilities vary, the greater is the risk of bias. The R-Indicator measures only the variability of the response probabilities, it does not provide us any information on the effects that this variability has on the risk of bias for a single statistic of interest. For this reason we suggest to use, in addition to it, a new indicator of the third type. This new indicator, referred to hereafter as R-Statistic-level-Indicator, assesses the risk of bias of a single survey statistic by measuring for the specific survey variable the variation of the respondents average across the percentiles of the response probabilities predicted in order to estimate the R-Indicator itself. When for a specific survey variable the respondents averages associated with the different percentiles of the response probabilities are different, the risk of bias increases both with the increase in the difference between these averages and with the increase in the difference between the percentiles of the response probabilities. The R-Statistic-level-Indicator allows to judge only the first increment, whereas the R-Indicator allows to judge only the latter one. Therefore, in order to judge the bias that the non-response or the non-probability sampling may cause on a single statistic of interest, it is advisable to use them both together.

In order to evaluate the performance of this combined use of the R-Indicator and the R-Statistic-level-Indicator, a simulation study is carried out under various scenarios.

The remainder of the paper is arranged as follows. In Sect. 2 the theoretical framework is introduced. In Sect. 3 the R-indicator and the R-Statistic-level-indicator are defined. A simulation study and results of that study are described in Sect. 4. Section 5 concludes and discusses future work.

2 Theoretical Framework and Notation

We assume that a sample survey is undertaken, where a probability sample s of n units is selected, from a population U of N units labelled $i (i = 1, \ldots N)$. The sample is drawn by employing a sampling design $p(s)$ and the first order inclusion probability for unit i is denoted π_i. The survey is subject to unit non-response, therefore only a subset $r \subset s$ of $n_r \leq n$ responding units is observed. We shall suppose that the target of inference is a population parameter (like the mean or the total) of a survey variable taking value y_i for unit i and that the data available for estimation purposes consist of the values $\{y_i; i \in r\}$ of the survey variable and the values $\{\mathbf{x}_i = (x_{1,i}, \ldots x_{K,i}); i \in s\}$ of a vector of auxiliary variables that may influence the non-response mechanism and/or the survey variable.

Unlike the sampling selection the survey sampler has no control over the response mechanism. Therefore, to account for it in the estimation of the parameters of interest, it becomes necessary to model it. To this end, a response indicator δ_i is defined so that $\delta_i = 1$ if unit i responds and $\delta_i = 0$ otherwise. The response mechanism corresponds to the distribution of the vector $\{\delta_i : i \in s\}$. We consider only the case where units respond independently from each other and from s. Moreover, taking advantage of the fact that a vector of auxiliary variables $\mathbf{x}_i$, which may influence the response, is know for both respondents and non-respondents, we define the response probability as the conditional expectation of δ_i given $\mathbf{x}_i$ i.e. we set $\rho_i \equiv \rho(\mathbf{x}_i) = E(\delta_i|\mathbf{x}_i)$, where $\rho(.)$ is an unspecified function of $\mathbf{x}_i$, with $\rho(.) \in (0, 1]$.

Since y_i is only observed for respondents, the response probability conditional on y_i is generally inestimable without further assumption. However, the response probabilities defined as conditional on $\mathbf{x}_i$ capture the feature of non response mechanism relevant for the target of inference only if δ_i is conditional independent of y_i given $\mathbf{x}_i$, that is only if the non-response is 'Missing at Random' (MAR) given the vector of auxiliary variables ([7], p. 12). Making this further assumption we finally have

$$pr(\delta_i = 1|i \in s, y, \mathbf{x}) = \rho(\mathbf{x}_i) \equiv \rho_i \tag{1}$$

for all $i \in U$.

3 Risk Indicators of Non-response Bias

When non-response occurs, quality indicators are needed to measure its impact. To this end, in this section, first we review the definition of two indicators: the R-indicator and the R-Statistic-level-Indicator. Next we suggest and discuss the opportunity in each survey of monitoring the risk of non-response bias though a combined use of them.

3.1 The R-Indicator

The basic idea of the R-indicator ('R' for representativeness) suggested by Schouten et al. [11] is that a response subset is representative with respect to a vector of auxiliary variables $\mathbf{x}$ when response propensities are constant for $\mathbf{x}$. Relying on this idea, it measures the extent to which the response probabilities $\rho_i = \rho(\mathbf{x}_i)$ vary as a function of their population standard deviation:

$$S_\rho = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2} \tag{2}$$

where $\bar{\rho} = \sum_{i=1}^{N} \rho_i / N$. The R-indicator is modelled in terms of S_ρ as follows:

$$R_\rho = 1 - 2S_\rho, \tag{3}$$

therefore, it will be higher when the variability among the response probabilities is lower.

Since it may shown that $S_\rho \leq \sqrt{\bar{\rho}(1-\bar{\rho})} \leq 0.5$, then R_ρ takes values on the interval [0, 1]. The value $R_\rho = 1$ indicates the most representativeness and the value $R_\rho = 0$ indicates the least representativeness.

In practice, the response propensities are unknown. However, when auxiliary data are available at a sample level, it is possible to estimate them for all sampled units and to replace R_ρ with the estimator:

$$\hat{R}_\rho = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i \in s} \frac{(\hat{\rho}_i - \hat{\bar{\rho}})^2}{\pi_i}} \tag{4}$$

where $\hat{\bar{\rho}} = \sum_{i \in s} (\hat{\rho}_i / \pi_i)/N$ [12].

The response propensities, $\hat{\rho}_i$, are commonly estimated with explicit or implicit models linking the response occurrences to the auxiliary variables, for instance, by using a logistic or a probit regression model, or the weighting within cell method.

Assuming that the implicit or explicit model used for estimating the response propensities is correct, $\hat{R}_\rho$ may be viewed as a lack of association measure between the response mechanism and the auxiliary variables used for estimating the response propensities. A value of $\hat{R}_\rho$ close to 1 which corresponds to a low variation of the estimated response propensities, $\hat{\rho}_i$, denotes a weak association. The weaker the association the better, as this implies there is no evidence that non-response has affected the composition of the observed data and, therefore, it implies that the risk of non-response bias is low whatever is the target statistic as long as it satisfies the 'MAR assumption'.

Unfortunately, however, a value of $\hat{R}_\rho$ not close to 1, which denotes an higher association between the response mechanism and the auxiliary variables used for estimating the response propensities, it is not equally able to detect the risk of bias for a survey estimate. Shlomo et al. [12] show that the variance of the response probabilities conditional on the vector of auxiliary variables may be split into two (unmeasurable) components. One represents the variation of the probabilities conditional on the survey variable, that is the probabilities which we should ideally like to use. The other one represents additional variation of the estimated response probabilities which is unrelated to the survey variable, that is a sort of noise due to the use for the estimation of the response probabilities of auxiliary variables predictive of the response probabilities but weakly associated to the survey variable. A low value of the R-Indicator indicates an high variability of the response probabilities, but if this variability depends on an high value of the first described component, the risk of bias for the statistic of interest is high; on the contrary, if it depends on an high

value of the second described component, the risk of bias for the statistic of interest is low. Moreover, since the second component increases as the association between the auxiliary variables and the variable of interest decreases we can state that, when the R-indicator denotes a level of association not negligible between the auxiliary variables and the response indicator, in order to judge the risk of bias for a specific statistics of interest it is necessary to evaluate the association between the auxiliary variables and the variable of interest. The R-Statistic-level-Indicator suggested in Sect. 3.2 allows to evaluate this association.

3.2 The R-Statistic-Level-Indicator

In order to judge the risk of bias for a single survey estimate when the estimated value of the R-Indicator is not close to 1, that is in order to evaluate the level of association between the variable of interest and the auxiliary variables when the association between these latter variables and the response indicator is not negligible, we suggest, to compare respondents averages of the specific variable across the percentiles of the response propensities predicted in order to estimate the R-Indicator itself. You could simply plot these means [9] or you can synthesize their difference through an indicator. We define such an indicator by considering the following steps:

1. respondents are ordered with respect to the estimated response propensities, $\hat{\rho}_i$, $i \in r$;
2. the ordered respondents set is then partitioned into H classes, $r_h (h = 1, \ldots H)$, of size $n_{r_h} =$ roughly n_r/H on the basis of $H - 1$ percentiles of the estimated response propensities;
3. the respondents mean, $\overline{y}_{r_h} = \sum_{i \in r_h} y_i / n_{r_h}$, of the target variable for each class is calculated;
4. finally, the indicator, named R-Statistic-level-Indicator, is defined as:

$$R_{\overline{y}_{p\rho}} = 1 - \frac{\sum_{h=1}^{H} (\overline{y}_{r_h} - \overline{y}_r)^2 n_{r_h}}{\sum_{i \in r} (y_i - \overline{y}_r)^2} \tag{5}$$

where $\overline{y}_r = \sum_{i \in r} y_i / n_r = \sum_{h=1}^{H} \sum_{i \in r_h} y_i / n_r$ is the total respondents mean.

Since, for the well known deviance decomposition formula

$$\sum_{h=1}^{H} (\overline{y}_{r_h} - \overline{y}_r)^2 n_{r_h} \leq \sum_{i \in r} (y_i - \overline{y}_r)^2, \tag{6}$$

then $R_{\overline{y}_{p\rho}}$ takes values on the interval [0, 1]. The value $R_{\overline{y}_{p\rho}} = 1$ indicates a risk of non-response bias negligible whatever is the value of the R-Indicator and is obtained when there is no association between the variable of interest and the auxiliary variables used for predicting the response probabilities. Moreover, the value of $R_{\overline{y}_{p\rho}}$

decreases when the homogeneity of the values of response variable within each class increases and this may happen either because the response probabilities are altogether less variable or because there is a significant association between the variable of interest and the auxiliary variables. Only a low value of $R_{\bar{y}_{p\rho}}$, due to this latter cause implies a high risk of non response bias for the statistic of interest. We can exclude to be in this situation if we calculate the R-Indicator too, and its value is close to 1. On the contrary, when the R-Indicator and the R-Statistic-level-Indicator are both low, this indicates that the association of the auxiliary variables is high both with the response indicator and with the study variable and consequently that the risk of non response bias for the statistic of interest is high.

3.3 The Combined Use of R-Indicator and R-Statistic-Level-Indicator

The R-indicator as well as all indicators of the second type in the 'Wagner classification' [13], provides an overall measure on the risk of non-response bias for the whole survey and does not give any direct information about the real bias of a single survey statistic. Therefore, in a multi-purpose survey $\hat{R}_\rho$ could be a better indicator for some survey statistics and a less effective one for others. In fact, in a survey with several survey variables it would be unlikely to identify a set of auxiliary variables correlated together with the response probability and with any survey variable. On the other hand in a multiple objective survey the estimation of a different indicator for each variable of interest could be unfeasible during the data collection phase. Therefore, according to us, the process for analyzing the survey data quality in presence of non-response must be organized in two stages:

1. the first, prior to the data processing, in which it is monitored the risk of self-selection bias for the survey as a whole;
2. the second, during the data processing, which aim is to evaluate the risk of bias for each survey estimate and the opportunity or not to adopt for it a non-response adjustment method.

The estimation of an indicator that allows to monitor the survey data collection process as a whole, like $\hat{R}_\rho$, may be useful for the following two reasons.

First, when some auxiliary variables, relevant for describing the survey population, are available, it is reasonable to ask whether the subset of respondents is representative, at least, with respect to these, and $\hat{R}_\rho$ can provide the answer. For example, for the official statistics produced by the national institutes of statistics, the compliance with some external coherence constraints regarding some socio-demographic variables, available from the frame is very important.

Moreover, if the model assumption on which $\hat{R}_\rho$ relies is good, and this indicator is used for adapting the data-collection process in order to achieve a highly representative response set and, this result is achieved, that is on the final set of respondents a

value of $\hat{R}_\rho$ close to 1 is estimated, this allows to predict a low risk of non-response bias for all the statistics of interest.

Assuming that the model adopted for predicting the response propensities is good, the second stage for evaluating the risk of bias needs only in the case in which $\hat{R}_\rho$ is not close to 1. For these cases we suggest to estimate separately for each target variable the indicator $R_{\overline{y}_{p\rho}}$. It allows for each survey variable to confirm or to deny the risk of non response bias detected by a value of $\hat{R}_\rho$ not close to 1. The risk is confirmed for all the target statistics corresponding to an estimated value of $R_{\overline{y}_{p\rho}}$ not close to 1.

The two indicators, R_ρ and $R_{\overline{y}_{p\rho}}$ may be seen as mutually complementary, neither, used alone allows to detect the risk of non-response bias for all the situations. On the contrary if they are used jointly they achieve this aim. The risk of non response bias is high when they are both not close to 1. Obviously, their efficacy is subordinated to the validity of the assumptions on the response process that have been specified in Sect. 2 and on which they both stem.

4 Simulation Study

In this section, we describe a simulation study which aims to empirically explore the ability of the combined use of $\hat{R}_\rho$ and $R_{\overline{y}_{p\rho}}$ to detect non-response bias under various scenarios. To this end we have reproduced the simulation setting that Little and Vartivarian [8] have used in order to provide an empirical proof of the fact that the non-response weighting adjustments based on adjustment cells are effective in reducing bias of an unweighted respondents mean only if the auxiliary information used for defining the adjustment cells is related to both the non-response mechanism and the outcome of interest.

Simulation setting:

1. to keep things simple it is assumed that: (a) the sample is selected by simple random sampling, (b) only the population mean of a target variable y is of interest and, (c) only one auxiliary variable x is available;
2. x is a categorical variable with 10 categories that identify 10 cells of adjustment;
3. conditional on the sample size, the sampled cases have a multinomial distribution over the (10×2) contingency table based on the classification of the response indicator, δ, and x, with cell probabilities

$$pr(\delta = 1, x = c) = pr(\delta = 1)\, pr(x = c|\delta = 1)$$
$$pr(\delta = 0, x = c) = (1 - pr(\delta = 1))\, pr(x = c|\delta = 0) \quad c = 1, \ldots, 10$$

   given in Table 1 for two marginal response rates, 70% and 52%, and three conditional distributions of δ given x corresponding to high, medium and low association between the two variables;
4. the simulated distribution of y given $\delta = h$, $(h = 0, 1)$, and $x = c$ have the form:

Table 1 Percent of samples in cell $x \times \delta$

Response rate = 52%											
Association x and δ	x	1	2	3	4	5	6	7	8	9	10
High	$\delta = 1$	0.55	1.00	4.01	4.52	5.04	5.55	6.06	6.58	9.14	9.96
	$\delta = 0$	8.69	9.00	6.01	5.53	5.04	4.54	4.04	3.54	1.02	0.20
Medium	$\delta = 1$	2.77	3.50	4.01	4.52	5.04	5.55	6.06	6.58	7.11	7.62
	$\delta = 0$	6.47	6.50	6.01	5.53	5.04	4.54	4.04	3.54	3.05	2.54
Low	$\delta = 1$	4.62	5.15	5.21	5.28	5.34	5.40	5.45	5.52	5.58	5.64
	$\delta = 0$	4.62	4.85	4.81	4.77	4.73	4.69	4.65	4.60	4.57	4.52
Response Rate = 70%											
Association x and δ	x	1	2	3	4	5	6	7	8	9	10
High	$\delta = 1$	0.55	3.00	6.51	7.04	7.55	8.07	8.59	9.11	9.64	9.96
	$\delta = 0$	8.69	7.00	3.51	3.02	2.52	2.02	1.52	1.01	0.51	0.20
Medium	$\delta = 1$	4.44	5.30	5.81	6.33	6.85	7.37	7.88	8.40	8.93	9.45
	$\delta = 0$	4.80	4.70	4.21	3.72	3.22	2.72	2.22	1.72	1.22	0.71
Low	$\delta = 1$	6.19	6.85	6.91	6.98	7.05	7.11	7.17	7.24	7.31	7.37
	$\delta = 0$	3.05	3.15	3.11	3.07	3.02	2.98	2.93	2.88	2.84	2.79

$$[y|\delta = h, x = c] \sim N(\beta_0 + \beta_1 x, \sigma^2), \tag{7}$$

and three sets of values of (β_1, σ^2) corresponding to high, medium and low association between y and x are considered and shown in Table 2. The intercept β_0 is chosen so that the overall mean of y is 26.3625 for each scenario;

5. 10,000 replicate samples of size 400 were simulated for each combination of parameters in Tables 1 and 2;
6. for each replica the following estimates have been produced: (a) the unweighed mean of the respondents; (b) the response probability, for each unit in the sample, by using the 'weighting within cell' method and identifying the cells with the 10 categories of x; (c) $\hat{R}_\rho$; (d) $R_{\bar{y}_{p\rho}}$, for which the respondents have been partitioned into the quintiles of the response probabilities (there is no rule for choosing the

Table 2 Parameters β_1 and σ^2 for outcome model (7)

Association between x and y	β_1	σ^2
High	4.75	46
Medium	3.70	122
Low	0.00	234

Table 3 Summaries of results based on 10,000 replicate samples. Response rate = 52%

Association between x and δ	Association between x and y	Emp. bias of unweighted mean (%)	$\hat{R}_\rho$	$R_{\bar{y}_{p\rho}}$
High	High	27.24	0.43	0.37
	Medium	21.23	0.43	0.65
	low	0.02	0.43	0.98
Medium	High	14.76	0.68	0.36
	Medium	11.48	0.68	0.62
	Low	0.05	0.68	0.98
Low	High	2.16	0.85	0.69
	Medium	1.68	0.85	0.81
	Low	0.00	0.85	0.99

Table 4 Summaries of results based on 10,000 replicate samples. Response rate = 70%

Association between x and δ	Association between x and y	Emp. bias of unweighted mean (%)	$\hat{R}_\rho$	$R_{\bar{y}_{p\rho}}$
High	High	19.10	0.44	0.36
	Medium	14.91	0.44	0.75
	low	0.08	0.44	0.99
Medium	High	11.32	0.70	0.33
	Medium	8.86	0.70	0.60
	Low	0.04	0.70	0.99
Low	High	2.16	0.86	0.69
	Medium	1.68	0.86	0.81
	Low	0.00	0.86	0.99

number of classes, in this choice we have taken into account the following two points: first, Cocran [1] shows that stratification into five subclasses removes approximately 90% of bias due to the stratifying variable; second, the use of five classes of response propensities is an accepted practice in the context of non-response weighting adjustments).

The empirical relative bias of the unweighted mean of the target variable, the median across the replications of $\hat{R}_\rho$ and the median across the replications of $\hat{R}_y$ are reported in Table 3 for simulation with a response rate of 52% and in Table 4 for simulation with a response rate of 70%.

The pattern of results is very similar in the two tables. From both tables we note that:

1. When the association between δ and x is low, the $\hat{R}_\rho$ value is high and the bias of the unweighted mean, even thought it decreases with the decreasing of association between y and x, is always very low.
2. On the contrary, a low value of $\hat{R}_\rho$ does not necessarily mean a high bias of the unweighted mean since when the association between y and x is low, the bias of the unweighted mean is negligible irrespective of the value of $\hat{R}_\rho$.
3. A value of $R_{\bar{y}_{p\rho}}$ close to 1 allows for identifying the situations in which, given a low association between y and x, the bias of the unweighted mean is negligible.
4. Finally, the two indicators, considered jointly, allow for discriminating between the statistics for which the risk of non-response bias is higher ($\hat{R}_\rho$ and $R_{\bar{y}_{p\rho}}$ are both close to 0) from those for which it is lower ($\hat{R}_\rho$ or $R_{\bar{y}_{p\rho}}$ is close to 1).

5 Final Remarks

Any analysis that deals with missing data must make some model assumption, either implicitly or explicitly. The quite general assumption that is made when $\hat{R}_\rho$ and/or $R_{\bar{y}_{p\rho}}$ are used, as indirect measure of the risk of non-response bias, is that the response process is MAR. The results highlight that, under the MAR assumption, for detecting the risk of non-response bias is important to consider the relation of the auxiliary variables, that make MAR the response process, with both the probability of response and the survey variable. $\hat{R}_\rho$ and $R_{\bar{y}_{p\rho}}$ measure one only of these relations, $\hat{R}_\rho$ the former and $R_{\bar{y}_{p\rho}}$ the latter. The joint use of $\hat{R}_\rho$ and $R_{\bar{y}_{p\rho}}$ allows to measure both.

The evidence for the importance of monitoring the risk of non-response bias by examining the relation of the auxiliary variables with both the probability of response and the survey variable is in agreement with known finding in the literature on the non-response weighting adjustment methods: [5, 7, 10], among others, show that an auxiliary variable used for a weighting adjustment must have two characteristics to reduce non-response bias: it needs to be related to the probability of response and it needs to be related to the survey variable.

We are all aware that non-response bias has long been a concern for surveys and, even more so over the past decades with the increasing decline of the response rates. Anyway, over the past decades it is also increased the availability of auxiliary information that may be used for indirectly evaluating the non-response bias and/or for exploring possible weighting or imputation adjustment methods. Unfortunately, however, when there are several survey variables and several auxiliary variables, may be very difficult to select a single set of auxiliary information for monitoring the risk of bias of all survey variables and/or for identifying a single method of adjustment for all of them. The possibility to investigate the quality of survey data in two stages (suggested in this paper), or in more stages, may be an opportunity. The results are encouraging, but more investigations need. The investigation of how this approach works in more complicated scenarios is a first area of research. A second area of research is to consider the possibility to extend this approach to cases in which non-

response is Not-Missing-at-Random (NMAR). An other area of research concern the relation between the indicators for monitoring the risk of non-response bias and the weighting or imputation methods used to deal with it, the two themes are both largely studied in literature, however, according to our research, their joint study is an option that is untouched for the most part.

References

1. Cocran, W.G.: The planning of observational studies of human populations. J. R. Stat. Soc. Ser. A **128**, 234–255 (1965)
2. Curtin, R., Presser, S., Singer, E.: The effects of response rate changes on the index of consumer sentiment. Public Opin. Q. **64**, 413–428 (2000)
3. Ehrenberg, A.S.C.: Data Reduction; Analysing and Interpreting Statistical Data. Wiley, New York (1975)
4. Groves, R.M., Peytcheva, E.: The impact of nonresponse rates on nonresponse bias: A meta-analysis. Public Opin. Q. **72**, 167–189 (2008)
5. Kim, J.K., Kim, J.J.: Nonresponse weigthing adjustment usuing estimated response probabilities. Can. J. Stat. **35**, 501–514 (2007)
6. Kruskal, W., Mosteller, F.: Representative sampling III: Current statistical literature. Int. Stat. Rev. **47**, 111–123 (1979)
7. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. Wiley, NJ (2002)
8. Little, R.J.A., Vartivarian, S.: Does weighting for nonresponse increase the variance of survey mean. Surv. Meth. **31**, 161–168 (2005)
9. Olson, K.: Survey partecipation, nonresponse bias, measurement error bias and total bias. Public Opin. Q. **70**, 737–758 (2006)
10. Rocco, E.: Using auxiliary information and nonparametric methods in weigthing adjustments. In: Torelli, N., et al. (eds.) Advanced in Theorethical and Applied Statistics, Studies in Theorethical and Applied Statistics, pp. 193–302. Springer, Berlin (2013)
11. Schouten, B., Cobben, F., Betleehem, J.: Indicators for the representativeness of survey response. Surv. Meth. **35**, 101–113 (2009)
12. Shlomo, N., Skinner, C., Schouten, B.: Estimation of an indicator for the representativeness of survey response. J. Stat. Plan. Inference **142**, 201–211 (2012)
13. Wagner, J.: A comparison of althernative indicators for the risk of nonresponsebias. Public Opin. Q. **76**, 555–575 (2012)

Data Science Methods for Social and Population Studies

The Propensity to Leave the Country of Origin of Young Europeans

Paolo Balduzzi, Alessandro Rosina and Emiliano Sironi

Abstract Using data from the "Youth Project", a survey carried out by the Toniolo Institute for Advanced Studies, we provide evidence of the determinants of the propensity to leave the native country by young Europeans and show how this phenomenon depends on the economic opportunities offered by the countries of origin. In addition, we underline the effect of individuals' trust in the economic development of the country of residence as a main predictor of the intention of moving away.

Keywords Mobility patterns · Brain drain · Migration

1 Introduction

The propensity to leave for young people around the world has always been very high. On one (sunny) side, going abroad means gaining new experiences; but on the other (dark) side, going abroad also means escaping from a country without opportunities.

This phenomenon is currently assuming great importance, because of the increasing involvement of highly skilled workers in Europe: an increasing flow of high skilled workers, who decide to migrate, enriches destination countries and is a drain on the home countries [4]. Media reports and political debate tend to confirm, or even to empower, this view. Nonetheless, the scientific literature has been much less pessimistic and tried to highlight also the positive effects for the home countries. Among the others, one of these effects is that highly skilled migrants may come back after some years and bring additional human—and possibly even monetary—capital with them [14]. Traditionally, migrant workers tend to transfer part of their earnings to

P. Balduzzi · A. Rosina · E. Sironi (✉)
Università Cattolica Del Sacro Cuore, Milan, Italy
e-mail: emiliano.sironi@unicatt.t

P. Balduzzi
e-mail: paolo.balduzzi@unicatt.it

A. Rosina
e-mail: alessandro.rosina@unicatt.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_19

their country of origin (or rather, to their families in the country of origin). In poorer counties, remittances may repay or even more than compensate for the educational cost. Some recent literature (see, for instance, [6–8, 10–12, 15]) also highlights how the prospects of migration are positively correlated with the increase in human capital in the country of origin: anticipating higher earnings abroad, a greater number of citizens in the country of origin will be willing to invest in education compared to the case in which this perspective does not exist (the so-called "brain effect"). Hence, given that the "effective" migratory behavior is lower than the "desired" migratory behavior, this brain effect is positive.

However, especially in the Italian case, the chronical lack of opportunities in the labor market [2, 3] undermines the chances to favor the return of high skilled workers after a positive experience spent abroad. Hence, what seems like a temporary experience becomes more and more a permanent choice that does not allow the migrants to come back in their country of birth [5].

Unfortunately, there is no useful, relevant, comparable and consistent database on this phenomenon. At first sight, the most important data sources are certainly AIRE and ISTAT. On the one hand, AIRE (Register of Italian Residents Abroad) provides a photograph of Italian residents abroad. However, these data only cover Italians who voluntarily signed up and provided information. The incompleteness and inconsistency of these data, due to the not mandatory nature of the registration to the source, are likely to provide only a partial—if not misleading—view. In addition, publicly accessible data are very limited and do not contain any information about the degree of education of Italian residents abroad or about the reason for their migration, nor about the work done or the income received. It is clear, in light of the above, that these data are necessary to evaluate any costs and benefits of our country's (possible) brain drain. On the other hand, ISTAT (Italian National Institute of Statistics) provides data on residence transfers, that is data on the annual variations in migration from Italy. An additional source of information is the Italian Minister for University and Research (MIUR), which provides interesting information on the number and origin of foreign students in Italy, as well as on the number of graduates.

In this paper, we use data from the "Youth Project", a survey carried out by the Toniolo Institute for Advanced Studies, the founding institution and promoter of the Catholic University of the Sacred Heart. Using that source of data, which provides a representative sample of young Europeans aged between 18 and 32, we implement an ordered logistic regression, in order to identify the determinants of the propensity to leave their native country in the five biggest European countries: Germany, France, UK, Italy and Spain.

Results should suggest helpful indications on the main drivers of the "brain drain", which are useful to prevent or reduce a phenomenon that is supposed to weaken the quality of the local labour force.

This paper is organized as follows: after a brief review of the literature in the introduction, we present our datasets and some descriptive statistics regarding the population under analysis (Sect. 2). In Sect. 3 we analyze and discuss the determinants of the propensity to leave and finally present our methodology and the empirical results about the propensity to return. Finally, Sect. 4 concludes.

2 A Descriptive Analysis of the Mobility of Young Europeans

This analysis is based on data describing young Europeans' propensity to leave their country of origin. Information were collected among 5,000 young people from Italy, France, Germany, Spain and the UK (1.000 individuals from each country) within the survey "Youth Project" carried out in July 2015 by the "Toniolo Institute for Advanced Studies" and jointly realized with the scientific cooperation of Ipsos. The survey was realized using quotas that are designed to reflect the population and are representative of the country's population after having corrected frequencies through a weighting procedure. The sampling variables are gender, age, employment status (employed/not employed), respondents' education (as classified in Table 1), geographical area (at the level NUTS 1) and size of the municipality of residence.[1] Descriptive statistics about that sample are reported in Table 1 and no missing data have been registered for the variables of interest.

The first striking difference among countries come from education: on average, 35.7% of the sample has a university degree; nonetheless, in Italy this share is below 18%; on the contrary, more than 45% of the UK sample has a degree. Relevant differences emerge from the employment status as well: on average, 45.8% of respondents are not employed: but young unemployed in Germany and the UK are only 37.5% and 36% of the population, respectively; whereas almost 60% of young Italians are unemployed. Finally, in UK 15.4% of the sample has at least one foreign parent and 7.0% of the entire sample is foreign. On average, 85% of the sample lives in the country of citizenship, with peaks in Italy (89.8%) and Spain (88.9%). Individuals with at least a foreign parent are above the average (9.7%) in France (12.7%) and Germany (10.7%).

Considering the high attitude to mobility that characterizes young Europeans in comparison with the older cohorts we are interested on the evaluation of their attitudes towards migration in a comparative perspective.

Table 2 presents answers of young Europeans to the question: "Is migrating the only opportunity to fully realize yourself?". Results shed light on this question and the results are quite dramatic for countries such as Italy and Spain. In almost all countries (see Table 2), more than 50% of the sample agrees "Very" or "Quite" with the statement that migrating is necessary to realize themselves, though with relevant differences: from 91% in Spain to 88.3% in Italy and to 53.9% in the United Kingdom. The only exception is Germany, where the percentage is 47.6% (almost half of Italy and Spain).

As regards real or desired willingness to move of young Europeans, it is interesting what emerges from another set of questions investigating the choices of possible destinations. Table 3 presents data regarding the opinion on the general attractiveness of other countries (first choice). Italians confirm their traditional destinations (Germany and the United States, along with the United Kingdom).

[1] For further details on the surveys from the "Youth Project", please see the methodological appendix of Istituto Toniolo (2016) [13].

Table 1 Frequencies in the sample population

	Italy (%)	France (%)	Germany (%)	UK (%)	Spain (%)
Gender					
Male	50.8	50.0	51.2	50.4	50.5
Female	49.2	50.0	48.8	49.6	49.5
Age					
18–20 years	18.3	20.0	17.3	19.0	16.6
21–23 years	19.4	20.0	19.5	20.3	18.0
24–26 years	20.1	19.6	21.2	20.4	19.3
27–29 years	20.5	20.1	20.6	20.1	21.4
30–32 years	21.8	20.3	21.3	20.1	24.7
Education					
High	17.9	44.7	28.5	45.8	41.5
Medium/Low	82.1	55.3	71.5	54.2	58.5
Employment					
Employed	40.5	56.1	62.5	64.0	47.9
Not employed	59.5	43.9	37.5	36.0	52.1
Nationality					
Same country	89.8	84.9	83.8	77.7	88.9
At least one foreign parent	5.0	12.7	10.7	15.4	4.9
Foreigner	5.1	2.4	5.5	6.9	6.2

Source Balduzzi and Rosina [4]

Table 2 Is migrating the only opportunity to fully realize yourself?

	Italy (%)	France (%)	Germany (%)	UK (%)	Spain (%)
Very much	43.0	19.4	10.7	11.1	35.5
Quite	45.3	51.2	36.9	42.8	55.5
Little	10.1	25.5	40.9	38.0	7.2
Not at all	1.6	3.9	11.5	8.1	1.8

Source Balduzzi and Rosina [4]

For the Germans, only the US can offer better opportunities. Italy is the last in the ranking of the favorite destinations of all young Europeans.

This element contributes to worsen the opinion of Italy as a possible destination for foreign European young adults.

Table 3 Country attractiveness (first choice)

Origin	Destination country						
	Italy	France	Germany	UK	Spain	US	Rest of the world
Italy	NA	3.5%	12.2%	14.0%	1.5%	17.5%	51.3%
France	0.2%	NA	3.6%	10.0%	0.9%	20.2%	65.1%
Germany	0.2%	1.2%	NA	4.8%	1.7%	21.4%	70.7%
UK	1.1%	1.3%	6.3%	NA	2.5%	22.9%	65.9%
Spain	2.3%	4.6%	14.8%	16.2%	NA	17.5%	44.6%

Note NA: Not Appliable. *Source* Balduzzi and Rosina [4]

3 A Multivariate Analysis of Young Europeans' Propensity to Leave

After the descriptive analysis presented in the previous section, we now focus with more detail on the propensity to move to increase employment opportunities. More precisely, we address the pattern of answers investigating the intentions of moving abroad within one year from the time of the interview. This item is worded as follows "Do you intend to move to another country next year to improve your job opportunities, for study or to reach other people?". The answer included four options: "Surely not" (coded as 1), "Probably not" (2), "Probably yes" (3) and "Surely Yes" (4).

The pattern of answers for each of five countries included in the survey are listed below in Table 4.

As we can see from the table Italy shows the highest proportion of young adults intentioned to move to another country in the next twelve months from the time of the interview. In more detail, more than 40% of the interviewed people do not exclude to move away against percentages below the threshold of 20% in France, Germany and UK. Spain generally performs worse than these last three countries but better than Italy: only 29% of individuals declare to be probably or surely intentioned to move away.

Table 4 Intentions to move to another country within the next year from the time of the interview

	Italy	France	Germany	UK	Spain
1—Surely not (%)	17.3	37.7	52.0	42.4	27.8
2—Probably not (%)	41.4	43.7	37.1	43.3	43.2
3—Probably yes (%)	34.4	14.2	7.9	12.1	24.5
4—Surely yes (%)	6.9	4.4	3.0	2.2	4.5

Source Original elaborations from Youth Project (2015)

Through an ordered logistic regression, it is possible to relate these answers with some important explanatory factors that can match both demo-social characteristics and the respondents' perception of their condition (current and future) in the country where they live. Unfortunately, the parallel lines assumption, which an ordered logit is based on, has been violated as described by the results of the Brant test. This means that the size of the coefficients of some explanatory variables depends on the cut-off points of the dependent variable when we collapse the four available answers in a binary outcome.

Hence, as suggested by Williams [16], estimates from an ordered logistic regression have been replaced by those from a generalized ordered logit (results are displayed in Tables 5 and 6).

This model allows to perform different strategies to treat variables that fail the parallel line assumptions and those not violating that condition.

More specifically, if the parallel line assumption holds, the coefficient estimates are the same as from an ordered logistic model. If the parallel line assumption fails, a series of cumulative logit models has been run: the original ordinal variable is collapsed into two categories and a series of binary logistic regressions are estimated. First, it is category 1 (Surely not) versus categories 2, 3, 4 (Probably not/Probably yes/Surely yes); then it is category 1, 2 (Surely not/Probably not) versus categories 3, 4 (Probably yes/Surely yes); then, finally categories 1, 2, 3 (Surely not/Probably not/Probably yes) versus category 4. In each dichotomization the lower values are recoded to zero, while the higher ones are recoded to one.

The explanatory variables for the propensity to labor mobility are the following: gender (male or female), age (in five-year classes), educational level (graduates, secondary school diploma in four or five years, other lower levels), employment status (student, employed, students that are currently working or Neet[2]), having already had experiences abroad (for training or work), the perceived attractiveness of a country for work and/or study experiences, the combination of the perception of opportunities in their own country with respect to other counties and the prospects for future improvement.

These last two factors are specifically derived from the following two questions: "Do you think the opportunities for young people in your country of origin are better or worse than the average of other developed countries?", and "How much confidence do you have in the possibility that in three years the opportunities for young people in your country of origin will be better than today?". The answers to each of these two items have been summarized in two ways: for the former "Very or somewhat lower" versus "Similar or better"; for the latter "Very or quite" versus "Little or nothing". The variable included in the model is formed by four categories derived from the combination of such response modes.

Hence, we obtained the following combinations: (1) people believing that in their country of origin there are actually "Few opportunities" ("Very or somewhat lower" in the first question) and that have "Little confidence in improvements" in the economic condition ("Little or nothing" in the second question); (2) people believ-

[2]Not in Employment, Education, or Training.

Table 5 Estimates from a Generalized Ordered Logit: partial parallel model for describing the intentions of moving abroad in the next year—Five biggest European countries. Categories of the dependent variable: 4 = "Surely yes", 3 = "Probably yes", 2 = "Probably not", 1 = "Surely not"

Explanatory variables	Ordered logit	4, 3, 2 versus 1	4, 3 versus 1, 2	4 versus 1, 2, 3
Gender				
Females		−0.31***	−0.16*	0.09
Males		Ref.	Ref.	Ref.
Age				
18–20	Ref.			
21–23	−0.04			
24–26	0.03			
27–29	−0.17			
30–32	−0.48***			
Education				
< High School Diploma (4–5 years)		0.05	0.05	0.05
High School Diploma (4–5 years)		0.03	−0.15*	−0.34*
Bachelor/Master Degree		Ref.	Ref.	Ref.
Employment				
Neither student nor employed (Neet)		Ref.	Ref.	Ref.
Student		0.30**	0.06	−0.37
Employed		−0.80	−0.43***	−0.69***
Student and employed		0.36**	0.13	−0.49*
Country of residence				
Italy		Ref.	Ref.	Ref.
France		−0.84***	−0.72***	−0.18
Germany		−1.33***	−1.32***	−0.71*
United Kingdom		−1.00***	−1.00***	−1.00***
Spain		−0.58***	−0.58***	−0.58***
Abroad experience				
No	Ref.			
Yes, to study	0.93***			
Yes, to work	1.10***			
Perceived country conditions (with respect to the other countries)				

(continued)

Table 5 (continued)

Explanatory variables	Ordered logit	4, 3, 2 versus 1	4, 3 versus 1, 2	4 versus 1, 2, 3
Few opportunities and little confidence in improvements		0.35***	0.22*	−0.28
Not few opportunities and little confidence in improvements		−0.15*	−0.27**	−0.63**
Few opportunities and confidence in improvements		−0.04	−0.04	−0.04
Not few opportunities and confidence in improvements		Ref.	Ref.	Ref.
Perceived country attractiveness for work and/or study experiences				
Very or sufficiently attractive	Ref.			
Little or nowise attractive	0.05			
Brant test	90.94***			
Observations	5000	5000	5000	5000

Notes: Standard errors in parentheses: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$
Source original elaborations from Youth Project (2015)

ing to actually have "Not few opportunities" but who have "little confidence in improvements"; (3) people feeling to have "Few opportunities" but some degree of "confidence in improvements" in the future; (4) finally, a residual group represented by all the individuals that declared to have "Not few opportunities" and a "confidence in improvements" (chosen as reference category).

The main interest of our analysis relates to this last variable. We wish to evaluate not only the burden of the gap between present-day in the own context and the one in other countries, but also the importance of the persistence of this gap in the future. The idea is that leaving does not necessarily regard those who are worse off today, but actually those who see more room for future chances in the future abroad. In addition, future opportunities are just as important as today's. Hence, the choice of leaving has both a spatial and a temporal dimension; in other words, the choice depends both on the comparison with other countries and on expectations for improvement. Our results tend to confirm this reading. All other factors fixed, the effect of the variable that combines the current conditions with the prospects of development appears to have an important and significant effect.

We consider the following categories: worse opportunities for young people in own country than the average in other developed countries and little confidence that

Table 6 Estimates from a Generalized Ordered Logit: partial parallel model for describing the intentions of moving abroad in the next year—Italy. Categories of the dependent variable: 4 = "Surely yes", 3 = "Probably yes", 2 = "Probably not", 1 = "Surely not"

Explanatory variables	Ordered logit	4, 3, 2 versus 1	4, 3 versus 1, 2	4 versus 1, 2, 3
Gender				
Females	−0.20			
Males	Ref.			
Age				
18–20		Ref.	Ref.	Ref.
21–23		0.76*	0.05	−0.67
24–26		−0.08	−0.08	−0.08
27–29		−0.50	0.03	−0.04
30–32		−0.23	−0.23	−0.23
Education				
< High School Diploma (4-5 years)		−0.56	0.01	0.39
High School Diploma (4-5 years)		−0.12	−0.12	−0.12
Bachelor/Master Degree		Ref.	Ref	Ref.
Employment				
Neither student nor employed (Neet)	Ref.			
Student	0.28			
Employed	−0.17			
Student and employed	0.09			
Region of residence				
North West		Ref.	Ref.	Ref.
North East		0.02	0.02	0.02
Centre		0.11	0.48*	−0.40
South + Islands		0.56*	0.81***	0.13
Abroad experience				
No	Ref.			
Yes, to study	0.80***			
Yes, to work	1.06***			
Perceived country conditions (with respect to the other countries)				
Few opportunities and little confidence in improvements	0.27*			

(continued)

Table 6 (continued)

Explanatory variables	Ordered logit	4, 3, 2 versus 1	4, 3 versus 1, 2	4 versus 1, 2, 3
Not few opportunities and little confidence in improvements	−0.26			
Few opportunities and confidence in improvements	−0.16			
Not few opportunities and confidence in improvements	Ref.			
Perceived country attractiveness for work and/or study experiences				
Very or sufficiently attractive	Ref.			
Little or nowise attractive	−0.04			
Brant test	56.41*			
Observations	1000	1000	1000	1000

Notes Standard errors in parentheses: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$
Source original elaborations from Youth Project (2015)

in the next three years they will improve; worse opportunities in own country but with reasonable confidence for improvement; no worse opportunities than in other countries but little expectation of improvement; no worse opportunities and confidence in further improvement. Specifically, the first category (worse than other countries and low future prospects) significantly differs from the others, with a significant and positive effect on the propensity to move for work. This provides empirical support to the idea that the current status of the country of departure is relevant, along with the low confidence for future improvement.

Conversely, a non-negative perception of the current situation is associated with a lower propensity to mobility.

However, these results require some caution in their interpretation: the expectations regarding the evolution of the economic scenario (the main explanatory variable) and the intentions towards individual future behaviors (the dependent variable) derive from same source of data, so they may be codetermined by the presence of unobservable confounders. Hence, endogeneity may affect the coefficient estimates for the variable describing perceived country conditions.

Anyway, this effect is captured after deducting all other variables included in the regression model and is consistent with the expectations. Nonetheless, it could be useful, as control factors, a reading of the effects of other variables. The intention to move is higher for males than for females in almost all the models (significance disappears in the last model, probably due to the decline in the sample size of responses

equal to one). The level of education shows a U-shaped effect in the models contrasting the answers "Probably/Surely yes" versus "Probably/Surely not" and "Surely yes" versus all the other options. This denotes a greater propensity to move for those with a lower and those with a higher level of education. The former would probably move out of necessity whereas the latter to obtain the most out of their human capital [1, 9].

The gender and the age of respondents also play a significant role in determining the intentions of moving abroad: females are less likely than males to show positive moderate intentions to move abroad, even if gender specificity does not mark a difference in determining who is surely intentioned to move (see model 4 vs. 1,2, 3 in Table 5). With respect to the effect of the respondents' age, people belonging to the older cohorts are less likely to move. Probably in this case we refer to individuals who have already completed the main steps of their transition to adulthood and that are therefore less inclined to start over in another country.

Other relevant variables in determining the intentions of moving abroad are the country of origin and past experiences in other countries.

Previous studying or working experiences abroad is positively related to a renewed mobility in the short run. It is also relevant the stage of training or professional where someone is. In particular, those who study tend to give greater consideration to the foreign option than those who already have a job. Finally, keeping all these factors fixed, it emerges how significant the country of residence actually is. Among the countries considered, Italians give greater importance to mobility to find a job, both within and outside the national borders.

Limiting the analysis to the Italian sample (Table 6) broadly confirms the results of the general model, even if the small sample size (1000 units vs. 5000 units of the pooled model) reduces the significance of some variables.

Nevertheless, the perception of Italy as a country characterized by few opportunities with little confidence in improvements is positively associated with the intentions of leaving.

The inclusion of a territorial breakdown (displayed in the output of Table 6) shows a propensity for higher mobility in the South in almost all the models. Nonetheless, actual behaviors may in fact deviate from intentions and plans. And, as data from other researches confirm, it is easier to turn intentions into actual behavior for those individuals with greater cultural and economic resources. This dramatically leads to an accentuation of human capital loss.

Finally, our results show that there is a higher propensity to move abroad for those who live in contexts which are perceived as less dynamic, more lacking in opportunities, and with less prospects for improvement. And Italy is the country where this phenomenon is more relevant.

4 Concluding Remarks

Exploiting an original dataset, we provide additional evidence on the propensity to leave the native country by young Europeans and show how this depends on the economic opportunities offered by the countries of origin. In more detail, Italy is the country showing, net of the effect of all individual predictors included in the analysis, the highest propensity of young adults desiring to leave. This is particularly true for younger cohorts, who seem to be more vulnerable in the early stages of working careers and that are also the most dynamic social category, with less obstacles in programming their future. However, the most valuable predictors for the choices of younger European generations are the expectations on the future development of the country of origin: people that show little confidence in future developments of the country of residence are more likely to show a positive intention of moving away. On the contrary, a positive view is associated to a lower risk of migrating.

References

1. Ackers, L.: Moving people and knowledge: scientific mobility in the European union. Int. Migr. **43**(5), 100–131 (2005)
2. Balduzzi, P., Rosina, A.: I giovani italiani nel quadro europeo: la sfida del "degiovanimento". Ricercazione **2**(2), 201–214 (2010)
3. Balduzzi, P., Rosina, A.: Le ragioni della rottamazione (2010). http://archivio.lavoce.info/articoli/pagina1002006.html
4. Balduzzi, P., Rosina, A.: Studio e lavoro senza confine: generazione mobile. In: La Condizione Giovanile in Italia. Rapporto Giovani 2016 . Istituto Giuseppe Toniolo (ed.), pp. 157–182. Il Mulino, Bologna (2016)
5. Biondo, A.E., Monteleone, S., Skonieczny, G., Torrisi, B.: The propensity to return: Theory and evidence for the Italian brain drain. Eco. Lett. **115**(3), 359–362 (2012)
6. Beine, M., Docquier, F., Rapoport, H.: Brain drain and economic growth: theory and evidence. J. Dev. Econ. **64**(1), 275–289 (2001)
7. Bertoli, S., Brücker, H.: Extending the case for a beneficial brain drain. J. Econ. Stat. **231**(4), 466–478 (2008)
8. Bhagwati, J.N., Hamada, K.: The brain drain, international migration of markets for professionals and unemployment. J. Dev. Econ. **1**(1), 19–42 (1974)
9. Cairns, D.: I wouldn't stay here: economic crisis and youth mobility in Ireland. Int. Migr. **52**(3), 237–249 (2012)
10. Carrington, W. J., Detragiache, E.: How big is the brain drain? IMF working paper 98/102 (1998)
11. Docquier, F., Lowell, B.L., Marfouk, A.: A gendered assessment of highly skilled emigration. Popul. Dev. Rev. **35**(2), 297–321 (2009)
12. Docquier, F., Rapoport, H.: Quantifying the Impact of Highly-Skilled Emigration on Developing Countries. PEGGED Policy Report n. 1 (2009)
13. Istituto Toniolo: La condizione giovanile in Italia. Rapporto giovani 2016. Il Mulino, Bologna (2016)
14. Saxenian, A.: From brain drain to brain circulation: transnational communities and regional upgrading in India and China. Stud. Comp. Int. Dev. **40**(2), 35–61 (2005)

15. Stark, O., Helmenstein, C., Prskawetz, A.: A brain drain with a brain gain. Econ. Lett. **55**(2), 227–234 (1997)
16. Williams, R.: Understanding and interpreting generalized ordered logit models. J. Math. Sociol. **40**(1), 7–20 (2016)

New Insights on Student Evaluation of Teaching in Italy

Francesca Bassi, Leonardo Grilli, Omar Paccagnella, Carla Rampichini and Roberta Varriale

Abstract This work focuses on the relationship between student evaluation of teaching and student, teacher and course specific characteristics, exploiting the richness of information collected by a new survey carried out among professors of the University of Padua. Data collected in this survey are able to highlight teacher needs, beliefs and practices of teaching and learning. This allows to introduce in the study some *subjective* traits of the teachers. The role of these new variables in explaining student evaluations is deeply investigated.

Keywords Multilevel modelling · Multi-source data · Student ratings · Teacher opinions

1 Introduction

Students' opinions and judgements of teaching performance play a substantial role in higher education, particularly as instruments for gathering information on the quality of education and evaluating university courses [5, 25]. The relationship between

F. Bassi · O. Paccagnella (✉)
Department of Statistical Sciences, University of Padua, Padua, Italy
e-mail: omar.paccagnella@unipd.it

F. Bassi
e-mail: bassi@stat.unipd.it

L. Grilli · C. Rampichini
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, Italy
e-mail: leonardo.grilli@unifi.it

C. Rampichini
e-mail: carla.rampichini@unifi.it

R. Varriale
ISTAT (Italian National Statistical Institute), Rome, Italy
e-mail: varriale@istat.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*, Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_20

student-, teacher-, course-specific characteristics and Student Evaluation of Teaching (SET) is the topic of a huge amount of works in the literature (see an extensive review provided by Spooren et al. [21]). It is also generally accepted that a multilevel analysis of the students' ratings is a satisfactory approach for investigating teaching evaluations, because of the hierarchical nature of the data, such as university students nested into classes [16, 17, 20].

However, findings concerning the relationship between SET and the characteristics of courses, students and teachers are sometimes contradictory [24]. Indeed, these characteristics usually explain only a small portion of the total variance in SET scores [1, 20].

On the one hand, there is a branch of the literature that has investigated the psychometric properties of the evaluation questionnaires, concluding that SET is reasonably good in terms of reliability and validity of measurement; these researches also highlight a positive correlation between SET measures and other indicators of teaching quality, such as student achievement, alumni ratings, and so on [21]. On the other hand, a growing amount of works criticises the use of course evaluations as a measure of teaching quality. First, there is no or minimal significant correlation between SET ratings and student learning [23]; second, SET causes grade inflation and lowering of academic standards [6, 11]; third, SET may be biased, because of the presence of some factors—not necessarily related to the teaching quality—that affects the student evaluation [3, 9]; in the end, SET has some intrinsic limitations, because students "can only evaluate what they can observe, and what they observe is mainly what occurs inside the classroom. But as stated previously, there are other very important components of teaching, such as course quality, instructor knowledge, quality of assignments, and curriculum development that cannot be measured by student ratings, and need to be assessed in some other way" [15].

Despite this discussion, the general consensus on the quality of teaching issue is the influential role played by the teachers, even if their most common observable characteristics often reveal weak effects [10]; furthermore, teacher quality may differ in many ways, this is not captured by observable qualifications or experience [18]. How to measure *teacher quality* is nevertheless a complicated task. By means of a comprehensive review, Goe et al. [8] stress the advantages of collecting teacher self-report methods: such data "can tap into a teacher's intentions, thought processes, knowledge, and beliefs better than other methods" (p. 38); then, teachers have the full knowledge of their abilities, classroom context, and so on; moreover, self-reported measures exhibit a positive trade-off between amount of retrieving information and cost of collecting them. But, collected teacher self-reported data substantially describe instructional practices [2, 12, 14], while they usually lack information on the teacher's beliefs and needs.

This work aims at filling in this gap, enriching the multilevel literature on the student evaluation of teaching proposing some original analyses based on a wider set of teacher-specific characteristics, including particularly teachers' opinions on their teaching activities. Indeed, this work exploits an innovative and original dataset available at the University of Padua, obtained through the linkage of survey and administrative data coming from three different sources: first, the conventional survey

on the student evaluation of teaching carried out among university students; second, administrative data related to the main features of the teachers and the Didactic Activities (DAs) they are involved in; third, a new CAWI survey carried out by means of the research project PRODID (Teacher professional development and academic educational innovation). This new survey started at the University of Padua in 2013, with the aim of developing strategies to support academic teachers and enhance their teaching competences. A specific questionnaire was then developed and addressed to all professors involved in almost all didactic activities of the University. This new survey collected opinions, beliefs and needs of the professors, with regard to their teaching activities developed in their classes.

This work is organised as follows. Section 2 introduces the data of this analysis, while Sect. 3 describes the empirical application (model specification and results). Section 4 ends the paper, highlighting the main conclusions and some suggestions for future works.

2 The Dataset

This work investigates data obtained by merging three different datasets coming from the University of Padua. The reference is the 2012–2013 academic year.

The first one is the standard online survey carried out by the University to measure students' opinions on the didactic activities. It involves all students who have been attending lessons of any degree courses of the Athenaeum. Students were asked to express their level of satisfaction on a scale from one to ten (being one the lowest level) to a set of 18 items (seven if the student attended less than 30% of the lessons).

The second one is the administrative dataset that collects information on the teachers and the didactic activities of all Padua academic institutions (the educational offer).

The third one is an innovative dataset, collected by means of a new online survey aiming at providing a picture of the teaching experiences developed in the university classrooms. Indeed, the University of Padua in 2013 promoted the PRODID project (Teacher professional development and academic educational innovation—in Italian "Preparazione alla professionalità docente e innovazione didattica") with the purpose of developing an integrated system to improve teaching competences and academic innovation. The PRODID project promoted a research-based approach to creating training programs, faculty learning communities, pilot experimental contexts where teaching innovation could be tested and monitored [7]. Following an evidence-based approach, the project aimed at highlighting teachers' needs, beliefs and practices of teaching and learning, which may constitute a privileged context for the development of innovative teaching activities within the institution.

The final questionnaire was developed according to the Framework of Teaching of Tigelaar et al. [22] and is composed by three sections. The first section focuses on *practices* developed by the Padua professors in their teaching activities. The teacher is thought as a facilitator of the learning processes and for this reason the section asks

for each DA (at most three, having a minimum of four University Educational Credits each) about the application (or not) of some specific practices in his/her activities. Eight items are collected. Six indicators are then constructed and five of them are obtained considering separately as dummy variables the first five items:

1. implementation of practices for actively getting involved students;
2. proposal of external contributions (i.e. stakeholders);
3. monitoring students learning during the course by means of specific tests/other ways;
4. assessment of students learning using various types of exams;
5. modification of teaching practices according to SET.

The sixth indicator is calculated summarising in a single dummy variable the last three items of the section (6. reporting at least one activity involving technology practices), since these three questions collect similar information on these practices.

The second section deepens teachers' *beliefs* about teaching in higher education. Differently from the first section, here the focus is on the person, with his/her way of thinking; therefore, regardless of the number of the DAs, only one set of answers is collected. By means of 20 questions, in a scale from one (fully disagree) to seven (fully agree), some general dimensions are investigated: the Person as Teacher, the Expert on Content Knowledge, the Facilitator of Learning Processes and the Scholar/Lifelong Learner. Considering also some questionnaire validation analyses (a factor analysis in particular), six factors are defined (they substantially replicate the aforementioned dimensions), computed as the average values of the answers within each factor. These measures may be summarised as a second group of subjective characteristics of the teachers:

1. passion for teaching;
2. passion for research;
3. feeling the need of support for improving teaching activities;
4. will to change teaching activities according to students needs;
5. features of teaching and learning methods (e-learning, English language and so on);
6. features of teaching and evaluation activities.

The third section focuses on teachers' *needs*, that are collected through some open-ended questions (however, they are not exploited in this analysis).

The PRODID questionnaire was addressed to all teaching staff of the University of Padua involved in any DA during the academic year 2012–2013; the response rate of this survey was slightly lower than 50%.

Further information on the questionnaire administration and item contents is available in [4].

3 The Empirical Application

3.1 The Model

The analysis of the dataset described in Sect. 2 is based on the estimation of a multilevel random intercept model [19]:

$$y_{ij} = \alpha + \beta' x_{ij} + \gamma' w_j + u_j + \varepsilon_{ij} \tag{1}$$

where $j = 1, \ldots, J$ indexes the level-2 units and $i = 1, \ldots, n_j$ the level-1 units, α is the intercept, β and γ are vectors of parameters to be estimated, while ε_{ij} and u_j are i.i.d. errors terms, distributed as:

$$\begin{aligned} \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \\ u_j &\sim N(0, \sigma_u^2) \\ \varepsilon_{ij} &\perp u_j \qquad \forall i, j \end{aligned}$$

The proportion of residual variance due to unobserved between-group factors is given by the Intraclass Correlation Coefficient (ICC):

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \tag{2}$$

The dependent variable y_{ij} is the overall level of satisfaction, measured on a scale 1 to 10 (1 for completely unsatisfied students and 10 for completely satisfied students). The student ratings are level-1 units, while the didactic activities (DAs) associated to the teachers are the level-2 units. For each course, the student is asked to evaluate the activities of each professor having a minimum number of hours taught in the course. The student degree programme is not considered as a further hierarchical level, but it is modelled as fixed effects.

In general, the rating of a student to a given item for a certain course may depend on course-related factors (class size and heterogeneity, course difficulty and so on), student-related factors (gender, age and so on) and teacher-related factors (age, gender, personal traits and so on) [21]. According to the aims of this work and the original features of our dataset, the set of considered explanatory variables may be divided in three groups:

1. Student (level-1 covariates)
   - Demographics: gender, age.
   - University career: year of enrolment, average (per year) number of passed exams in the whole career, average grade of the passed exams in the considered academic year.

2. Didactic Activity (level-2 covariates)
   - Course characteristics: compulsory course, total number of hours, more than one teacher involved in the DA, location (in Padua or outside), shared course (i.e. students belonging to different degree programs).
3. Teacher (level-2 covariates)
   - Demographics (objective characteristics): gender, age.
   - University career (objective characteristics): academic position (full professor, associate professor and so on).
   - Practices (subjective characteristics): according to Sect. 2, the six indicators of teaching practices.
   - Beliefs (subjective characteristics): according to Sect. 2, the six factors of teacher beliefs.

This model specification allows to investigate the role of both *objective* and *subjective* teacher characteristics.

3.2 Some Descriptive Statistics

In this analysis we consider only ratings expressed by students attending at least 50% of lessons. The considered Didactic Activities belong to bachelor degree programmes (3-years undergraduate degrees), except those of the Medical school. We excluded Didactic Activities with a number of student ratings smaller than five, in order to avoid comparisons based on too few ratings.

According to these criteria, the linkage of the different sources led to a final dataset composed by 29175 complete records, corresponding to student ratings. The total number of level-2 units is equal to 548, with an average number of observations per group of about 53 (ranging from 5 to 371); this value of level-2 units results from 450 DAs (nested in 69 degree programmes) and 472 teachers.

The students in the dataset are 52.3% females, 20.5 years old on average. Half of them are enrolled to the first year of their degree programme, while less than 4% are not regularly enrolled. In the considered academic year, on average these students passed about six exams, with an average grade a bit larger than 25 out of 30.

Figure 1 shows the distribution of the dependent variable (the overall student satisfaction) in the analysed sample. This variable is characterised by a left-skewed distribution, about one fourth of respondents reports an evaluation equals to eight; the positive ratings (i.e. from six to ten) are nearly 90% of all evaluations. However, there is a large heterogeneity across degree programmes.

The sample of professors is mainly composed by males (63.6%) and are 49.8 years old on average. The majority of them are assistant professors (about 39%), while the role of full professors refers to exactly a quarter of the teacher sample.

Figure 2 reports the proportion of teachers who claim to have experienced the specific practice in his/her activities. It is interesting to note that teachers did some

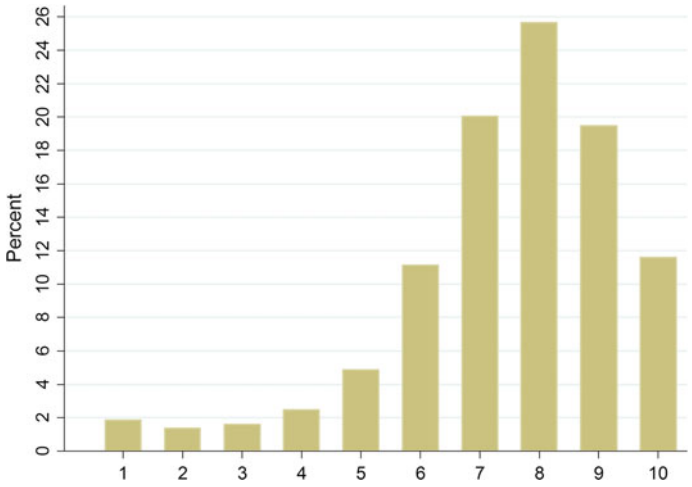


Fig. 1 Distribution of the SET ratings in the analysed sample

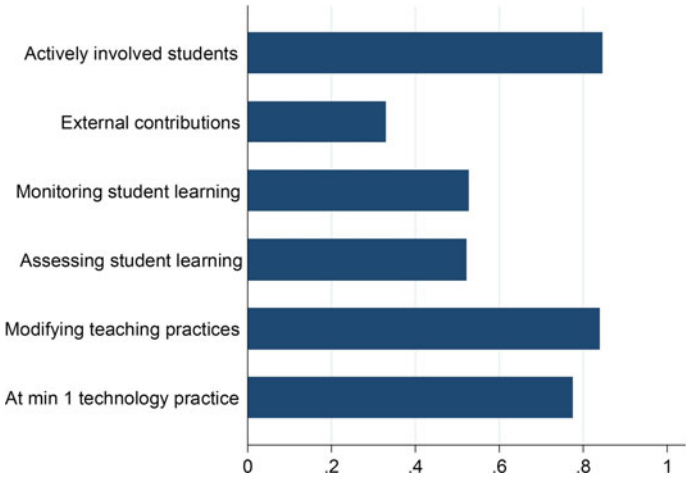


Fig. 2 Distribution of the indicators obtained from the *Practices* section of the PRODID questionnaire

modifications of their teaching practices according to the SET evaluations collected in the previous academic years in more than 80% of the DAs. It is also quite large the number of DAs where some technology practices are exploited by the professors. On the other hand, the use of external contributions for improving teaching activities is still rather low.

Table 1 summarises the main features of the teacher beliefs and needs, according to their self-evaluations collected in the second part of the PRODID questionnaire. Not surprisingly, teachers highlight a high level of passion for research (more than

Table 1 Descriptive statistics of the subjective teacher factors obtained from the *Beliefs* section of the PRODID questionnaire

Subjective teacher covariate	Median	Mean ± s.d.
Passion for teaching	6	5.71 ± 1.19
Passion for research	7	6.31 ± 1.07
Feeling the need of support for improving teaching activities	4.25	4.16 ± 1.74
Will to change teaching activities according to students needs	5	4.70 ± 1.27
Features of teaching and learning methods	4.67	4.63 ± 1.17
Features of teaching and evaluation activities	4.50	4.55 ± 0.80

Note Each distribution ranges from 1 (fully disagree) to 7 (fully agree)

half of them indicate a fully agreement with these statements), but also the passion of teaching appears very important in this sample of professors. The other four factors are characterised by much more symmetric distributions and show the presence of relevant proportions of disagreed answers (particularly strong for the beliefs that summarised the need of a support to improve the teaching activities).

3.3 Main Results

Results from the estimation of model (1) are reported in Table 2. Several model specifications have been used, gradually adding some sets of covariates to the null model (i.e. the model where only the intercept is specified, which shows an ICC equal to 24.3%), in order to investigate and highlight the role of level-2 information. As specified in Sect. 2, fixed effects for the student degree programmes are introduced in all estimated models.

In column A we control for student characteristics only: all estimated parameters are significant, but gender. It is worth noting the positive relationship between SET and grade, as well documented in the literature [6]. However, the ICC value is close to 25%, which indicates the presence of a very large between-group heterogeneity.

From model B to E, sets of level-2 covariates are introduced in the model one by one and the role of course and teacher characteristics for explaining such heterogeneity is therefore investigated. However, in all of these models, parameter estimates of the student variables do not change, both in magnitude and in significance, as well as the estimation of the level-1 variance.

Course characteristics reveal very low effects in explaining level-2 variability (from model C to E): only one parameter is weakly (and negatively) significant, that is the presence of more than one teacher involved in the DA. Similar conclusions may be reached when *objective* traits of the teachers are added to the model specification: the age of the professors is the only variable reporting a statistically significant estimate (the older the teacher, the worse he/she is evaluated, ceteris paribus). No difference appears according to the academic position. Overall, the introduction

Table 2 Estimates of the random intercept models on the students' overall satisfaction

Variable	Model				
	A	B	C	D	E
Student characteristics					
Female	−0.036	−0.036	−0.036	−0.036	−0.037
Age	0.297***	0.296***	0.297***	0.297***	0.297***
First year of enrolment	0.098*	0.127**	0.122**	0.132**	0.131**
Being regularly enrolled	−0.160**	−0.149**	−0.152**	−0.148**	−0.147**
Average number of passed exams	0.107***	0.106***	0.106***	0.105***	0.105***
Average grade of passed exams	0.397***	0.398***	0.398***	0.398***	0.397***
Course characteristics					
Compulsory course		−0.044	−0.040	−0.026	−0.035
Number of hours		0.052	0.098	0.172	0.233
More than one teacher		−0.164**	−0.182**	−0.182**	−0.171**
Location of courses in Padua		−0.389	−0.379	−0.365	−0.409
Shared course		−0.103	−0.102	−0.112	−0.093
Teacher characteristics					
Female			−0.112	−0.134	−0.089
Age			−0.219***	−0.216***	−0.179***
Full professor			0.079	0.090	−0.025
Associate professor			0.083	0.066	0.017
Teacher practices					
Actively getting involved students				−0.003	0.019
Proposal of external contributions				0.206**	0.185**
Monitoring students learning ongoing				−0.014	−0.039
Assessing learning using different exams				−0.214**	−0.257***
Modification of practices according to SET				−0.047	0.007
Reporting at least 1 activity on technology				0.118	0.046
Teacher beliefs					
Passion for teaching					0.134***
Passion for research					−0.045
Need support to improve teaching activities					−0.120***
Changing activities with student needs					0.074*
Features of teaching and learning methods					0.166***
Features of teaching and evaluation activities					−0.032
constant	6.196***	6.638***	7.737***	7.614***	6.507***
Level-2 variance (σ_u^2)	0.876	0.859	0.823	0.806	0.731
Level-1 variance (σ_ε^2)	2.701	2.701	2.702	2.702	2.702
ICC	24.5%	24.1%	23.4%	23.0%	21.3%

Note *** = 1% of level; ** = 5% of level; * = 10% of level

of several level-2 variables capturing course and *objective* teacher characteristics enables to reduce the ICC by about 1% only.

Findings are different when *subjective* teacher features are taken into account: two indicators of *practices* and even four factors of *beliefs* are statistically significant. According to practices, it is worth noting that, ceteris paribus, students positively evaluate the DA when external contributions are proposed, while they judge less favourably the use of many types of exams to assess their learning. However, these two indicators are also characterised by the least diffusions among teachers.

The set of teacher beliefs comprises the most important level-2 covariates, particularly those related to the sensitivity and the aptitude of teaching. For instance, according to the PRODID questionnaire the factor "Feeling the need of support for improving teaching activities" may highlight those teachers who feel some difficulties or inadequacies in their teaching activities/performances and for this reason they need help from experts. Students are able to perceive such difficulties and then reporting a lower evaluation of the course (other things being equal). On the other hand, students recognise those teachers with a high passion for teaching or the will to propose suitable and helpful instruments in their DAs to improve the student learning: such traits may be able to enhance the transmission of knowledge from the teacher to the student. Therefore, their evaluations are higher, ceteris paribus. Even if weak, it is important to underline the positive relationship between SET and the teachers' willingness at changing their teaching activities according to the needs of the students. It is worth noting two other interesting findings: i) the different role that comes to light between the *passion for teaching* and the *passion for research* dimensions in explaining SET evaluations; ii) the introduction of the *beliefs* covariates seems to lead to some non trivial changes in the estimation of some *objective* and *subjective* (practices) teacher characteristics.

Summing up, student characteristics are strongly associated with the overall satisfaction rating of the DA, particularly those related to the academic experience of these students. On the other hand, the main features of the courses play no or a weak role. Instead, there are some interesting results on the relationship between SET and teacher characteristics: *objective* teacher' traits are not related to SET ratings, while *subjective* features of the teachers disclose a stronger role in explaining SET ratings.

After controlling for a large number of level-2 covariates, there is still evidence of a high between-group variance (the ICC is always larger than 20%). Overall, this finding may support the claim that the SET is a biased measure of the DA quality, as highlighted in the Introduction: the psychological literature has widely documented the phenomenon of over-reporting individual self-assessments, because of the tendency of presenting themselves in a more favourable light [13]. However, in the educational literature the extent of this potential bias is still to determine. Moreover, on the basis of the fixed effect estimates, there are significant differences among some degree programmes.

4 Conclusions

Exploiting the richness of information provided by an innovative survey on teaching experiences and beliefs of professors working at the University of Padua, the role of the teacher perceptions and needs on their DA evaluations is deeply investigated. Findings clearly show that *subjective* characteristics of the teachers play an important role in explaining SET ratings.

This work may be seen as a first step for enhancing the relationship between quality of a course (or university) and students' opinions. Indeed, teaching is a complex and multidimensional concept, so a future research strand could be the analysis of a multidimensional indicator of course quality, based on a battery of items.

Moreover, results should be improved taking into account the missing data problem affecting the sample of professors: because of the unit non-response phenomenon in the data collection of the PRODID project, about half of the students' ratings cannot be used in the current analysis. This reduction in sample size severely reduces the power of statistical tests. Future research will investigate how to impute the missing values originated by teacher non-responses.

Acknowledgements The authors thank two anonymous referees for their helpful comments. This work was supported by the University of Padua through the project SID2016 "Advances in Multilevel and Longitudinal Modelling".

References

1. Beran, T., Violato, C.: Ratings of university teacher instruction: how much do student and course characteristics really matter? Assess. Eval. High. Educ. **30**, 593–601 (2005)
2. Blank, R.K., Porter, A., Smithson, J.: New Tools for Analyzing Teaching, Curriculum and Standards in Mathematics & Science. Results from Survey of Enacted Curriculum Project Final Report, Council of Chief State School Officers, Washington, DC (2001)
3. Centra, J.A., Gaubatz, N.B.: Is there gender bias in student evaluations of teaching? J. High. Educ. **71**, 17–33 (2000)
4. Dalla Zuanna, G., Clerici, R., Paccagnella, O., Paggiaro, A., Martinoia, S., Pierobon, S.: Evaluative research in education: A survey among professors of University of Padua. Excel. Innov. Learn. Teach. **1**, 17–34 (2016)
5. Emerson, J.D., Mosteller, F., Youtz, C.: Students can help improve college teaching: A review and an agenda for the statistics profession. In: Rao, C.R., Székely, G.J. (eds.) Statistics for the 21st Century: Methodologies for Applications of the Future. Marcel Dekker, New York (2000)
6. Ewing, A.M.: Estimating the impact of relative expected grade on student evaluations of teachers. Econ. Educ. Rev. **31**, 141–154 (2012)
7. Felisatti, E., Serbati, A.: The professional development of teachers: from teachers practices and beliefs to new strategies at the university of Padua. In: Proceedings of the ICED conference Educational development in a changing world, Stockholm, 16–18 June 2014 (2014)
8. Goe, L., Bell, C., Little, O.: Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. National Comprehensive Center for Teacher Quality, Washington (2008)
9. Goos, M., Salomons, A.: Measuring teaching quality in higher education: assessing selection bias in course evaluations. Econ. Educ. Rev. **58**, 341–364 (2017)

10. Hanushek, E.A., Rivkin, S.G.: Teacher quality. In: Hanushek, E.A., Welch, F. (eds), Handbook of the economics of education, vol. 1, pp. 1050–1078. North Holland, Amsterdam (2006)
11. Johnson, V.E.: Grade Inflation: A Crisis in College Education. Springer, New York (2003)
12. Mayer, D.P.: Measuring instructional practice: can policymakers trust survey data? Educ. Eval. Policy Anal. **21**, 29–45 (1999)
13. Moorman, R.H., Podsakoff, P.M.: A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. J. Occup. Organ. Psychol. **65**, 131–149 (1992)
14. Mullens, H.G.: Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Followup Survey, NCES Working Paper No 95–15, National Center of Education Statistics Washington (1995)
15. Murray, H.G.: Student Evaluation of Teaching: Has It Made a Difference? Paper presented at the Annual Meeting of the Society for Teaching and Learning in Higher Education, Charlottetown, Prince Edward Island (2005)
16. Rampichini, C., Grilli, L., Petrucci, A.: Analysis of university course evaluations: from descriptive measures to multilevel models. Stat. Methods Appl. **13**, 357–373 (2004)
17. Rantanen, P.: The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students evaluation of teaching. Assess. Eval. High. Educ. **38**, 224–239 (2013)
18. Rivkin, S.G., Hanushek, E.A., Kain, J.F.: Teachers, schools, and academic achievement. Econometrica **73**, 417–458 (2005)
19. Snijders, T.A.B., Bosker, R.J.: Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling. Sage, London (2012)
20. Spooren, P.: On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. Stud. Educ. Eval. **36**, 121–131 (2010)
21. Spooren, P., Brockx, B., Mortelmans, D.: On the validity of student evaluation of teaching: The state of the art. Rev. Educ. Res. **83**, 598–642 (2013)
22. Tigelaar, D.E.H., Dolmans, D.H.J.M., Wolfhagen, I.H.A.P., Van Der Vleuten, C.P.M.: The development and validation of a framework for teaching competencies in higher education. High. Educ. **48**, 253–268 (2004)
23. Uttl, B., White, C.A., Wong Gonzalez, D.: Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Stud. Educ. Eval. **54**, 22–42 (2017)
24. Wachtel, H.K.: Student evaluation of college teaching effectiveness: a brief review. Assess. Eval. High. Educ. **23**, 191–212 (1998)
25. Zabaleta, F.: The use and misuse of student evaluations of teaching. Teach. High. Educ. **12**, 55–76 (2007)

Eurostat Methodological Network: Skills Mapping for a Collaborative Statistical Office

Agne Bikauskaite and Dario Buono

Abstract Collaboration, interaction and exchange of knowledge among staff are important components for development and enriching of the scientific intelligence within a statistical office. Eurostat methodological network has been built as a skills mapping tool aiming identify in-house competencies for innovation and affordability of diffusion of knowledge, promotion of collaboration on methodological issues, and processes within statistical office. In this exercise we mainly focus on staff knowledge and working and academic experience on statistics and econometrics. Quantitative network analysis metrics are used to measure the strengths of methodological competencies within Eurostat, to identify groups of people for collaboration in providing results on specific tasks, or characterise areas that are not fully integrated into methodological network. By combining network visualisation and quantitative analysis, we able easily assess competency level for each dimension of interest. Network analysis helps us in making decisions related to improvement of staff communication and collaboration, by building mechanisms for information flows, filling competency gaps. Data represented as mathematical graph makes readily visible general view, absorbs its structure, permits us to focus on persons, competencies and relations between them. Modernisation of ways of working leads to a more cost effective use of resources.

Keywords Complex network · Data analysis · Network visualisation · Bipartite graph · Network projection · Ego network · Network analysis

A. Bikauskaite · D. Buono (✉)
Eurostat, European Commission, 5, Rue Alphonse Weicker,
2721 Luxembourg City, Luxembourg
e-mail: dario.buono@ec.europa.eu

A. Bikauskaite
e-mail: agne.bikauskaite@ec.europa.eu

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_21

1 Introduction

Collaboration, interaction and exchange of knowledge among staff are important components for development and enriching of scientific intelligence within a statistical office, especially when this exchange happens across areas of interest by both interacting sides. Methodological network has been built as a skills mapping tool aiming identify in-house competencies for innovation and affordability of diffusion of knowledge and information, and promotion of collaboration on methodological issues and processes within organisation. We mainly focus on staff knowledge and working and academic experience in methodological areas, domains and tools on statistics and econometrics. This paper gives a set of mathematical network analysis measures used for the analysis, from basic ones as size and degree to more complex as clustering coefficient and their correlation with degree that evaluates and makes better understandable the methodological knowledge network structure. Those metrics help us to:

- detect specific network communities,
- identify the facility in knowledge sharing and contribution in providing support on emerging and changing needs of methodological tasks that to fulfil methodological objectives,
- characterise critical methodological areas and dimensions that are not fully integrated into network,
- build tools for accessibility, efficient exchange, innovation of existing skills and performance of methodological tasks promotion,
- bring people competent in the same area into contact,
- map experts on statistical competencies and establish a methodology for easy distribution of information across the organisation and knowledge network development.

Each individual accumulates new knowledge in two ways: through a process of individual learning; and/or through a process of interactive learning. Studying the structure of the networks formed may be a way to know more in depth the availability of competencies and possible knowledge diffusion processes. Network analysis calculations and visualisations obtained using the `R` packages `igraph`, `tnet`, `bipartite`, `shiny`, and some additional functions developed specially for this particular exercise.

2 Survey Methodology

Aiming to map existing methodological skills within Eurostat that to increase productivity and to set up to function Methodological network the "Eurostat methodological skills—staff survey" has been conducted. The survey was open to all Eurostat staff on a voluntary basis. The questionnaire focused on staffs' knowledge in statistics

and econometrics. The respondents were inquired to indicate up to five methodological areas, three statistical domains and three tools in which they could contribute fulfilling specific tasks or while sharing information with others.

The findings presented in this paper are derived from input of 67 respondents, identified as population of Eurostat methodological network. Data presented in this paper has been anonymised to each respondent giving ID from 1 to 67 seeking to assure confidentiality.

3 Network Analysis Methods

The purpose of the conducted survey is to build Eurostat methodological network and to highlight important dimensions through the network analysis techniques. Quantitative network metrics are used to measure the strengths of Eurostat methodological network members' competencies, to identify groups of people for collaboration in providing results on specific tasks, and characterise areas that are not fully integrated into methodological network.

3.1 Bipartite Graphs

Network data consists of a set of elements with relations on those elements and it may be represented as a graph. Our research subjects, individuals, form links that characterise their competencies in statistics and econometrics. Formally we have graph $G = (V, E)$, where G is a relational structure consisting of set of vertices V and set of edges E [2]. We say that a graph is bipartite when the vertex set V is divided into two finite, disjoint $V_1 \cap V_2 = \emptyset$ sets [4]. When V_1 composed of the first mode vertices and V_2 of the second mode vertices, we have the bipartite graph $G = (V_1, V_2, E)$ where ties map the elements of different modes only.

3.2 Incidence Matrix

The two basic parameters of graph are the number of vertices and the number of edges [1]. $n_{V_1} = |V_1|$ and $n_{V_2} = |V_2|$ are the numbers of vertices in the first and in the second sets respectively, where $n = n_{V_1} + n_{V_2}$ is a number of full set of vertices in the graph G and is defined as size of the network. $m = |E|$ gives the number of edges (links).

A bipartite graph data is represented in the form of incidence matrix, which allows mathematical calculations that to summarise the information of the graph. In our particular case data are arranged as person by skill matrix, where the rows

correspond to methodological network members and the columns to the dimensions of statistical competencies. We obtain binary matrix $\mathbf{A}$ of size $n_{V_1} \times n_{V_2}$:

$$\mathbf{A}_{ij} = \begin{cases} 1, & \text{if person } i \text{ has methodological skill } j \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

Simply by looking at the columns, we can see who is similar to whom in terms of having common knowledge and scientific interest.

3.3 Mathematical Measures

In order to understand network and its structure, network analysis statistical models have been employed in this study. Centrality is a family of concepts characterising the structural importance of a vertex position in a graph (see Table 1). The most important descriptive statistics of centrality is degree. This measure shows how central node i is in the network and is defined as the number of edges formed to it. This metric helps to identify the most known competencies, and to diagnose the knowledge gaps within the methodological network. Also helps coordinating the work, perform raised methodological issues by available internal resource and in the end notify the tasks for which external support would be required.

Competencies degree ranges from 0 to 39, with a mean of 10.2 for Eurostat methodological network, what indicates, that a certain competence is chosen by 10 respondents in average. The minimum degree 0 indicates isolated vertices, that do not have any links within the network. We found out that in the Eurostat methodological network exists one isolated vertex, which belongs to methodological area Micro-data access, what means that within Eurostat methodological network significant lack of experts in Micro-data access have been observed.

Degree sequence of statistical competencies points that Eurostat methodological network members are mostly familiar to Data Analysis, very well competent in Social Statistics domain and experienced in R statistical software. While the gap exists of people knowledgeable on Micro-data access and Statistical confidentiality, experienced in Transport and Energy statistics, and capable on Hadoop tool. Other competencies are more or less covered and known by Eurostat methodological network members. The degree sequence of statistical competencies is depicted in the Fig. 1.

The average degree of vertex sets V_1 and V_2 is commonly used summary of how well connected the network is and defined as proportion of number of links and number of nodes $k_{V_b} = \frac{m}{n_{V_b}}$, where $b = 1, 2$.

While the average degree of overall network is obtained from the total numbers of nodes and edges by $k = \frac{2m}{n_{V_1} + n_{V_2}}$.

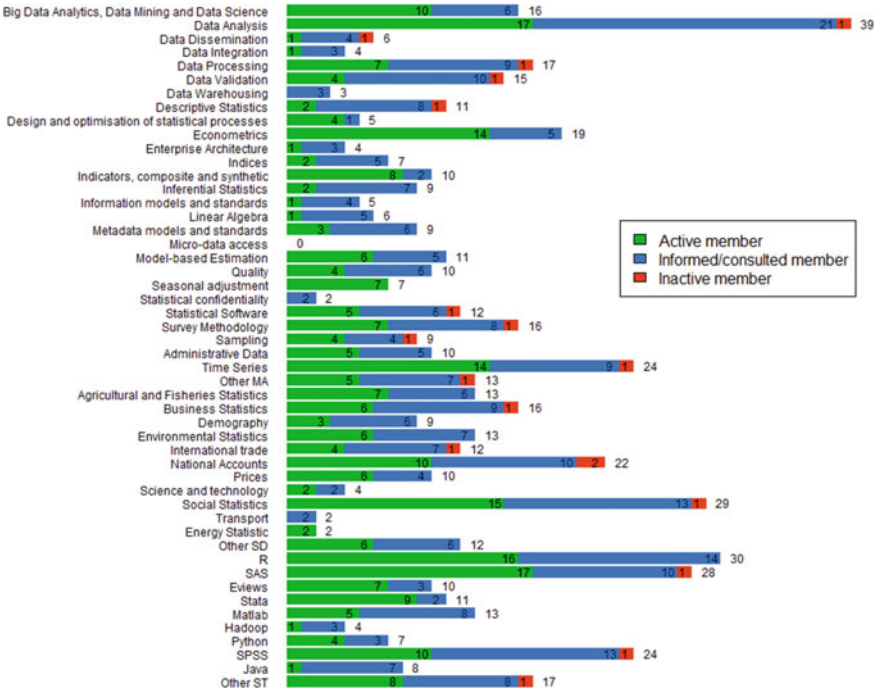


Fig. 1 The degree sequence of statistical competencies

The density δ of the bipartite graph G measures average ratio of the actual degree of the nodes in the network and the maximum possible degree, which corresponds to the number of nodes in the set of different mode nodes $\delta(G) = \frac{m}{n_{V_1} n_{V_2}}$.

This index is equal to 1 in the fully connected case (i.e. G has one component) and takes value of 0 when G is fully disconnected (i.e. G is composed entirely of isolates). Density can be interpreted as the marginal probability of an edge from any given vertex of individuals to any vertex of competencies.

For Eurostat methodological network data the standard density measure gives a value 0.17, which shows a fairly sparse network with presence of 17% of the possible links for average node. However in this particular case the standard denominator is clearly not appropriate defining methodological network members' competencies. Due to restriction of choice of maximum 11 dimensions out of 50 possible, it cannot be interpreted as actual possible density. Using modified denominator, our network has density 0.79, which tells us that respondents' competences level is high.

One of the most important properties of the network is the clustering coefficient which concerns link correlation. The clustering coefficient of a node i is the proportion of links between the nodes within its neighborhood divided by the number of edges that could be possibly exist between the nodes. The equation of it is $cc_{ijl} = \frac{q_{ijl}}{(k_j - \eta_{ijl}) + (k_l - \eta_{ijl}) + q_{ijl}}$ where j and l are a pair of neighbors of node i, q_{ijl} is the

number of squares which include these three nodes, and $\eta_{ijl} = 1 + q_{ijl} + \theta_{jl}$ with $\theta_{jl} = 1$ if neighbors j and l are connected with each other and 0 otherwise.

The clustering coefficient gives an idea of how compact is the network. Correlation of links allows us to sustain cooperation between Eurostat methodological network members, that otherwise would not be able to function. If respondents i and h forms a links to common competencies j and l, then efficient collaboration between them is very possible. In the methodological network studied, the clustering coefficient of competencies vertices set is not so high, above 20%. There is detected a moderate correlation between clustering coefficient and degree.

Another important mathematical measure of the graph is structural equivalence. A pair of nodes are structurally equivalent if they are connected to exactly the same others. Structurally equivalent nodes are identical with respect to all structural properties. As a result one approach to identify structurally equivalent nodes is to compute a similarity measure among rows and columns of the adjacency matrix defining the graph. The minimum value of similarity for nodes i and h is 0, it captured when none of the node's neighbors are neighbors with each other, while the maximum is 1, and it means that all of two nodes' neighbors are overlapped. The value is in between when partial overlap is captured, closer to 1 when the overlap is large compares to their degrees. In Eurostat methodological network overlapping is not significant, the similarity equivalence of competencies vertices set is very low, above 10%, for persons' nodes slightly higher, up to 20%.

Isolated nodes in the sub-network of Methodological areas belongs to set of competencies, while in the sub-networks of Statistical domains and tools it refers to the set of members of Eurostat methodological network.

The highest proportion of the amount of existing edges to the maximum possible amount of links belongs to sub-network of Statistical tools, what informs, that respondents are highly competent in statistical software. The lowest density detected in sub-network of methodological areas, but the reason of it could be that the variety of choices was almost three times bigger, what makes network less connective.

Results in the Table 1 provide quantitative evidence that respondents are qualified in different fields and there are no any overlapping nodes, similarity of the respondent's competencies is low. Generally given measures ensure possibility of well performance of Eurostat methodological network. Network is connected, and gap of competencies is detected only in one methodological area from the defined list.

Finally, there is a group of factors not related to knowledge itself, that also influence knowledge diffusion within the network and functionality of itself. To obtain successfully working network, we are not enough to have a list of existing knowledge, in addition is important to know if person is interested in knowledge sharing and taking of active or consultative role within the work organised by network. Two respondents having the same knowledge may have different wills spreading it. Nodes in this network may present different preference for being active members of the network while other passive.

Table 1 Statistics for the whole Eurostat methodological network and for sub-networks

	Whole network	Methodological areas	Statistics domains	Statistics tools
n_{V_1}	67	67	67	67
n_{V_2}	50	28	12	10
m	595	299	144	152
k_{V_1}	8.9	4.5	2.1	2.3
k_{V_2}	11.9	10.7	12.0	15.2
k	10.2	6.3	3.8	4.4
δ	0.18	0.16	0.18	0.23
cc_{V_1}	0.18	0.25	0.36	0.37
cc_{V_2}	0.13	0.40	0.11	0.17
$r_{k_{V_1},cc_{V_1}}$	0.38	−0.21	−0.79	−0.60
$r_{k_{V_2},cc_{V_2}}$	0.50	−0.77	−0.15	−0.35
Isolated nodes (%)	1	1	4	10
SE_{V_1}	0.14	0.14	0.14	0.17
SE_{V_2}	0.08	0.06	0.06	0.09

3.4 Network Visualisation and Evaluation

Representing data as mathematical graph makes readily see general view, absorb its structure, and permits us to focus on Eurostat methodological network members, competencies and relations between it. This network established constructing person by methodological area, statistical domain and tool expertise graphs. Our built interactive graphs (see Fig. 2) has two-mode nature, it consist two disjoint sets of vertices.

The different mode and qualitative differences among variables we represent by different colours. The vertices of Eurostat staff who has participated in the survey are coloured in green, blue, and red depending on the interest in involvement, while the second set corresponds to 28 methodological areas, 12 statistical domains and 10 tools, which are coloured in yellow (see Fig. 3). In two mode network edges exist only between the vertices belonging to different sets. If person i has certain competence j undirected link between those two components exist, otherwise does not.

Network data gives a complete picture of relations within the network. Figure 3 displays a graphical representation of bipartite graph of statistical competencies by Eurostat methodological network members. In order to highlight some characteristics, of the network structure, visual effects are added to the graph. We distinguish the two node sets by colors, so that nodes of the same type have the same color. The size of the label and vertex is proportional to its degree (number of links), meaning that the more links the node has the larger it appears in the graph. Lines in the network represent links, meaning that person i is competent in skill j.

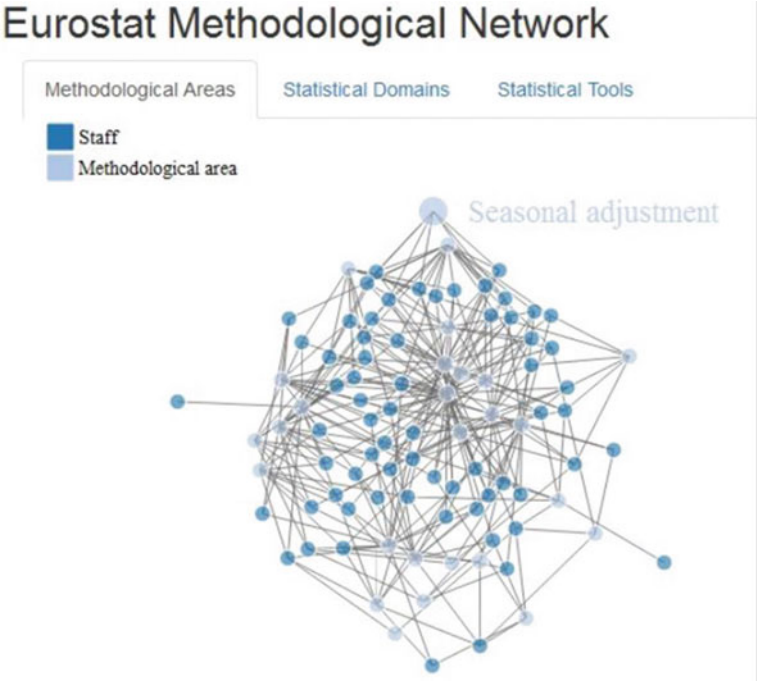


Fig. 2 Interactive navigation tool build for Eurostat methodological network

Nodes are distributed according to degree as well. The nodes of the competencies with high degree are located in the center of the graph, while the nodes with low degree are spread more on the sides.

Figure 3 depicts the network among Eurostat methodological network members and list of statistical methodological skills. This network consists of 117 vertices and 595 edges, the average degree is equal to 10.2 with the density of 0.2. The lowest average degree belongs to set of staff vertices and is equal to 8.9. The set of competencies obtain slightly higher average degree. The reason of this difference is size of the set of vertices. The higher size of sets itself ensure lower average degree. The highest in-degree values within the network detected at the vertices of Data Analysis (degree 39), software R (degree 30), and Social statistics (degree 29) which are the largest and particularly central in the graph with most of connections to respondents. It means that 58% of respondents are competent in Data Analysis, 45 in statistical software R, and 43% are specialised in Social Statistics. The lowest in-degree belongs to methodological areas Statistical Confidentiality, Data Warehousing, Data Integration, and Enterprise Architecture, to Transport and Energy statistical domains, and tool Hadoop. The minimum and maximum out-degree of respondents is 3 and 11 respectively. The measure of density is quite low, however considering that some restrictions have been introduced allows us to state that network is well connected.

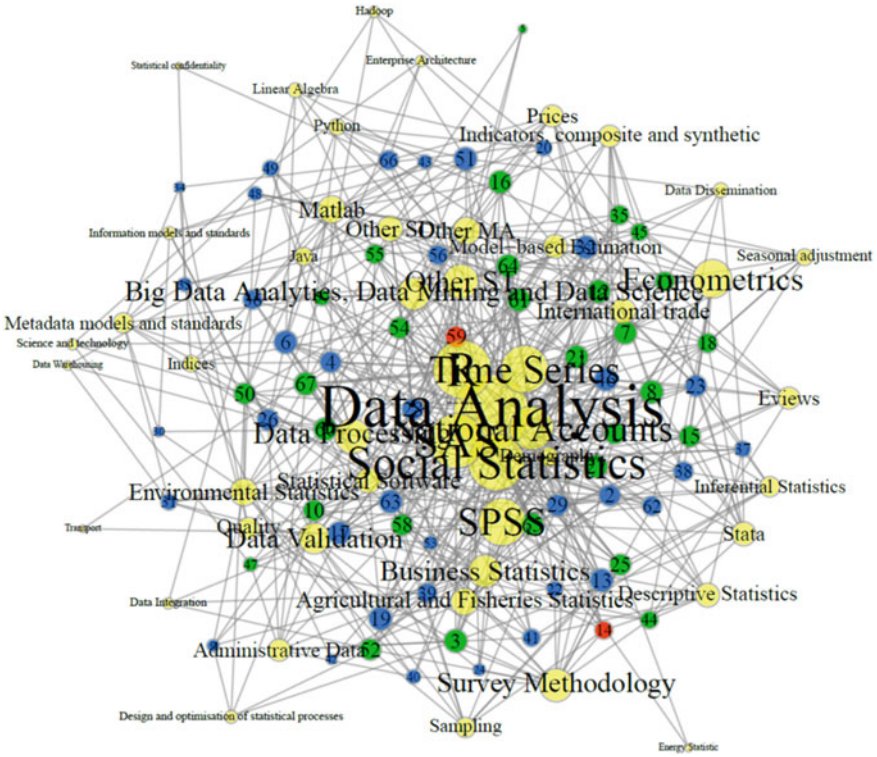


Fig. 3 Eurostat methodological network

At first sight the graph of the full network appears quite heavy and not well readable because of too much information. Some of the visual information is inevitably lost as the nodes and links overlap and obscure each other. The way to make it clearer was by splitting the network in several sub-networks by competencies, dividing the vertex set of competencies into three subsets: of methodological areas; statistical domains; and statistical tools. For $V_{MA} \subseteq V_2$, we say $G_{MA} = (V_1, V_{MA}, E_{MA})$ is the sub-network of G in V_{MA} if $E_{MA} \subseteq E$ contains the links in G that connects individual and competence from the list of methodological areas (see Fig. 4). While for $V_{SD} \subseteq V_2$ graph $G_{SD} = (V_1, V_{SD}, E_{SD})$ is the sub-network of G in V_{SD} if $E_{SD} \subseteq E$ contains links in G that connects individual and competence from the list of statistical domains (see Fig. 5). By the same approach we obtain the sub-graph for statistical tools (see Fig. 6).

However, for the presentation purpose it makes clearer picture of how wide connected graph is and nodes are not located in the same part of space what lets us to avoid chaos in the graph. In the Methodological areas sub-network the number of staff remains the same, only number of competencies decreases to 28 and it contains 299 ties. Sub-network based on methodological areas has one isolated vertex

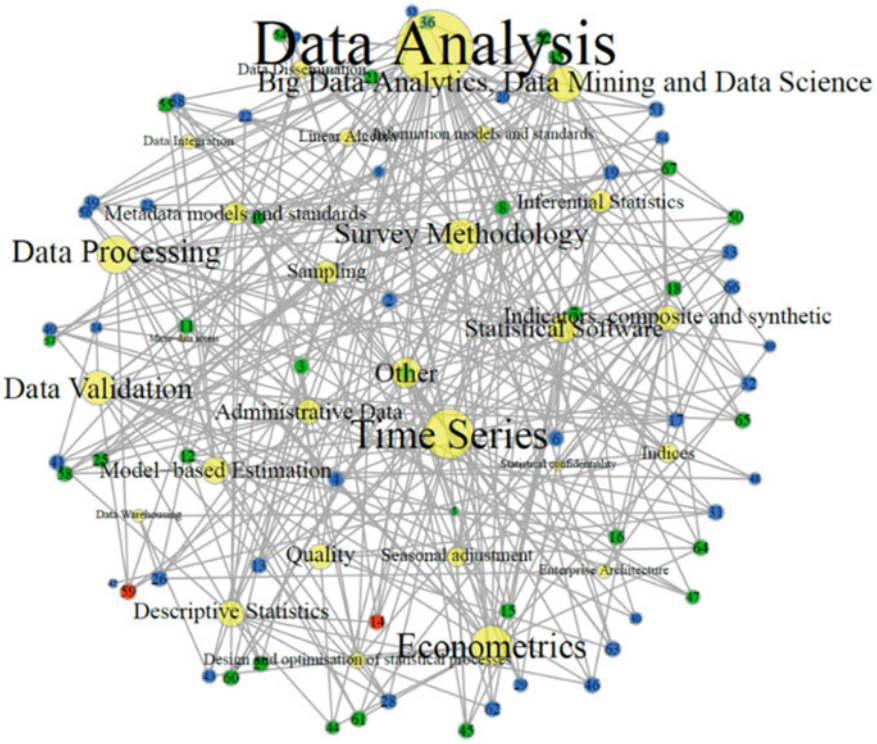


Fig. 4 Competency sub-network on methodological areas by methodological network members

which represents the competence of Micro-data access (with degree 0). The correlation between clustering coefficient and degree is negative, that means clustering coefficient of low degree nodes is quite large, but the one of high degree nodes is small. The staff competencies vary differently, as overlap of the skills do not overlap significantly.

By using one of the force based algorithms (for instance Fruchterman and Reingold) for data visualisation we get more informative graphs. In the Fig. 5, graph's two nodes are near each other roughly to the extent that the geodesic distance between them is short. Simply saying positions of vertices from the some or different sets are sets are near each other if they have link to each other or to the same dimensions. For instance vertices of people are near each other if they are connected by the choosing common statistical domains. In this example the representation makes clear that there is a set of people shown in the lower part of the graph who have experience in Business statistics, National accounts and Prices, while the other cluster of people (represented on the top) are related by domains like Science and Technologies, Environmental and Agricultural statistics. The vertex of Social statistics is in between of both clusters, as it shares the list of methodological network members from the one and another cluster. We can see that this domain is principle and has direct links to

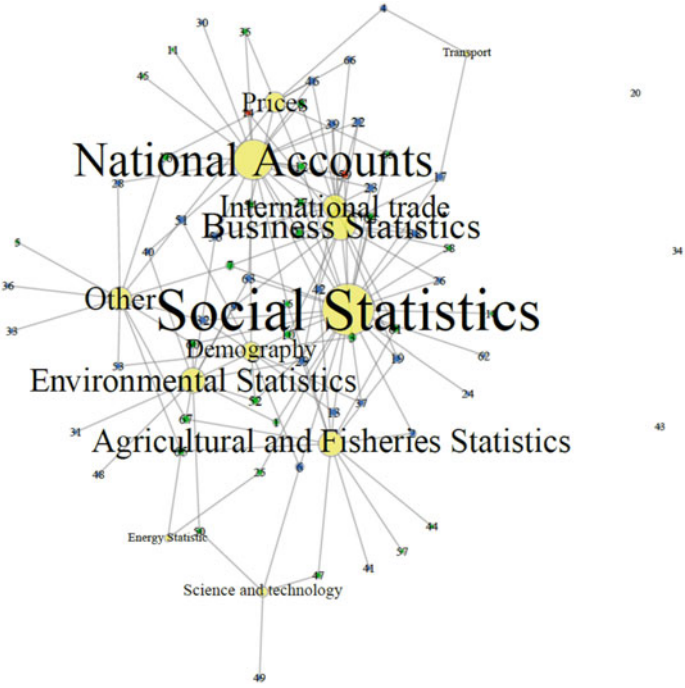


Fig. 5 Competency sub-network of statistical domains by methodological network members

all other statistical domains, except Science and technologies. Sub-network based on statistical domains has three isolates nodes, which belong to the set of staff vertices. It means that there are three people who have not indicated any statistical domain from the list as being competent in. The average degree of domain sub-network is equal to 3.8. The lowest degree value is 0 by frequency of three and belongs to the set of vertices of staff. While the highest degree is equal to 29 which lets us assume Social Statistics as best known domain. The least known domains between our respondents are Transport and Energy statistics (with degree of 2). The density of this sub-network is 0.18, which means that 18% of all the potential ties between respondents and competencies are actually present.

By the same algorithm we visualise the sub-network of competencies in Statistical tools (see Fig. 6). Here we have eight isolates nodes, which belong to the vertices set of staff. It means that there are eight people who haven't indicated any statistical tool as being experienced in. The average degree of statistical tool sub-network is 4.4. The lowest degree value is 0 by frequency of the aforementioned eight people.

While the highest degree is equal to 30, which belongs to software R as the most used tool between respondents. The large the vertex is, the more central the competence in the network is. The least known tool is Hadoop. Even though the density

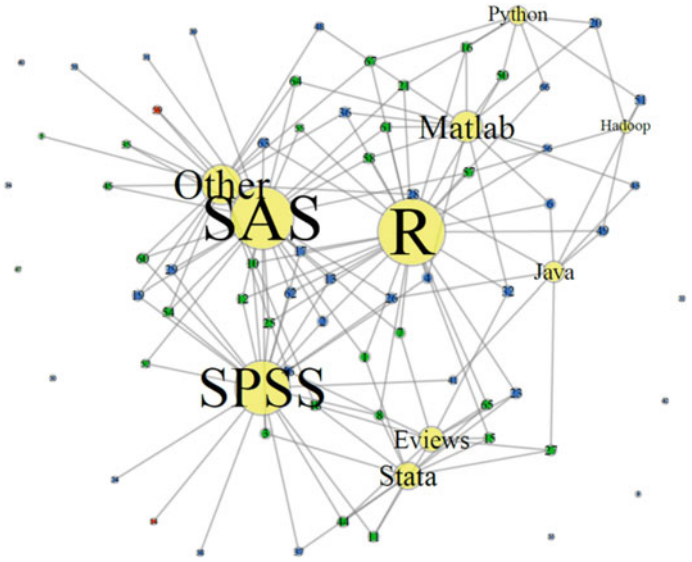


Fig. 6 Competency sub-network of statistical tools by methodological network members

of this sub-network is low 0.23, it is still higher than of the whole methodological network.

4 Projections

Thus far we had discussed two mode graphs, putting people and competencies into a network itself. But in fact it is important to view competencies as existing "outside" the network, by making two mode network's projection into two one mode networks. From the Eurostat methodological network we generate two relevant network projections. In the staff network nodes represent people who replied to Eurostat methodological network—staff survey, and two persons form link to each other if they have at least one common field of competencies. In this section only active members of the Eurostat methodological network are considered and analysed.

4.1 Projection to Person by Person Methodological Network

Given matrix $\mathbf{A}$, is reconstructed by the multiplication of matrix $\mathbf{A}$ and its transpose $\mathbf{A}^\top$, that produce a person by person matrix whose ijth cell indicates the number of methodological fields both persons i and k are competent in. This value is interpreted

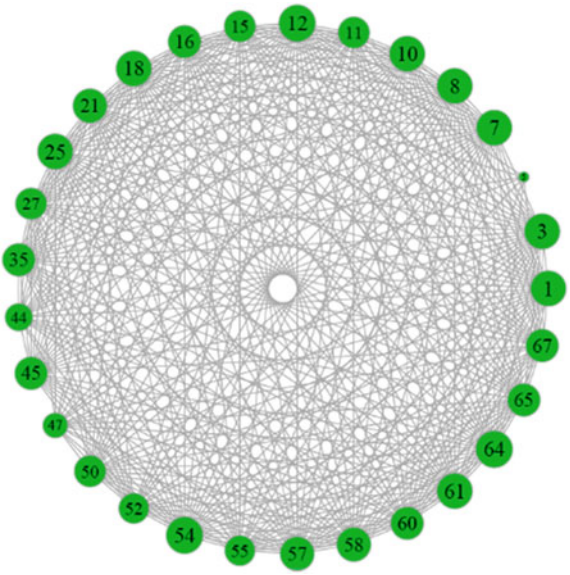


Fig. 7 Eurostat methodological network active member by member network

as weight or an index of the strength of knowledge proximity of the two persons. The fact that people who are competent in the same statistical methodological skill have a link to each other, is a statement about the structure of methodological collaboration network. This matrix can be interpreted as an index of possible staff interactions. The higher the number of the common competencies is, the more significant overlapping of existing competencies is and more likely those two people could collaborate together efficiently. The visualisation of person by person network is given in Fig. 7.

The size of each node is proportional to the number of people with whom person has common field of knowledge. Mathematical measures of the network notes that active members of Eurostat methodological network are well connected by existence of common skills. The network is a composition of 29 nodes and 359 weighted edges. The degree ranges from 8 up to 28 with a mean of 24.8. The density is equal to 0.88 what is considered as very high, meaning that 88% of possible links appear in the network.

4.2 Projection to Competency by Competency Methodological Network

In the same way as defined in the Sect. 4.1, just multiplying the transpose $\mathbf{A}^\top$ of matrix $\mathbf{A}$ by the original matrix, we obtain competence by competence matrix, where each cell gives the number of people who is experienced in both, the row and the column competences. The cells in the principle diagonal indicate the degree of that

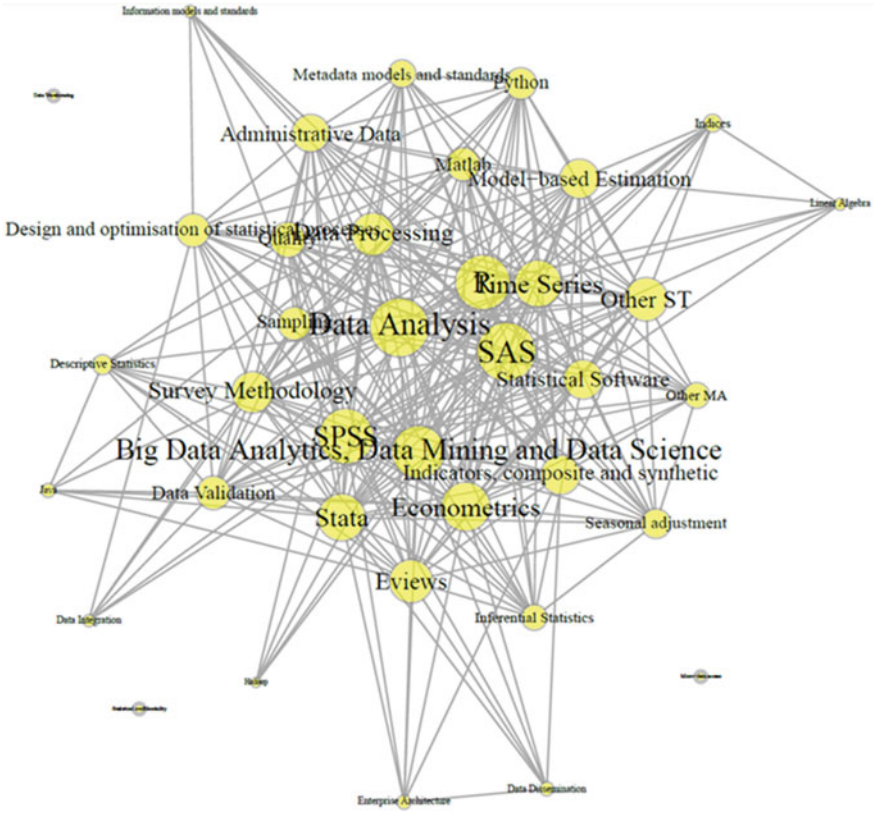


Fig. 8 Competence by competence network of active methodological network members

competence. The visualisation of competency by competency network is given in Fig. 8.

Strongly connected components are centralised in the graph. Components with weaker connections are put in the boundaries. The size of the node increases depending on the number of linked competencies.

The competencies network consists of four components: one connected and three of isolates. From the graph we can see that the most central nodes are Data analysis, SAS, R, etc. While the isolates are the nodes of Data warehousing, Micro-data access, and Statistical Confidentiality, what means, that active Eurostat methodological network members have not declared is as their main experience in.

This visual observation is confirmed by mathematical measures of the network. It is a one mode undirected sub-network consisting of 50 nodes and 301 weighted edge. The average degree ranges from 5 up to 30 with a mean of 17.2. The density is equal to 0.43 what is considered as relatively high.

5 Ego-Network

The methodology and visualisation tolls analysed and developed under this specific project, was adapted and reused for other initiatives within the Eurostat. In response to the request from Eurostat R users group that is a network of staff interested in R programing and innovations, the two mode competency's R ego-network has been modelled. Ego-network is composition only of those nodes which are related to one specific node, all other nodes have been eliminated [3]. In ego-network based on competence of statistical software R, only members of the Eurostat methodological network and information on competencies in methodological areas chosen by them, who are competent in this particular tool (Fig. 9).

R ego-network consists of 30 Eurostat methodological network members, of which 14 are active, 15 informative and 1 inactive, and 26 methodological areas. The degree of R ego-network ranges from 1 to 21 with a mean of 4.9. The most central nodes are of Data analysis and Time series. The density is equal to 0.17, which is slightly lower than the one of the whole network of methodological areas.

The active members of Eurostat methodological network provide support and share knowledge with the in-house R users and statistics producers. This saves costs of the organisation allowing reuse existing resources, instead of purchasing consultations.

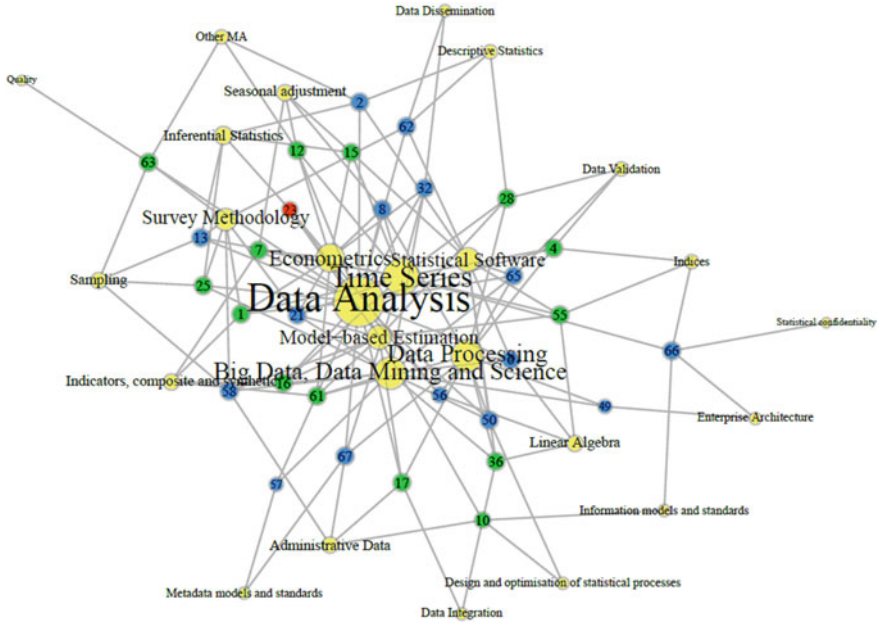


Fig. 9 R software ego network competencies by methodological network members

6 Conclusions and Discussion

In this paper, we evaluated Eurostat methodological network applying network analysis techniques. Networks as analytical and visualisation tools provided a number of useful outcomes. Our research object is respondents to the 'Eurostat methodological network—staff survey'. 10% response rate have been considered as quite high and as a good starting point for evaluation and development of newly made up methodological network, assessing the fact that the survey is based on voluntary basis.

Applying network analysis and visualisation techniques we were able to study the structure of Eurostat methodological network. The distribution diagrams, tables of mathematical measures and plotted graphs have displayed important information about the methodological network components. Similar network diagrams are being produced regularly for any breakdown under request.

Eurostat methodological network consists of 67 respondents that constructed 595 links to 50 statistical methodological areas, domains, and tools. We noticed that in the network exist only one isolated node, meaning that gap of skills in that particular area exist within the network members. Almost half of respondents expressed an interest in taking an active role on upcoming in-house methodological initiatives and projects.

Results show high competence staff with density 0.18 of the Eurostat methodological network, what is quite high considering that some restrictions has been introduced for filling in the survey. We have learned that almost all methodological areas, statistical domain and tools would be covered by people interested in collaboration and contribution on upcoming processes and projects with interest rate from 2 up to 39. The competencies known by majority of respondents are in Data Analysis, Time Series, National Accounts, Social Statistics, R, SAS, and SPSS. The lack of knowledge within Eurostat methodological network members noticed in area of Micro-data access. Going deeper and looking at the indicators of sub-networks we notice the tendency on increase of the density when average degree decreases. Overlapping of the structure of the nodes is very small, what points that there is large variety of the respondents with different knowledge.

One of the key methods for addressing skills gaps is the provision of appropriate training courses. Monitoring of skills gaps and already existing knowledge lets us organise training in efficient way for better staff knowledge development, leading to productive performance of daily duties and methodological network functions.

We can outline the importance of monitoring existing in-house knowledge. Two employees could affect each other if they are aware about each other common competencies. Or while looking for specific information an efficient communication within the organisation is possible only when we know with whom we could potentially contact. Network is a key source in helping and supporting of knowledge diffusion and expanding, enriching professional and personal skills and filling in the gaps.

Moreover by obtained high response rate and statistics produced in the described study we were able to confirm the possibility of efficient network functionality and

stress out the importance of the methodological network further development and implication for diffusion of knowledge.

As actual and most visible results at Eurostat the Methodological strategy has been adopted. The methodological network launched and already finalised seven specific methodology related projects, some of them are now developed for production and dissemination as part of Eurostat experimental statistics. The regular in-house R users support is being performed as well.

References

1. Borgatti, S.P., Everett, M.G.: Network analysis of 2-mode data. Soc. Netw. **19**, 243–269 (1997)
2. Butts, C.T.: Social network analysis: a methodological introduction. Asian J. Soc. Psychol. **11**, 13–41 (2008)
3. Hanneman, R.A., Riddle, M.: Introduction to social network methods. http://faculty.ucr.edu/~hanneman/nettext/index.html (2005)
4. Latapy, M., Magnien, C., Del Vecchio, N.: Basic notions for the analysis of large two-mode networks. Soc. Netw. **30**, 31–48 (2008)

The Evaluation of the Inequality Between Population Subgroups

Michele Costa

Abstract This paper illustrates the advantages to evaluate inequality between population subgroups with respect to a maximum compatible with the observed data, thus going beyond the traditional approach to the analysis of inequality between, where the maximum corresponds to total inequality. The new proposal improves both the measurement and the interpretation of the contribution of inequality between to total inequality.

Keywords Inequality evaluation · Inequality decomposition · Inequality between subgroups

1 Introduction

Inequality decomposition can be extremely helpful into evaluating and understanding the individual distribution of economic and social variables. Moreover it provides powerful insights on the comparisons across time and space.

Inequality between population subgroups represents perhaps the most important component of total inequality. By means of inequality between, different sources of inequality are evaluated and compared, with the twofold goal to detect the main determinants of inequality and to implement socio-economic policies able to reduce or alleviate its consequences. Inequality-reduction policies will address poverty and social exclusion, but also gender or race gaps as well as many other themes of economics. Policy interpretations of inequality decompositions are a challenging topic and rise many questions, some of which still unanswered [12].

The measurement of inequality between can be achieved following different approaches, since inequality literature presents a wide collection of contributions on inequality decomposition. However, the size of inequality between is usually evaluated with respect to its theoretical maximum, which corresponds to total inequality,

M. Costa (✉)
Department of Economics, University of Bologna, Bologna, Italy
e-mail: michele.costa@unibo.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_22

when the inequality within subgroups is equal to 0. The case of null inequality within is a quite unrealistic situation, which can be essentially considered as a theoretical reference, without a proper phenomenal correspondence. That is, we really do not expect to achieve a situation where each unit of each subgroup possesses the subgroup mean.

Furthermore, by comparing inequality between to total inequality, we can observe two unfortunate effects. First, the size of inequality between is frequently unreasonably small, thus suggesting a too low influence of the underlying inequality factor. Second, the measure of inequality between is strongly influenced by the number of subgroups used into the partition of the total population, thus preventing a direct comparison between different inequality factors when the number of subgroups is not the same.

In order to overcome these drawbacks, we propose a new framework for the evaluation of the inequality between, where the basis for comparison is not represented by total inequality, but by the maximum which can be obtained given the observed data. More specifically, we do not allow to the inequality within to be 0, but we refer to the minimum inequality within compatible with the data.

We build on [7] and develop new indicators for the evaluation of the inequality between. The new indexes allow to assess the importance of the different inequality factors into the observed data, thus improving our knowledge of inequality.

We illustrate the effects of the number of subgroups on inequality between and on its evaluation by means of a Monte Carlo study, which also allows to compare the new indexes to the traditional approach. We also propose a case study on real data with a typical income inequality decomposition based on two different inequality factors.

2 Methodology

In the following we will adopt as our inequality measure one of the most used and widespread inequality measure, the Gini index [8]:

$$G = \frac{1}{2n^2\bar{y}} \sum_{i=1}^{n} \sum_{r=1}^{n} |y_i - y_r| \tag{1}$$

where $\bar{y}$ is the arithmetic mean of Y in the overall population, y_i is the value of Y in the ith unit and, accordingly, y_r is the value of Y in the rth unit.

Introduced with the purpose to measure the inequality in the individual distribution of income and wealth, the Gini index has experienced and extraordinary success, with a wide variety of different formulations and extensions (see e.g. [15]) proposed during more than a century. From the original area of economic inequality, the use of the Gini index has expanded to poverty, well-being and many other fields of economics.

In the following we will refer to the expression of the Gini index for the case of a population disaggregated into k subgroups

$$G = \frac{1}{2n^2\bar{y}} \sum_{j=1}^{k} \sum_{h=1}^{k} \sum_{i=1}^{n_j} \sum_{r=1}^{n_h} |y_{ji} - y_{hr}| \tag{2}$$

where y_{ji} is the value of Y in the ith unit of the jth subgroup and, accordingly, y_{hr} is the value of Y in the rth unit of the hth subgroup, while n_j and n_h, are the size of the jth subgroup and of the hth subgroup, respectively.

For a detailed discussion of the Gini index see, e.g., [4, 9, 10, 15].

Notwithstanding the great importance of the Gini index, here it represents an example of a measure of inequality, and for any other inequality indicator the same observations regarding the evaluation of the inequality between would apply.

2.1 The Dagum's Gini Index Decomposition

The literature on the Gini index decomposition is extremely wide, but as for the choice of the inequality indicator, also the choice of the method of decomposition is not a crucial aspect, since the same development proposed for a method can be extended to all the others. Among the many contributions which allow to decompose the Gini index (see [5, 11, 14] among the others), we use the decomposition proposed by Dagum [6], who builds on a previous work of Mehran [13].

The Dagum's contribution is developed on the basis of three components: the inequality within the k subgroups G_w, the inequality between the k subgroups G_b and the overlapping between the k subgroups G_t.

The inequality within can be easily derived as a weighted average of the Gini indexes of each subgroup:

$$G_w = \sum_{j=1}^{k} G_{jj} p_j s_j \tag{3}$$

where

$$p_j = n_j/n$$

and

$$s_j = (n_j \bar{y_j})/(n\bar{y})$$

are the population share and the character share of the jth subgroup, respectively.

The contribution to total inequality related to the differences between the subgroups is evaluated on the basis of Gini index between subgroups j and h, G_{jh}, as

$$G - G_w = G_b + G_t = \sum_{j=1}^{k} \sum_{h=1, j \neq h}^{k} G_{jh} p_j s_h$$

where

$$G_{jh} = \frac{1}{n_j n_h (\bar{y}_j + \bar{y}_h)} \sum_{i=1}^{n_j} \sum_{r=1}^{n_h} |y_{ji} - y_{hr}|.$$

Since the original version of both G_b and G_t require an heavy computational effort, a simplified version of the Dagum's decomposition is available by Costa [2] as

$$G_b = G_b^* + 0.5(G - G_w - G_b^*) \tag{4}$$

$$G_t = 0.5(G - G_w - G_b^*) \tag{5}$$

where

$$G_b^* = \sum_{j=1}^{k-1} \sum_{h=j+1,k}^{k} \frac{p_{hj}^* - s_{hj}^*}{p_{hj}^* s_{jh}^* + p_{jh}^* s_{hj}^*} (p_j s_h + p_h s_j)$$

$$p_{hj}^* = p_h / (p_h + p_j)$$

$$s_{hj}^* = s_h / (s_h + s_j).$$

The Dagum's decomposition has an immediate link to the Gini index expression for the case of k subgroups since it assigns the differences $|y_{ji} - y_{hr}|$ in (2) to G_w when $j = h$, to G_b when $j \neq h, \bar{y}_j \geq \bar{y}_h, y_{ji} \geq y_{hr}$, and to G_t when $j \neq h, \bar{y}_j \geq \bar{y}_h$, $y_{ji} < y_{hr}$. Globally we have $G = G_w + G_b + G_t$.

For a detailed description of the Dagum's decomposition see [2, 6].

2.2 *The Traditional Evaluation of the Inequality Between*

The measurement of the contribution to total inequality attributable to the differences between the subgroups represents the main argument into the debate on inequality decomposition. Since the pioneering work of Bhattacharia and Mahalanobis [1] many Authors proposed different proposal for the measurement of G_b: even if an exhaustive list would be a challenging task [10], we cite, besides the Dagum's papers previously illustrated, the contribution by Yitzhaky and Lerman [14].

However of great interest, the debate on the measurement of G_b is not relevant here, since our focus is not on the measurement but on the evaluation of G_b.

In the framework of the Dagum's decomposition, as well as following any other approach to the Gini index decomposition, or to the decomposition of any other

inequality indicator, inequality between is usually evaluated with respect to its maximum, which is achieved when two conditions are verified.

First, the k subgroups should not overlap, that is, in our case, the component G_t is equal to 0.

Second, the variability within the k subgroups should be equal to 0, that is the component G_w is equal to 0 and each subgroup unit possesses the subgroup mean: $y_{ji} = \bar{y}_j$, $j = 1, \ldots, k$; $i = 1, \ldots, n_j$.

On the basis of these two conditions, we have $G_w = 0$, $G_t = 0$ and then

$$G_{bmax} = G - G_w - G_t = G.$$

By referring to the case $G_{bmax} = G$, the evaluation of G_b is generally obtained by means of the ratio

$$I_{G_b} = G_b / G_{bmax} = G_b / G \tag{6}$$

which is used to measure the weight of inequality between on total inequality and to determine the importance of different inequality factors.

2.3 A New Proposal for the Evaluation of the Inequality Between

With the aim to provide new insights on the evaluation of the inequality between, in this paper we propose to modify $G_{bmax} = G$, that is the traditional reference for the analysis of G_b. More specifically, we propose to relax the condition $G_w = 0$ and to compare G_b not to its theoretical maximum G, but to the maximum G_{bmax} which can be achieved conditionally to the observed data. That is, we propose to compare G_b not to the unrealistic case of equidistributed subgroups, but to a case more coherent and compatible with the data.

We preserve the condition $G_t = 0$ since it is less unrealistic than $G_w = 0$. For example, if we divide total population in 2 subgroups by gender, the hypothesis of no overlapping, that is the richest female unit is poorer than the poorest male unit, however extreme, seems less unrealistic than the hypothesis of null inequality within, that is all the female units have the same income $\bar{y}_f$ and all the male units have the same income $\bar{y}_m$.

Moreover, the presence of overlapping influences [2, 3] the role of the inequality factors and, therefore, in order to achieve a better understanding of their importance, it is more opportune to set $G_t = 0$, thus removing this source of potential differences.

By maintaining the condition of no overlapping, we have $G_t = 0$, but, by relaxing the hypothesis of null inequality within, the minimum of G_w is no longer 0, but G_{wmin}, that is the minimum inequality within, which can be obtained partitioning the observed data into k non overlapping subgroups. In this way we get

$$G_{bmax} = G - G_w - G_t = G - G_{wmin}.$$

We have many ways to divide n units into k non overlapping subgroups: with the aim of preserving the structure of the original partition, we propose two possible solutions. First, we obtain the k subgroups by using the original $p_i = n_i/n$ values, thus keeping the same population shares of the original partition. Second, we obtain the k subgroups by using the original $s_i = (n_i \bar{y}_i)/(n\bar{y})$ values, thus keeping the original income shares.

The next step of our method refers to the calculus of G_{wmin}, the minimum inequality within compatible with the new k subgroups. We propose to permute the sequence of the p_i (or s_i for the second solution), to get a set of k subgroups for each permutation, to calculate the related G_w and to chose the minimum value among all disposable G_w. Let be $G_{wmin(p)}$ the minimum inequality within, which can be obtained by permutating the values p_i and, correspondingly, $G_{wmin(s)}$ the minimum inequality within, which can be obtained by permutating the values s_i.

In the last step we derive the new indexes for the evaluation of G_b, obtained as

$$I_{G_{b(p)}} = G_b/(G - G_{wmin(p)}) = G_b/G_{bmax(p)} \tag{7}$$

$$I_{G_{b(s)}} = G_b/(G - G_{wmin(s)}) = G_b/G_{bmax(s)} \tag{8}$$

The new indexes depend on the minimum inequality within compatible with the observed data and, therefore, are not strongly affected by k as for I_{G_b}.

3 The Simulation Study

In this section we present a Monte Carlo study aimed at analysing the effects of inequality between measurement related to the number of subgroups k and the number of observations n. Furthermore, the Monte Carlo study also allows to assess how these effects influence the evaluation of inequality between and to compare the traditional framework based on I_{Gb} to the new proposals I_{Gp} and I_{Gs} presented in (7) and (8).

The simulated samples are randomly extracted from a beta or a gamma distribution: in order to achieve a wide coverage, for each subgroup beta distribution parameters $B(a, b)$ are randomly selected with $0.5 < a < 2$ and $0.5 < b < 4$ or gamma distributions parameters $G(c, d)$ are chosen within the intervals $0.5 < c < 10$ and $0.5 < d < 10$.

We consider the cases of $k = 2, 3, 4, 5$ for the number of subgroups and $n = 500, 1000, 5000$ for the number of observations.

For each combination of k and n, 10,000 samples are generated, 50% from a beta distribution and 50% from a gamma distribution. For each sample has been calculated the overall Gini index, the three terms decomposition proposed by Dagum and the index I_{G_b} for the traditional evaluation of the inequality between subgroups.

Table 1 Inequality decomposition characteristics for the simulated samples, average ratios G_w/G, G_b/G and G_t/G by k and n

k n	500	1000	5000	500	1000	5000	500	1000	5000
	Gw/G			Gb/G			Gt/G		
2	0.48	0.50	0.55	0.47	0.45	0.40	0.05	0.05	0.05
3	0.28	0.34	0.35	0.65	0.62	0.60	0.07	0.05	0.05
4	0.19	0.22	0.26	0.74	0.72	0.67	0.07	0.07	0.07
5	0.13	0.15	0.21	0.80	0.78	0.71	0.07	0.07	0.08

Furthermore, from the sequences of p_i and s_i of each simulated sample have been obtained all possible permutations and with respect the new non overlapping subgroups obtained by means of each permutation has been calculated the inequality within G_w. The minimum of all G_w is used to obtain the index $I_{G_{b(p)}}$ when the permutation refers to the p_i and $I_{G_{b(s)}}$ when the permutation refers to the s_i. The condition of no overlapping is introduced when calculating $G_{bmax} = G - G_w - G_t = G - G_{wmin}$, while simulated samples allow overlapping to achieve more realistic situations.

Table 1 illustrates the Dagum's Gini index decomposition by means of the ratios G_w/G, G_b/G and G_t/G by number of observations n and by number of subgroups k. Each value on Table 1 refers to the average of the 10,000 ratios obtained for the 10,000 samples generated for each k and n. For example, for the 10,000 samples calculated for $k = 2$ and $n = 500$ we have that, on average, the 48% of total inequality is given by G_w, the 47% by G_b and the 5% by G_t.

The ratio Gw/G illustrates the weight of inequality within in total inequality: while the number of observations n seems to have only a slightly influence, the number of subgroups k strongly affects the importance of inequality within on overall inequality, with an inverse relation between Gw/G and k.

The analysis of the role of the inequality between by means of the ratio Gb/G follows the traditional way to evaluate the relevance of inequality between. It is possible to observe how Gb/G strongly depends on the number of subgroups k, thus confirming one of the main criticism to its use to evaluate the inequality between.

The weight of the overlapping component, evaluated by means of the ratio Gt/G, is quite small, as in many samples it is only marginal and the average has the effect to water down its overall importance. However, as expected, we can observe a direct relation between overlapping and the number of subgroups.

The traditional evaluation of the inequality between, obtained by means of $I_{G_b} = Gb/G$, is complemented by the new indices $I_{G_{b(p)}}$ and $I_{G_{b(p)}}$ reported in Table 2. Each value of Table 2 is obtained as the average of the 10,000 samples calculated for each case: for example, when $k = 2$ and $n = 500$, 0.71 is the average of the 10,000 $I_{G_{b(p)}}$ obtained in the 10,000 simulated samples extracted for this combination of k and n.

From Table 2 it is possible to observe how the number of observations n still does not affect the evaluation of G_b. However, unlike what happens for I_{G_b}, the new indexes $I_{G_{b(p)}}$ and $I_{G_{b(p)}}$ are quite robust with respect to the number of subgroups k:

Table 2 Inequality between evaluation by means of I_{G_b}, $I_{G_{b(p)}}$ and $I_{G_{b(s)}}$, average values in simulated samples by k and n

k n	500	1000	5000	500	1000	5000	500	1000	5000
	I_{G_b}			$I_{G_{b(p)}}$			$I_{G_{b(s)}}$		
2	0.47	0.45	0.40	0.71	0.70	0.67	0.69	0.67	0.64
3	0.65	0.62	0.60	0.77	0.76	0.74	0.77	0.75	0.74
4	0.74	0.72	0.67	0.81	0.81	0.78	0.81	0.80	0.77
5	0.80	0.78	0.71	0.84	0.84	0.80	0.82	0.84	0.77

Table 3 Inequality decomposition characteristics for the simulated samples, mean of IGb, IGp, IGs by k and by deciles of Gw/G (Gb/G, Gt/G)

k deciles of Gw	I	II	III	IV	V	VI	VII	VIII	IX	X
I_{G_b}										
2	0.73	0.63	0.57	0.51	0.48	0.43	0.40	0.36	0.33	0.25
3	0.85	0.77	0.74	0.71	0.68	0.64	0.6	0.56	0.53	0.45
4	0.89	0.84	0.81	0.78	0.77	0.74	0.71	0.68	0.64	0.58
5	0.91	0.87	0.85	0.84	0.82	0.80	0.77	0.73	0.72	0.66
$I_{G_{b(p)}}$										
2	0.91	0.85	0.83	0.77	0.76	0.71	0.62	0.59	0.56	0.46
3	0.94	0.88	0.85	0.83	0.80	0.76	0.74	0.69	0.66	0.58
4	0.94	0.90	0.87	0.85	0.84	0.80	0.78	0.75	0.72	0.64
5	0.95	0.92	0.89	0.87	0.86	0.85	0.80	0.78	0.77	0.70
$I_{G_{b(s)}}$										
2	0.93	0.87	0.81	0.78	0.76	0.67	0.64	0.55	0.52	0.39
3	0.94	0.89	0.86	0.84	0.82	0.77	0.72	0.68	0.65	0.57
4	0.96	0.92	0.89	0.87	0.86	0.84	0.81	0.78	0.72	0.71
5	0.98	0.96	0.95	0.95	0.94	0.92	0.91	0.87	0.87	0.86

from Table 2 we still can note a direct relation with k, but the increasing rate is really lower than for I_{G_b}.

The results of the simulation study offer many possibilities to investigate the behaviour of I_{G_b}, $I_{G_{b(p)}}$ and $I_{G_{b(s)}}$. Given the importance of size of inequality within into the simulation study and its possible effects on the evaluation of inequality between in Table 3 we analyse the traditional and the new indices with respect to G_w. The results are sorted by increasing value of G_w, 10 groups are constituted on the basis of the deciles of G_w and Table 3 reports the average of I_{G_b}, $I_{G_{b(p)}}$ and $I_{G_{b(s)}}$ in each group.

From Table 3 it is possible to determine the influence of inequality within on the evaluation of inequality between: by increasing k all indices converge to similar values, but the new proposal show clearly an higher degree of robustness particularly when G_w is lower.

A further goal which could be achieved by means of a simulation study refers to the correspondence of the inequality factors classification between the traditional framework and the new proposal: since it is possible to obtain different classifications, a future extension of this study could check the extent of these differences and propose an analysis with respect to the components of the decomposed Gini and the number of subgroups.

4 Case Study

In order to illustrate the advantages of our proposal, we present a case study related to the Italian households for the 2014. The data are from the Survey on Households Income and Wealth, a multidimensional survey on Italian households performed every two years by the Bank of Italy. The study analyses the income inequality among the Italian households, divided into subgroups by means of two of the main determinants of inequality: the area of residence of the household and the educational level of the head of household. In order to evaluate the effect of the number of subgroups on inequality between, we consider the cases $k = 2, 3, 5$.

Table 4 illustrates the income mean, the population share and the income share for the two partitions. From Table 4 it is possible to observe some well known stylized facts of income inequality in Italy, clearly evident from the values $\bar{y}_i$ and from the differences $(p_i - s_i)$. In the case of equidistribution we have $p_i = s_i$, while increasing differences $(p_i - s_i)$ suggest increasing levels of inequality, with $p_i > s_i$ $(p_i < s_i)$ indicating that the ith subgroup is a poor (rich) subgroup.

Our focus is on the effects of the differences between the subgroups on total inequality. Table 5 illustrates the Dagum's Gini index decomposition by area of residence. For the case of $k = 2$ subgroups, North and Center form one subgroup, and South and Island the other. When $k = 3$ we divide the North from the Center and for $k = 5$ we divide the North into North West and North East and we separate the South from the Islands.

Table 4 Mean income, population share and income share for Italian households divided by area of residence and by educational level of the head of household, 2014

Area	Mean	p	s	Education	Mean	p	s
North West	33750	0.254	0.279	None	14676	0.03	0.02
North East	35150	0.221	0.229	Elementary	22329	0.20	0.16
Center	32636	0.202	0.226	Middle school	26753	0.37	0.31
South	23365	0.244	0.173	High school	35893	0.26	0.31
Islands	24095	0.081	0.093	University	46641	0.13	0.20

Table 5 Income inequality decomposition by area of residence[a], Italian households 2014

	k	Gw	Gb	Gt
NC, SI	2	0.194	0.107	0.049
N, C, SI	3	0.125	0.139	0.086
NW, NE, C, S, I	5	0.073	0.168	0.109

[a]N north, NW north-west, NE north-east, C center, S south, I islands

Table 6 Inequality between evaluation; area of residence[a], Italian households 2014

	k	I_{G_b}	$I_{G_{b(p)}}$	$I_{G_{b(s)}}$
NC, SI	2	0.306	0.355	0.359
N, C, SI	3	0.397	0.562	0.574
NW, NE, C, S, I	5	0.479	0.568	0.566

[a]N north, NW north-west, NE north-east, C center, S south, I islands

By increasing k, we can observe the usual pattern in inequality decomposition: the decrease of inequality within G_w and the consequent greater importance of inequality between G_b and of overlapping component G_t.

The evaluation of G_b, for the area of residence, is reported on Table 6. When we refer to I_{G_b} it is possible to observe how the evaluation of G_b strictly depends on k: for $k = 2$ we have that the area of residence contributes for the 31% to total inequality, while for $k = 5$ its importance rises to the 48%.

From Table 6 we can also observe how $I_{G_{b(p)}}$ and $I_{G_{b(s)}}$ are not a monotone function of k, since they depend on the minimum inequality within. The new indexes show quite similar results, with the contribution of the geographical dimension ranging from the 36% for $k = 2$ to the 50–57% for $k = 5$.

By means of $I_{G_{b(p)}}$ and $I_{G_{b(s)}}$, that is by using an empirical maximum, we are able to obtain an evaluation less influenced by k, more robust to the number of subgroups, and therefore more able to highlight the contribution to the overall inequality attributable to the area of residence.

Furthermore, by using an empirical maximum we are able to assess the relevance of the inequality factor, in this case the area of residence, on real data which we are analysing, not with respect to a theoretical situation, where the risk is to have a relevant underestimation of the inequality between.

The results related to the decomposition by educational level of the head of household are reported on Table 7. For the case of $k = 2$ we distinguish between without or with high school diploma, when $k = 3$ we split high school and university degree, while for $k = 5$ we add two further subgroups, one for elementary school and one for the absence of an educational level. We can observe how the components G_w, G_b, G_t show a behaviour similar to the previous case, however we can note how G_b has a greater importance, while G_t is smaller: two signals of a stronger relevance of the educational dimension.

Moving to Table 8 for the evaluation of the inequality between, I_{G_b} confirms the importance of the educational level, showing higher levels with respect to Table 2.

Table 7 Income inequality decomposition by educational level[a] of the head of household, Italian households 2014

	k	Gw	Gb	Gt
NEM, HU	2	0.162	0.149	0.038
NEM, H, U	3	0.130	0.171	0.049
N, E, M, H, U	5	0.081	0.200	0.069

[a]N none, E elementary, M middle school, H high school, U university

Table 8 Inequality between evaluation: educational level[a] of the head of household, Italian households 2014

	k	I_{G_b}	$I_{G_{b(p)}}$	$I_{G_{b(s)}}$
NEM, HU	2	0.426	0.590	0.626
NEM, H, U	3	0.487	0.613	0.568
N, E, M, H, U	5	0.570	0.615	0.613

[a]N none, E elementary, M middle school, H high school, U university

Also the new indexes are higher, but their increase with respect to the results of Table 6 is less accentuated.

By comparing I_{G_b} to $I_{G_{b(p)}}$ and $I_{G_{b(s)}}$, it is clearly possible to observe one of the advantages of our proposal: the number of subgroups k only slightly affects $I_{G_{b(p)}}$ and $I_{G_{b(s)}}$, while it more strongly influences the traditonal evaluation I_{G_b}.

By comparing the results related to the two decompositions, we get that the educational dimension is considered an inequality factor more important than the geographical dimension by all indexes. It is however important to observe how, within the new proposals, the difference between the two factors is not so high as on the basis of I_{G_b}.

Finally, it is relevant to stress how, in both cases the new indexes attribute to the inequality factors a stronger role, overcoming the usual underestimation and truly reflecting the effective importance of these determinants of total inequality.

5 Conclusions

We propose to modify the traditional evaluation of the inequality between population subgroups by introducing a maximum compatible with the observed data. Our purpose is to assess the determinants of inequality with respect to the observed data, and not by referring to the unrealistic case of equidistributed subgroups.

Two new indexes are illustrated and their behaviour is analysed by means of a simulation study and also with respect observed data from the income distribution of the Italian households. Our proposal allows to strongly reduce the effect of the number of subgroups on the evaluation of inequality between and to overcome the usual underestimation of the importance of the inequality factors.

We believe that the foundation of the new indexes on the observed data represents an improvement for our knowledge of the inequality structure and a relevant complement to the traditional evaluation of the inequality between subgroups.

References

1. Bhattacharya, N., Mahalanobis, B.: Regional disparities in household consumption in India. J. Am. Stat. Ass. 143–161 (1967)
2. Costa, M.: Transvariation and inequality between subpopulations in the Dagum's Gini index decomposition. Metron **67**, 120–134 (2009)
3. Dagum, C.: Inequality measures between income distributions with applications. Econometrica 1791–1803 (1980)
4. Dagum, C.: Gini ratio. The New Palgrave Dictionary of Economics. Mac Millian Press, London (1987)
5. Dagum, C., Zenga, M.: Income and Wealth Distribution Inequality and Poverty. Springer, Berlin (1990)
6. Dagum, C.: A new decomposition of the Gini income inequality ratio. Empir. Econ. **22**, 515–531 (1997)
7. Elbers, C., Lanjouw, P., Mistiaen, J.A., Ozler, B.: Reinterpreting between-group inequality. J. Econ. Inequal. **6**, 231–245 (2008)
8. Gini, C.: Sulla misura della concentrazione e della variabilit dei caratteri. Atti R Ist Veneto Sci Lett Arti **73**, 1203–1248 (1914); English translation in Metron **63**, 3–38 (2005)
9. Giorgi, G.M.: Bibliographic portrait of the Gini concentration ratio. Metron **48**, 183–221 (1990)
10. Giorgi, G.M.: Gini's scientific work: an evergreen. Metron **63**, 299–315 (2005)
11. Giorgi, G.M.: The Gini inequality index decomposition. An evolutionary study. In: Deutsch, J., Silber, J. (eds.) The Measurement of Individual Well-Being and Group Inequalities. Routledge, London (2011)
12. Kanbur, R.: The policy significance of inequality decompositions. J. Econ. Inequal. **4**, 367–374 (2006)
13. Mehran, F.: A statistical analysis of income inequality based on a decomposition of the Gini index. In: Proceedings of the 40th ISI Session International Statistical Institute, Warsaw (1975)
14. Yitzhaki, S., Lerman, R.: Income stratification and income inequality. Rev. Income Wealth **37**, 313–329 (1991)
15. Yitzhaki, S.: More than a dozen alternative ways of spelling Gini. Res. Econ. Inequal. **8**, 13–30 (1998)

Basketball Analytics Using Spatial Tracking Data

Marica Manisera, Rodolfo Metulini and Paola Zuccolotto

Abstract Spatial tracking data are used in sport analytics to study the players' position during the game in order to evaluate game strategies, players' roles, performance, also in prospect. From the broad fields of statistics, mathematics, information science and computer science it is possible to draw theories and methods useful to produce innovative results based on speed, distance, players' separation trajectories. In basketball, spatial tracking data can be combined with play-by-play data, joining results on spatial movements to team performance. In this paper, using tracking data from basketball, we study the spatial pattern of players on the court in order to contribute to the literature of data mining methods for tracking data analysis in sports, with the final objective of suggesting new game strategies to improve team performance.

Keywords Sport science · Performance analysis · Players' position · Players' trajectories · Convex hulls · Cluster analysis

1 Introduction

The study of the players' position during the game is gaining relevance in the discipline of sport science [13, 39] due to the availability of spatial tracking data, that are used to investigate game strategies, players' roles and performance. Information Technology Systems (ITS) permit to collect a large amount of different types of spatio-temporal data from a game: play-by-play data, which report a sequence of

M. Manisera · R. Metulini (✉) · P. Zuccolotto
Department of Economics and Management, University of Brescia,
C.da S. Chiara, 50, 25125 Brescia, Italy
e-mail: rodolfo.metulini@unibs.it

M. Manisera
e-mail: marica.manisera@unibs.it

P. Zuccolotto
e-mail: paola.zuccolotto@unibs.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_23

relevant events occurring during a game, and tracking data, capturing the movements and trajectories of players on the court (or the ball).

On one hand, play-by-play data report events that can be broadly categorized as player events, such as passes and shots as well as technical events, for example fouls and time-outs. These can be used, adopting a data-driven approach, to identify the drivers that affect the probability to win a game, study the interactions among players, identify central players in a team or investigate the impact of specific situations on the performance [40]. Carpita et al. [4, 5] used machine learning tools and principal component analysis in order to identify the drivers that affect the probability to win a football game. Social network analysis has been used to capture the interactions between players [38]. Passos et al. [24] used centrality measures with the aim to identify central (or key) players, and to estimate the interaction and the cooperation between team members in water polo. In soccer, Cintia et al. [8] observed players' behaviour on the pitch. They predict the outcome of a long-running tournament such as the Italian major league using simple network measures. Moreover, Cintia et al. [7] proposed and computed a pass-based performance indicator that strongly correlates with the success of the team.

On the other hand, tracking data are collected using optical- or device-tracking and processing systems. Once tracking data are available, the analysis of players' movements should consider several aspects, for example the interdependency of one player's trajectory with the other players' movements. A branch of literature focuses on the analysis of synchronized movements. The trajectory of a single player, in fact, depends on a large amount of factors and on the trajectories of all the other players on the court, both team-mates and opponents. These interactions among players have been studied from the perspective of physical psychology [37], where players in court represent agents that face with external factors [25, 36]. In addition, typically, players' movements are determined by their role in the game. Predefined plays are used in many team sports to achieve some specific objectives. Moreover, team-mates who are familiar with each other's playing style may develop ad-hoc productive interactions that are used repeatedly and experts want to explain why, when and how specific movement behaviour is expressed because of tactical behaviour. Brillinger [3] addressed the question of how to analytically describe the spatio-temporal movement of particular sequences of passes.

Another complex task is to translate results into suggestions for sports experts, on how to improve game strategies in order to win a game. Analysts want to explain and observe cooperative movement patterns in reaction to a variety of factors, such as coach strategies and specific play-books. A useful approach in this regard consists in segmenting a game into phases, as it facilitates the retrieval of relevant moments of the game. Perin et al. [26] visually segmented a football game into different phases while Metulini et al. [20] segmented a basketball game into phases using a cluster analysis. A key factor in relation to teams' performance is how players control space. Many works are devoted to analyse how the space is occupied by players—when attacking and when defending—or in crucial moments of the game. Examples can be found in football [9, 23] or in futsal [10, 35].

In order to communicate the information extracted from the spatio-temporal data, visualization tools are required. Perin et al. [26] developed a system for visual exploration of phases in football, Sacha et al. [29] present a visual analysis system for interactive recognition of football patterns and situations. Notable works include data visualization in ice hockey [27] and tennis [28]. The most common approach to give a graphical description of spatio-temporal data is to use heat maps. Typical examples in the literature show the spread and range of a shooter [12] or count how many times a player lies in specific court zones. More recently, dynamic approaches have been proposed to visualize aggregated information displaying the time dimension: for example, Theron and Casares [33] employed tools for the analysis of players' movements and Metulini [18] investigated the use of motion charts for visualizing movements of basketball players' on the court.

The aim of this paper is to study the players' position on the court and contribute, with our results, to the literature of data mining methods for tracking data analysis in team sports, with the final objective of suggesting new useful strategies to improve team performance. Using a basketball case study, and having the availability of the players' spatio-temporal trajectories extracted from Geographical Positioning Systems (GPS), we (1) visualize the synchronized movements of players around the court and (2) identify different game phases using a cluster analysis, in which each cluster defines a game phase (because it groups all the moments being homogeneous in terms of spacings among players). We then characterize each cluster in terms of players' position on the court, define whether each cluster corresponds to defensive or offensive actions, and compute the transition matrices in order to examine the probability of switching from one cluster to another one from time t to time $t + 1$.

The paper is organized as follows. Section 2 presents a description of our data (Sect. 2.1) and outlines our research questions (Sect. 2.2). The proposed methodology, data analysis and results are in Sect. 3 while conclusions and future developments are in Sect. 4.

2 Data, Methods and Research Questions

In this section, we describe the data used in this paper (Sect. 2.1) and present our research questions (Sect. 2.2).

2.1 Tracking Data and Play-by-Play Data

Object trajectories capture the movement of players and the ball. Players' trajectories are retrieved using optical- or device-tracking and processing systems. Optical tracking systems use fixed cameras to collect the players' movements, and the images are then processed to compute the trajectories [1]. There are several commercial vendors who supply tracking services to professional sport teams and leagues [17, 34].

Device-tracking systems rely on devices that infer their location, and are attached to the players' clothing or embedded in the ball or puck. These systems are based on GPS [6]. The resulting dataset is dense, because GPS collect data at high temporal resolution. The adoption of this technology and the availability to researchers of the resulting data depend on various factors, particularly commercial and technical, such as the costs of installation and maintenance and the legislation adopted by the sport associations. This data acquisition may be partially restricted in some diffused team sports (as it was for example in soccer until 2015) while allowed for others.

Play-by-play is a sequence of significant events that occur during a game. Events can be broadly categorized as player events such as passes and shots; and technical events, for example fouls, time-outs, and start/end of period. Event logs are qualitatively different from the player trajectories in that they are not dense since samples are only captured when an event occurs. However, they can be semantically richer as they include details like the type of event and the players involved. Typically, in basketball, play-by-play data consist of a collection of about five hundreds events per game. The collection includes events such as made shots, missed shots, rebounds, fouls, start/end of the period, etc... Play-by-play data can be obtained, for example by means of webscraping procedures run on specific sport league websites.

Basketball is a sport generally played by two teams of five players each on a rectangular court (28 m $\times$ 15 m). The game, according to International Basketball Federation (FIBA) rules, lasts 40 min, and is divided in four periods of 10 min each. The objective is to shoot a ball through a hoop 46 cm in diameter and mounted at a height of 3.05 m to backboards at each end of the court.

The data we used in the analysis refer to a friendly game played on March 22th, 2016 by two Italian teams in the C-gold league, the fourth league in Italy. Data are referred to the home team. MYagonism (https://www.myagonism.com/) was in charge to set up a system to record the players' position on the court during the game. Each player worn a microchip that, having been connected with machines built around the court, collected his position (in pixels of 1 m^2) in both the x-axis (court length) and the y-axis (court width), as well as in the z-axis (i.e. how high the player jumps). The position of the players has been detected with an average frequency of about 37 Hz (i.e. 37 times every second). During the match, a total of six players rotated on the court. The system recorded a series of 133,662 measurements, each one referring to one among positioning, velocity or acceleration in one among x-, y- or z-axis, for a specific player in a specific time instant. Tracking systems retrieve data with a potential margin of error. In order to clean data in the (possible) presence of outliers and noise, we refer to the approach of the Kalman filter. The Kalman filter is an algorithm that predicts, using a set of measurements observed over time and containing statistical noise, values of a variable that tend to be more accurate than the single measurements. It is traditionally used in sport applications [16]. In our data, x-, y- and z-axes measurements have been smoothed with a Kalman filter. Measurements are detected with a non-constant frequency; in addition, measurements of different players are recorded at different time instants. As a consequence, the data matrix contains every millisecond of the game (a row of the data matrix identifies a millisecond), and we attributed the last measurement

available to players not detected in that millisecond. Moreover, the players' positions are detected also during the moments when the game is off: these rows have to be removed from the data matrix. However, there is no variable labelling milliseconds when the game is off, so we needed rules to identify moments to be filtered out. We filtered the rows of the data matrix by dropping the pre-game, the half-time break and the post-game periods, using the procedure described in Metulini [19].

The final data matrix $\mathbf{X}$ counts for 3, 485, 147 total rows, where each row correspond to an active millisecond. The data matrix $\mathbf{X}$, furthermore, is made by several variables (in column), each variable reporting the values of one among positioning, velocity or acceleration in one among x-, y- or z-axis, for one among the six players.

2.2 Research Questions

The overall objective of our research is to visualize and characterize the movement of basketball players around the court by finding relevant types of movement patterns that could affect the team performance.

Going into detail, the first specific objective is to find and demonstrate the usefulness of a visual tool approach in order to extract preliminary insights from trajectories. In this respect, we aim to visualize the synchronized movement of players and to characterize their position around the court in order to supply experts and analysts with a useful tool in addition to traditional statistics, and to confirm the interpretation of evidence from other methods of analysis. Some preliminary results, obtained using motion charts to visualize the movements of players around the court, allowed to identify differences in spacing structure among offensive and defensive plays [18, 21]. Such interesting results must be further developed, by analysing both team-mate and opponents trajectories and adding in the tracking data the ball's position.

Another research aim is to segment the game into phases. Specifically, our idea is to find, through a cluster analysis, a number of groups each identifying a specific spatial pattern, in order to find any regularities and synchronizations in players' trajectories, by decomposing the game into homogeneous phases in terms of spatial relations. In this paper, we will show results from an exploratory analysis using tracking data from one basketball game. We plan to extend the analysis to multiple games. Moreover, we aim to match play-by-play data and trajectories, in order to extract insights on the relations between particular spatial pattern and the team performance, and to include the effect of the ball's position on the players' movements.

3 Basketball Data Analysis and Results

We first use motion charts to visualize the synchronized spatio-temporal movements of players around the court. There are several softwares providing the possibility to reproduce motion charts, more or less intuitive, open source or requiring a license

(*Gapminder world*, *Google docs gadget*, *Trend compass*, and JMP from SAS institute). In addition, motion charts can be created through web programming languages using *Google application programming interface*, *Google* API, *Flash* or *HTML5*. We decided to opt for `gvisMotionChart` in *R* [11], because it outperforms alternatives in terms of open source and friendliness and allows to easily import data. A video tutorial showing players' trajectories using motion charts can be found at: http://bodai.unibs.it/bdsports/Ricerca3.htm. Motion charts have been applied in several fields, as students' learning processes [32] and linguistic changes [15], insurance [14] and development economics [30], medicine [31] and hydrology [2]; to the best of our knowledge, they have never been applied to basketball within a scientific study. In this paper, the evidence drawn from motion charts, that cannot be shown here for obvious reasons, is summarized by average distances among players and statistics computed on convex hulls areas, distinguishing defensive from offensive plays. To compute average distances and convex hulls areas, we add new variables to the data matrix **X**. The average distance (in meters) is defined, for player i and player j, and for a specific millisecond, as:

$$dist_{ij} = \sqrt{(pos\_x_i - pos\_x_j)^2 + (pos\_y_i - pos\_y_j)^2}$$

where *pos* stays for position, x and y stay for the the axes. The convex hulls areas, for a specific millisecond, is defined as the area computed on the following set of values:

$$(pos\_x_1, pos\_x_1), (pos\_x_2, pos\_x_2), (pos\_x_3, pos\_x_3), (pos\_x_4, pos\_x_4), (pos\_x_5, pos\_x_5)$$

where 1,2, …,5 denote the players on the court in that millisecond. Motion charts applied to our data show differences in the spacing structure of players among offensive and defensive plays. We defined whether each time instant corresponds to an offensive or a defensive play looking to the average coordinate of the five players on the court. More in detail, we separate the court in two sides along the half court line ($x - axis = 0$); In the first half of the game, rows of the data matrix are assigned to defense if the average $x - axis$ of the five players on the court has negative sign, to offense if the average $y - axis$ of the five players on the court has positive sign. In the second half of the game, teams change court side, so we invert the rule to assign rows to game phases. The evidence is summarized in Table 1, which reports the statistics describing the convex hulls areas and the average distances.

Results clearly highlight that average distances among players and convex hulls areas are larger in offensive plays than in defensive plays.

To confirm previous evidence, Figs. 1 and 2 report the convex hulls for selected snapshots from, respectively, the first offensive play and the first defensive play of the game. Once again, players are more spread around the court in offensive plays.

Then, we applied a k-means Cluster Analysis in order to segment the game into phases. Cluster analysis is a method of grouping a set of objects in such a way the objects in the same group (cluster) are more similar to each other than to those in other groups. In our case, the objects are represented by the rows of the data matrix

Table 1 Average distances among players and convex hulls areas for the full game, for defensive and offensive plays

	Average distances		Convex hull area	
	Attack	Defense	Attack	Defense
Min	2.296	0.400	1.000	1.000
1st Qu.	6.372	4.309	30.000	14.000
Median	7.235	5.086	41.000	20.500
Mean	7.250	5.680	42.590	28.550
3rd Qu.	8.132	6.523	53.000	33.500
Max	13.947	14.260	138.500	180.000

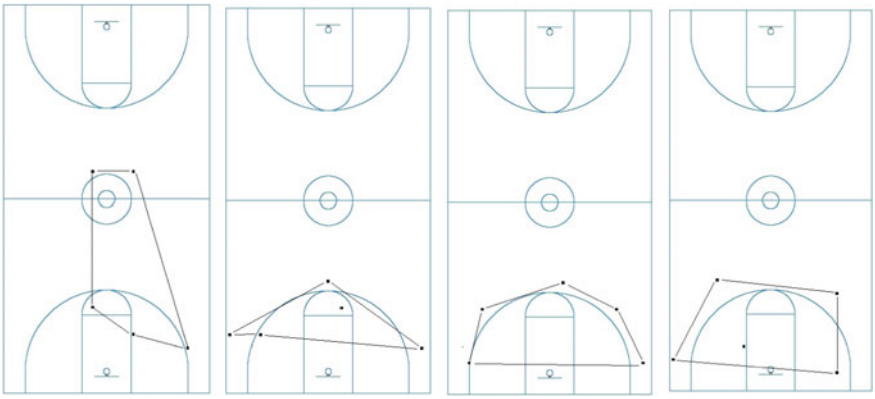


Fig. 1 Convex hull for selected snapshots related to the first offensive play of the game

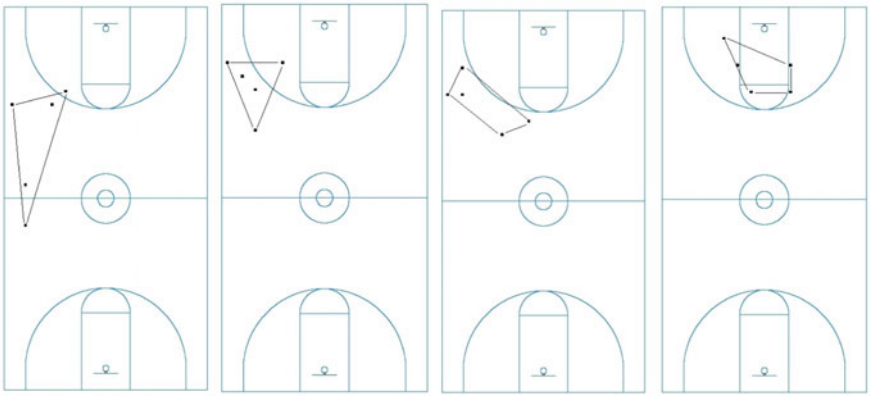


Fig. 2 Convex hull for selected snapshots related to the first defensive play of the game

X, expressing time instants (milliseconds), while the similarity among time instants is computed using, as variables, the ten distances between the two players of each pair of players. In details, we start with a subset of the data matrix **X** containing all the rows corresponding to time instants where players 1, 2, 3, 4 and 5 were on the court, and the ten variables reporting the distances between the two players of each pair:

$$dist_{12}, dist_{13}, dist_{14}, dist_{15}, dist_{23}, dist_{24}, dist_{25}, dist_{34}, dist_{35}, dist_{45},$$

where subscripts denote the players' pair. Clusters centroids are k randomly chosen time instants. The criteria to assign a time instant to a centroid is based on their similarity in terms of the ten variables defined above.

Our aim is to characterize the spatial pattern of the players on the court. We define different game phases, each considering moments being homogeneous in terms of spacings among players. We choose $k = 8$ clusters, based on the value of the between deviance (BD) / total deviance (TD) ratio for different number of clusters ($BD/TD = 50\%$ and relatively low increments for increasing k, for $k \geq 8$). The first cluster (C1) embeds 13.56% of the observations (i.e. 13.56% of the total game time). The other clusters, named C2, …, C8, have size of 4.59, 14.96, 3.52, 5.63, 35.33, 5.00 and 17.41% of the total sample size, respectively.

First, we characterize each cluster in terms of players' position on the court. We used Multidimensional Scaling (MDS) in order to plot the differences between the groups in terms of their position on the court. Using the MDS algorithm we aim to place each player in N-dimensional space such that the between-player average distances are preserved as well as possible. Each player is then assigned coordinates in each of the N dimensions. Since the basketball court have two dimensions (width and length), we choose $N = 2$ in order to guarantees the best visual interpretability. In detail, for each cluster, we apply a MDS on a 5×5 matrix M reporting the average distance computed averaging over the distances between two players of each of the 5^2 pairs. We obtained a scatterplot showing each player in a 2-dimensional space such that the average distances between players are preserved (Fig. 3). We observe remarkable differences among different game phases (clusters) in the players' position on the court. In C1 and C5 players are equally spaced along the court. C6 also highlights an equally spaced structure, but the five players are more closed by. In other clusters we can see a spatial concentration: for example in C2 players 1, 5 and 6 are closed by while in C8 this is the case of players 1, 2 and 6.

Figure 4 reports cluster profile plots and helps us to better interpret the spacing structure in Fig. 3, characterizing groups in terms of average distances among players. Profile plot for C6 confirms that players are more close by, in fact, all the distances are smaller than the average distance. At the same way, C2 presents distances among players 1, 5 and 6 smaller than the average.

After having defined whether each moment corresponds to an offensive or a defensive action looking to the average coordinate of the five players on the court, we also found that some clusters represent offensive actions rather than defensive. More precisely, we found that clusters C1, C2, C3 and C4 mainly correspond to

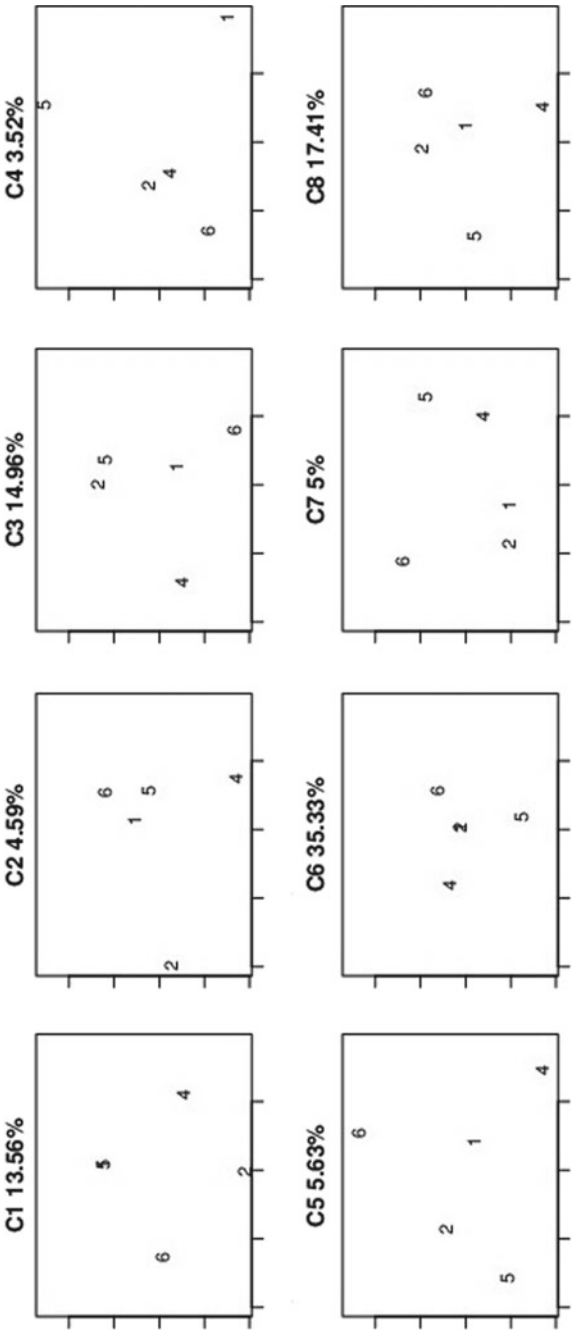


Fig. 3 Map representing, for each of the 8 clusters, the average position in the $x - y$ axes of the five players, using MDS. Percentages report the proportion of instants in the dataset belonging to each cluster. Dimension 1 is reported in the x-axis and dimension 2 in the y-axis

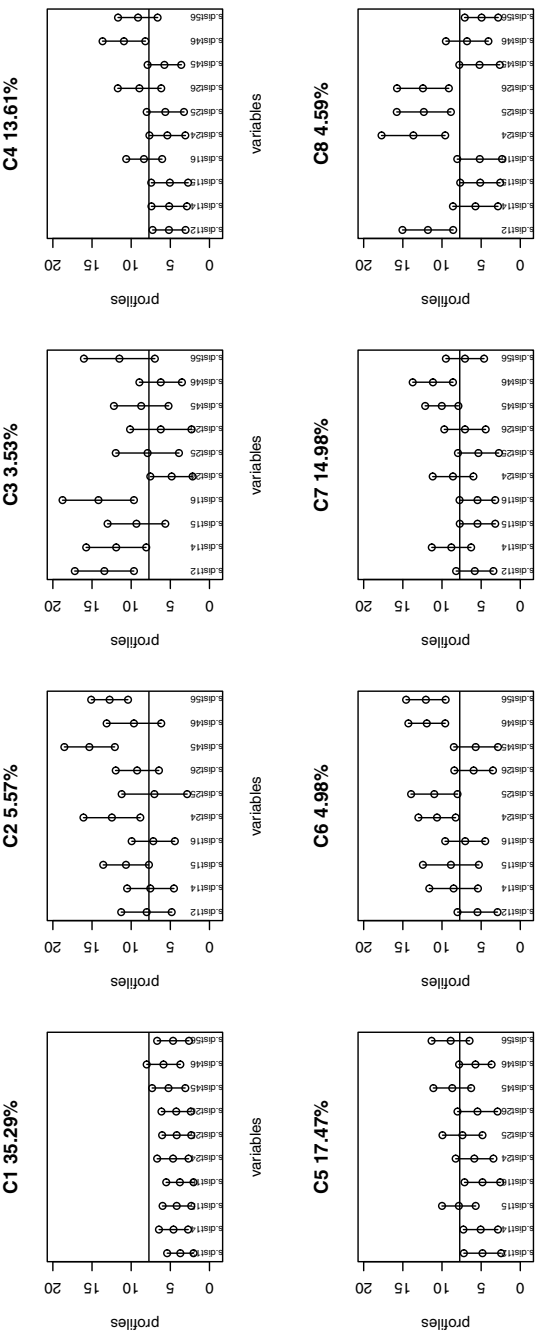


Fig. 4 Profile plots representing, for each of the 8 clusters, the average distance among each pair of players. Variables are in the x-axis, profiles in the y-axis

NA	1	2	3	4	5	6	7	8
1	0	10.71	23.53	47.83	0	20.83	31.25	20.23
2	0.77	0	9.15	0	1.85	2.08	8.33	2.89
3	31.54	42.86	0	8.7	44.44	20.83	18.75	20.23
4	6.15	3.57	1.96	0	7.41	0	10.42	1.16
5	0.77	3.57	16.99	17.39	0	0	16.67	8.09
6	27.69	7.14	18.95	0	1.85	0	0	43.93
7	15.38	21.43	3.92	4.35	18.52	0	0	3.47
8	17.69	10.71	25.49	21.74	25.93	56.25	14.58	0

Fig. 5 Transition matrix reporting the relative frequency subsequent moments $(t, t + 1)$ report a switch from a group to a different one

offensive actions (respectively, for the 85.88, 85.91, 73.93 and 84.62% of the times in each cluster) and C6 strongly corresponds to defensive actions (85.07%). Offensive clusters show larger players' spacings than in the defensive cluster. A motivation for this behaviour could be that players in defense have the objective to narrow the opponents' spacings in order to limit their play, while the aim of the offensive team is to maintain large distances among team-mates, to increase the propensity to shot with good scoring percentages. Anyhow, these findings go on the same direction of those of the convex hulls.

Figure 5 shows the transition matrix, which reports the relative frequency in which subsequent milliseconds report a switch from a cluster to a different one. It emerges that for the 31.54% of the times C1 switches to a new cluster, it switches to C3, another offensive cluster. C2 switches to C3 for the 42.85% of the times. When the defensive cluster (C6) switches to a new cluster, it switches to C8 for the 56.25% of times.

4 Conclusions and Future Developments

In recent years, spatial tracking data have been used in sport analytics to study the players' position during the game in order to investigate game strategies, players' roles, players' and teams' performance. In particular, coaches, sports experts and analysts have received benefits from the availability of large amounts of data to use in team sports analysis. This has increased the possibility to extract important information on team performance from every single game. The advent of information technology systems permits to match play-by-play data and players' trajectories and to analyse teams' performance with a variety of approaches. Having the trajectories of the players and the play-by-play available, and inspired by the literature based on the data-driven methods as well as by the increasing interest in visualizing data, we analysed the movement and the players' position using visual tools and data-mining techniques, with the aim of finding regularities and patterns.

First, we summarized results from the use of motion charts and, after having separated offensive plays from defensive plays, we computed average distances between

players and convex hulls areas. The most promising result relates to convex hulls' analysis. We found that players are more spread around the court in offensive plays rather than in defensive plays. At the moment, we are carrying out further analysis in order to better understand the logic underpinning this regularity, by examining the time series of the convex hulls areas of both teams together. This will answer the question whether the defensive team has success in limiting the spacing of the offensive team. Results of such analysis, aiming to assess whether and how the two teams pursue their strategies, and how the achievement of their strategy affects their performance, may be of interest for coaches and experts.

Second, we used a cluster analysis approach to group spatial tracking data in order to identify specific patterns of movement. We segmented the game into phases of play and we characterized each phase in terms of spacing structure among players, relative distances and whether they represent an offensive or a defensive action, finding substantial differences among different phases. These results shed light on the potentiality of data-mining methods for tracking analysis in team sports.

Results are promising. Future research will aim at finding regularities between trajectories and players' and team performance [22] by analysing tracking data of both team-mates, opponents, and the ball, for multiple games. This is essential to enhance the understanding of the multivariate and complex structure of trajectories in association with team performance but requires the availability of a big amount of high quality tracking data.

Acknowledgements Research carried out in collaboration with the Big & Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016, title: "Big Data Analytics in Sports", www.bodai.unibs.it/bdsports/), granted by Fondazione Cariplo and Regione Lombardia. Authors would like to thank MYagonism (https://www.myagonism.com/) for having provided the data.

References

1. Bradley, P., O'Donoghue, P., Wooster, B., Tordoff, P.: The reliability of ProZone MatchViewer: a video-based technical performance analysis system. Int. J. Perform. Anal. Sport **7**(3), 117–129 (2007)
2. Bolt, M.D.: Visualizing water quality sampling-events in Florida. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. **2**(4), 73 (2015)
3. Brillinger, D.R.: A potential function approach to the flow of play in soccer. J. Quant. Anal. Sport. **3**(1), 3 (2007)
4. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Football mining with R. Data Min. Appl. R (2013)
5. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Discovering the drivers of football match outcomes with data mining. Qual. Technol. Quant. Manag. **12**(4), 561–577 (2015)
6. Catapult USA Sports Ltd. - Wearable Technology for Elite Sports (2015). http://www.catapultsports.com/
7. Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., Malvaldi, M.: The harsh rule of the goals: data-driven performance indicators for football teams. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA) 36678, pp. 1–10 (2015)

8. Cintia, P., Rinzivillo, S., Pappalardo, L.: A network-based approach to evaluate the performance of football teams. In: Machine Learning and Data Mining for Sports Analytics Workshop, Porto, Portugal (2015)
9. Couceiro, M.S., Clemente, F.M., Martins, F.M., Machado, J.A.T.: Dynamical stability and predictability of football players: the study of one match. Entropy **16**(2), 645–674 (2014)
10. Fonseca, S., Milho, J., Travassos, B., Araujo, D.: Spatial dynamics of team sports exposed by voronoi diagrams. Hum. Mov. Sci. **31**(6), 1652–1659 (2012)
11. Gesmann, M., de Castillo, D.: Package 'googleVis'. Interface between R and the Google chart tools (2013)
12. Goldsberry, K.: Courtvision: new visual and spatial analytics for the NBA. In: 2012 MIT Sloan Sports Analytics Conference (2012)
13. Gudmundsson, J., Horton, M.: Spatio-temporal analysis of team sports. ACM Comput. Surv. (CSUR) **50**(2), 22 (2017)
14. Heinz, S.: Practical application of motion charts in insurance (2014)
15. Hilpert, M.: Dynamic visualizations of language change. Int. J. Corpus Linguist. **16**(4), 435–461 (2011)
16. Kim, J.Y., Kim, T.Y.: Soccer ball tracking using dynamic Kalman filter with velocity control. In: Sixth International Conference on Computer Graphics, Imaging and Visualization, CGIV'09, pp. 367–374. IEEE (2009)
17. Impire, A.G.: (2015). http://www.bundesliga-datenbank.de/en/products/
18. Metulini, R.: Spatio-temporal movements in team sports: a visualization approach using motion charts. Electron. J. Appl. Stat. Anal. **10**(3), 809–831 (2017)
19. Metulini, R.: Filtering procedures for sensor data in basketball. Stat. Appl. **15**(2), 133–150 (2017)
20. Metulini, R., Manisera, M., Zuccolotto, P.: Space-time analysis of movements in basketball using sensor data. In: Statistics and Data Science: New Challenges, New Generations SIS2017 Proceeding. Firenze Uiversity Press. e-ISBN: 978-88-6453-521-0 (2017)
21. Metulini, R., Manisera, M., Zuccolotto, P.: Sensor analytics in basketball. In: Proceedings of the 6th International Conference on Mathematics in Sport. ISBN 978-88-6938-058-7 (2017)
22. Metulini, R., Manisera, M., Zuccolotto, P.: Modelling the dynamic pattern of surface area in basketball and its effects on team performance. J. Quant. Anal. Sport **14**(3), 117–130 (2018)
23. Moura, F.A., Martins, L.E.B., Anido, R.D.O., De Barros, R.M.L., Cunha, S.A.: Quantitative analysis of Brazilian football players' organisation on the pitch. Sports Biomech. **11**(1), 85–96 (2012)
24. Passos, P., Davids, K., Araujo, D., Paz, N., Minguens, J., Mendes, J.: Networks as a novel tool for studying team ball sports as complex social systems. J. Sci. Med. Sport **14**(2), 170–176 (2011)
25. Passos, P., Araujo, D., Volossovitch, A.: Performance Analysis in Team Sports. Routledge, London (2016)
26. Perin, C., Vuillemot, R., Fekete, J.D.: SoccerStories: a kick-off for visual soccer analysis. IEEE Trans. Vis. Comput. Graph. **19**(12), 2506–2515 (2013)
27. Pileggi, H., Stolper, C.D., Boyle, J.M., Stasko, J.T.: Snapshot: visualization to propel ice hockey analytics. IEEE Trans. Vis. Comput. Graph. **18**(12), 2819–2828 (2012)
28. Polk, T., Yang, J., Hu, Y., Zhao, Y.: Tennivis: visualization for tennis match analysis. IEEETrans. Vis. Comput. Graph. **20**(12), 2339–2348 (2014)
29. Sacha, D., Stein, M., Schreck, T., Keim, D.A., Deussen, O.: Feature-driven visual analytics of soccer data. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 13–22 (2014)
30. Saka, C., Jimichi, M.: Inequality evidence from accounting data visualisation (2015)
31. Santori, G.: Application of interactive motion charts for displaying liver transplantation data in public websites. Transplant. Proc. **46**(7), 2283–2286 (2014)
32. Santos, J.L., Govaerts, S., Verbert, K., Duval, E.: Goal-oriented visualizations of activity tracking: a case study with engineering students. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 143–152. ACM (2012)

33. Theron, R., Casares, L.: Visual analysis of time-motion in basketball games. In: International Symposium on Smart Graphics, pp. 196–207. Springer, Berlin Heidelberg (2010)
34. Tracab Corporation. Player Tracking System (2015). http://chyronhego.com/sports-data/player-tracking
35. Travassos, B., Araujo, D., Duarte, R., McGarry, T.: Spatiotemporal coordination behaviors in futsal (indoor football) are guided by informational game constraints. Hum. Mov. Sci. **31**(4), 932–945 (2012)
36. Travassos, B., Davids, K., Araujo, D., Esteves, P.T.: Performance analysis in team sports: advances from an ecological dynamics approach. Int. J. Perform. Anal. Sport **13**(1), 83–95 (2013)
37. Turvey, M.T., Shaw, R.E.: Toward an ecological physics and a physical psychology. The Science of the Mind: 2001 and Beyond, pp. 144–169 (1995)
38. Wasserman, S., Katherine, F.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)
39. Wu, S., Bornn, L.: Modeling offensive player movement in professional basketball. Am. Stat. **72**(1), 72–79 (2018)
40. Zuccolotto, P., Manisera, M., Sandri, M.: Big data analytics for modeling scoring probability in basketball: the effect of shooting under high-pressure conditions. Int. J. Sport Sci. Coach. **13**(4), 569–589 (2017)

New Fuzzy Composite Indicators for Dyslexia

Isabella Morlini and Maristella Scorza

Abstract Composite indicators should ideally identify multidimensional concepts that cannot be captured by a single variable. In this paper, we suggest a method based on fuzzy set theory for the construction of fuzzy synthetic indexes of dyslexia, using the set of manifest variables measured by means of reading tests. A few criteria for assigning values to the membership function are discussed, as well as criteria for defining the weights of the variables. An application regarding the diagnosis of dyslexia in primary and middle school in Italy is presented. In this application, the fuzzy approach is compared with the crisp approach actually used in Italy for detecting dyslexic children in compulsory school.

Keywords Fuzzy composite indicators · Learning disabilities · Membership function · Reading performances · Threshold values

1 Introduction

Dyslexia is a functional deficit that affects the ability to decode a text. In academic learning, the normal acquisition of the process of writing of dyslexic children is affected by an underlying neurobiological disfunction. Thus, dyslexia is typically diagnosed from the end of the second grade, when the process of reading and writing acquisition has been given enough time to be completed. Since this learning disorder has a great impact on the individual's academic achievement and on his/her social life, it is important to detect dyslexic students especially in primary and secondary schools.

I. Morlini (✉)
Department of Economics Marco Biagi, University of Modena and Reggio Emilia, Viale Berengario 51, 41121 Modena, Italy
e-mail: isabella.morlini@unimore.it

M. Scorza
Department of Social Sciences, University of Modena and Reggio Emilia, Viale A. Allegri 9, 42121 Reggio Emilia, Italy
e-mail: maristella.scorza@unimore.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_24

Decoding ability in primary school in Italy and in countries with transparent orthography is currently assessed with the aid of standardized tests requiring the students to read aloud a selected list of words and non-words or a text. The most widely used standardized tests in Italy have been introduced in [15]. Recently, a new screening procedure for identifying impaired decoders in elementary grades has been proposed in [12, 13]. What is important in the use of tests and screening procedures is the way the results are interpreted. One of the defining characteristic of a skilled decoder is that he or she not only is able to spell written words (or non-words) accurately, but also does so rapidly and automatically. An individual who spells accurately but very slowly cannot be considered a skilled decoder. Slow rate of word reading is then characteristic of impaired decoding as well as low accuracy, especially in transparent languages [17]. In Italy, decoding ability is assessed without taking into account both aspects and an individual can be classified as impaired because he or she is able to read words (or non-words) very rapidly, even though he or she misspells a fairly large number of words (or non-words). Individuals with weak decoding skills who are able to read a large number of words, provided they are given ample time, can be erroneously classified as adequate decoders. Many authors have outlined the necessity of considering both speed or fluency and accuracy for a valid assessment of decoding skills and a new challenge in learning disability research is to develop composite indicators that incorporate measures of speed as well as of accuracy [13]. An other challenge is to estimate the dyslexia prevalence in school-age children in Italy. Although disorder in reading is one of the most common neurodevelopmental disorders affecting children, there is still high variability in dyslexia prevalence estimates due to the lack of univocal diagnostic criteria [11].

Since dyslexia is a vague concept and the rigid partition between impaired and not impaired readers does not always reflect reality, in this paper we use the fuzzy set theory for defining new composite indicators that can be used in clinical practice for diagnostic issues and for estimating the disorder prevalence.

The paper is organized as follows. In Sect. 2, we deal with the general problem of obtaining a synthetic fuzzy measure of a latent phenomenon like dyslexia from a set of metric variables. We present two criteria for transforming the values of a variable into fuzzy values. In Sect. 3, we discuss the problem of weighting the variables and aggregating them into a composite indicator. Clearly, the weights should reflect the contribution of each variable to the latent phenomenon. In Sect. 4, we focus on the specific application of measuring dyslexia in compulsory schools in Italy. The gradual transition from skilled to impaired readers can be captured by the fuzzy indexes, as well as the level of risk of being dyslexic. We apply the method to a sample of 3932 students attending elementary and middle schools in Italy. The fuzzy indicators of dyslexia allow us to obtain membership functions that can be compared with the results of one of the currently used diagnostic procedure, which, of course, strictly identifies a student as being dyslexic or not dyslexic. In Sect. 5 we give concluding remarks and outline necessary future work.

2 The Fuzzy Approach

The fuzzy approach was originally proposed by Zadeh [18] in order to model the degree of membership to a certain set. Some applications of this original theory in the social sciences are shown in [16]

Let X be a set of elements $x \in X$. A fuzzy subset A of X is a set of ordered pairs

$$[x, \mu_A(x)] \qquad \forall x \in X \tag{1}$$

where $\mu_A(x)$ is the membership function (m.f.) of x to A in the closed interval $[0, 1]$. If $\mu_A(x) = 0$, then x does not belong to A, if $\mu_A(x) = 1$, then x completely belongs to A. If $0 < \mu_A(x) < 1$, then x partially belongs to A and its membership to A increases according to the values of $\mu_A(x)$. Let us assume that the subset A defines the position of each element with reference to the achievement of the latent concept, e.g. dyslexia. In this case, $\mu_A(x) = 1$ identifies a situation of full achievement of the disease, whereas $\mu_A(x) = 0$ denotes the absence of the disease (a very skilled decoder). A value of $\mu_A(x)$ in the interval $(0, 1)$ represents the degree of uncertainty of being dyslexic. Consider a set of n individuals $i = 1, \ldots, n$ and p metric variables X_s $(s = 1, 2, \ldots, p)$ reflecting the latent phenomenon. In case of dyslexia, these variables are measures of reading performances in standardized tests like the time of reading in seconds, the number of misspelled words or the number of syllables read in a second. Without loss of generality, let us assume that each variable is positively related with that phenomenon, i.e. it satisfies the property the larger the more impaired. If a variable X_s shows a negative correlation (like the number of syllables read in a second) we substitute it with the simple decreasing function transformation

$$f(x_{si}) = \max(x_{si}) - x_{si}. \tag{2}$$

In order to define the m.f. for each variable, it is necessary to identify the extreme situations such that $\mu_A(x) = 0$ (non membership) and $\mu_A(x) = 1$ (full membership) and to define a criterion for assigning the m.f. to the intermediate values. Many criteria have been proposed in literature, in the field of social sciences, for measuring latent concepts like, for example, well-being, satisfaction and poverty [5–7, 10, 19–21]. For the specific purpose of measuring dyslexia, we will consider two specifications. The first one is characterized by the simplicity and for this reason can be straightforward used by all professionals involved in the diagnosis and in the management of specific learning disorders, like neuropsychiatrists, psychologists and education specialists. The second specification is drawn by making an hypothesis about the shape of the function relating the empirical reading performances and the amount of underlying dyslexic deficit. It is a more flexible function requiring the choice of two parameters influencing its shape.

Let us assume that X_s is a metric variable. In the following, for simplicity of notation, we will omit index s. For that variable X, we choose a lower threshold l and an upper threshold u and we define the first m.f. as follows:

$$\begin{cases} \mu_A(x_i) = 0 & x_i \leq l \\ \mu_A(x_i) = \frac{x_i - l}{u - l} & l < x_i < u \\ \mu_A(x_i) = 1 & x_i \geq u \end{cases} \tag{3}$$

In (3) the m.f. is a linear function between the values of the two thresholds. The upper threshold u can be set equal to the normative cut-off used to identify poor performances on academic tests for asymmetric variables. This value is $x_{95\%}$, that is the 95th percentile. The lower threshold l can be set equal to $x_{\min}$, that is the minimum value, or to $x_{5\%}$ (the 5th percentile). The choice of these thresholds will be discussed in the next section, analysing the empirical distribution of the variables.

Alternatively, we may consider the distance $d(x)$ between the value x and dyslexia. If $d(x) = 0$, there is full membership to A, then $\mu_A(x) = 1$. If $d(x) > 0$ then $\mu_A(x) < 1$. Hence, we can write:

$$\mu_A(x) = \frac{1}{1 + d(x)}, \tag{4}$$

If we assume that the relationship between empirical reading performances and learning disorder takes an exponential form, then the distance $d(x)$ can be expressed as

$$d(x) = e^{-a(x-b)}, \tag{5}$$

and the m.f. can be defined as follows:

$$\mu_A(x) = \frac{1}{1 + e^{-a(x-b)}}. \tag{6}$$

Zimmerman [22] highlights that, in general, the relationship between physical measures and perception takes an exponential form. Baliamoune-Lutz [1, 2] uses m.f. (6) to measure human well-being with a fuzzy approach. It is worth noting that in (6) the parameter a ($a \in \Re^+$) represents the extent of uncertainty and the parameter b (with $x_{min} < b < x_{max}$) may be viewed as the point in which the tendency of the subject's attitude changes from rather positive to rather negative. The choice of the parameters in m.f. (6) is somehow more subjective than the choice of u and l in m.f. (3). Moreover, in the application of Sect. 4, different specifications for the parameters a and b have been shown to lead to different results, while slightly chances in u and l (for example, $u = x_{90\%}$ or $u = x_{99\%}$ instead of $u = x_{95\%}$ and $l = x_{\min}$ or $l = x_{10\%}$ instead of $l = x_{5\%}$) have been shown to lead to similar results. Then, the choice of a and b should be made with caution and will be discussed in Sect. 4, also considering that in the literature there are no proposals of estimation procedures for these parameters.

3 The Fuzzy Composite Indicator

The most simple aggregation function is the weighted arithmetic mean [8]:

$$\mu_A(i) = \sum_{s=1}^{p} [\mu_A(x_{si})] \cdot w_s \tag{7}$$

where $w_s > 0$ is the normalized weight that expresses the relative importance of the variables X_s and $\sum_{s=1}^{p} w_s = 1$. In general, the weighting criteria in (7) are:

- equal weights, that imply a careful selection of the variables in order to assure a balance of the different aspects of the latent phenomenon;
- factor loadings, obtained by principal components analysis (PCA) when the first component accounts for a high percentage of the total variance;
- weights obtained from expert judgements;
- weights determined by an Analytic Hierarchy Process [9].

Since the relative importance of each variable measuring the empirical performance in a reading test is still an open question and among professionals involved in the diagnosis of dyslexia there is not consensus about the relevance of these variables (and, in particular, about the relevance of variables measuring accuracy and variables measuring fluency), for fuzzy composite indicators of dyslexia we choose as weights the normalized factor loadings. This weighting method is appropriate since reading tests and, in general, psychometric tests, are designed in order to have an high internal validity. With high internal validity, the first principal component accounts for a high percentage of the total variance. We also suggest an other criterion for the determination of the weights, considering for each variable X_s the fuzzy proportion $g(X_s)$ of the achievement of the target:

$$g(X_s) = \frac{1}{n} \sum_{i=1}^{n} \mu_A(x_{si}). \tag{8}$$

Formula (8) may be viewed as an index of the proportion of the units having (totally or partially) the latent phenomenon [4]. The normalized weights may be determined as an inverse function of $g(X_s)$, in order to give higher importance to rare features in the n units. To avoid excessive weights to the variables with low value of $g(X_s)$ we propose the following weights [3]:

$$w_s = ln\Big[\frac{1}{g(X_s)}\Big] / \sum_{s=1}^{p} ln\Big[\frac{1}{g(X_s)}\Big] \tag{9}$$

Using (9), each variable has a weight sensitive to the fuzzy membership.

Table 1 Frequency distributions of students in each grade

	Elementary school				Middle school		
Grade	II	III	IV	V	VI	VII	VIII
N.	715	472	621	519	922	311	372

4 Fuzzy Indicators of Dyslexia: An Application

We administer the standardized tests Batteries for the Diagnosis of Reading and Spelling Disabilities [15] to 3932 students attending elementary (from grade II) and middle school. We randomly choose schools in Lombardia and Emilia Romagna regions (Northern Italy) and administer the batteries to all students attending these schools (except for the first grade). Table 1 reports the frequency distribution of the students in each grade. In the Batteries for the Diagnosis of Reading and Spelling Disabilities, the metric variables measuring decoding performances are:

- X_1: time (in seconds) in reading the list of words
- X_2: number of words mispronounced in reading the list of words
- X_3: time (in seconds) in reading the list of non-words
- X_4: number of incorrect pronunciations in reading the list of non-words

Figure 1 shows the empirical distribution of the variables. We perform a PCA on the correlation matrix. The first component accounts for 66.5% of the total variance. It is highly correlated with all variables and it is the only component with eigenvalue greater than one. We construct the following fuzzy indicators:

- F_{11}: using m.f. (3) with $l = x_{5\%}$ (the fifth percentile) and $u = x_{95\%}$ (the 95th percentile) in each grade and weights proportional to the factor loadings of the first PCA.
- F_{12}: using m.f. (3) with $l = x_{5\%}$ and $u = x_{95\%}$ in each grade and weights (9).
- F_{21}: using m.f. (6) with $a = 0.5$ and $b = x_{90\%}$ (the 90th percentile) in each grade and weights proportional to the factor loadings of the first PCA.
- F_{21}: using m.f. (6) with $a = 0.5$ and $b = x_{90\%}$ in each grade and weights (9).

In m.f. (3) we use $x_{5\%}$ as the lower threshold instead of $x_{\min}$, since Fig. 1 reveals some outliers also in the left hand side of two empirical distributions. However, further analyses conducted with slightly different choices of u and l show that results are not affected by little changes in these two parameters. In m.f. (6), we choose $a = 0.5$ and $b = x_{90\%}$. As an example, Fig. 2 reports graphical representations of the membership function (for variables X_2 in the third grade) with different sets of parameters. With $a = 0.5$, the m.f. in maximally diversified for values of the variables close to b. We choose $a = 0.5$ in order to have the membership function values more spread out for values of X close to the point in which the performances start changing from positive to rather negative. Analysing the empirical distribution of the variables, we identify this last point as the 90th percentile. However, as Fig. 2 shows, little changes

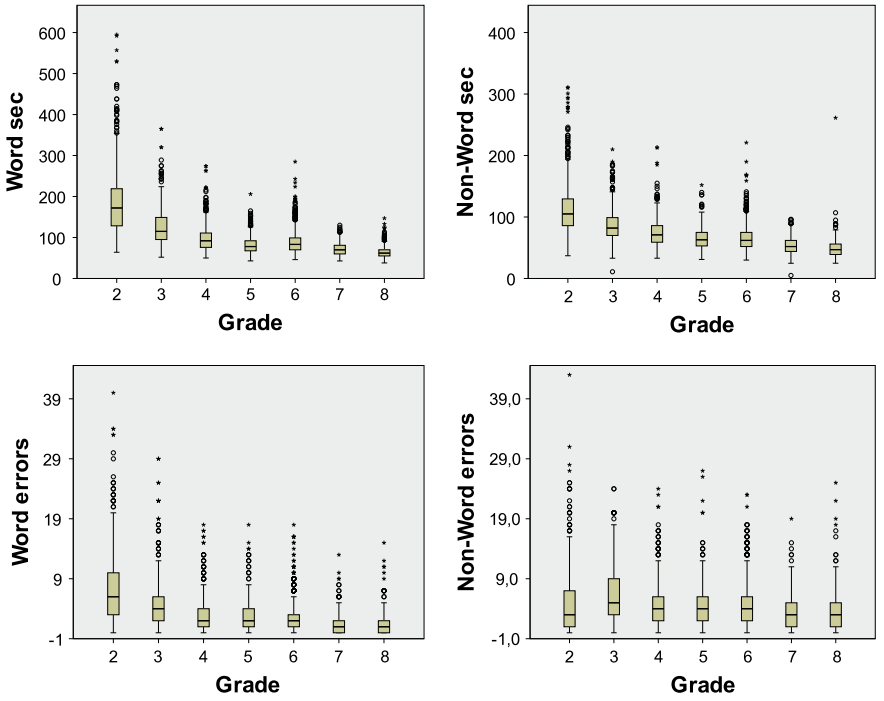


Fig. 1 Boxplots of variables X_1 (upper left), X_2 (lower left), X_3 (upper right) and X_4 (lower right)

in this parameter cause little changes in the shape of the membership function. Table 2 reports the frequency distribution of the values of the fuzzy indices. We may note that the differences in F_{11} and F_{12} and in F_{21} and F_{21} are negligible and thus the choice of the weighting system do not substantially change the values of the fuzzy indicator. On the other hand, the choice of the membership function does influence the results. The indicator is robust to weighs but not to the membership functions.

Applying the diagnostic criterion actually used in Italy for which a student is classified as impaired if he or she shows a value above normative cut-off in two or more variables, 4.8% of the students is classified as dyslexic. The fuzzy indicators give more insight into this percentage. According to F_{11} and F_{12}, about 2% is definitely dyslexic, while should be considered at high risk of impairment the 2.9%. Another approximately 4% may be viewed as being at medium risk. According to F_{21} and F_{21}, about 1% of the students are definitely dyslexic, while 1% is at high risk and approximately 1.6% at medium risk of impairment. We may also identify the prevalence of very skilled readers (64% according to F_{11} and F_{12} and 89% according to F_{21} and F_{22}) and the percentages of normal readers (given by the frequencies of the values ranging from 0.4 to 0.7).

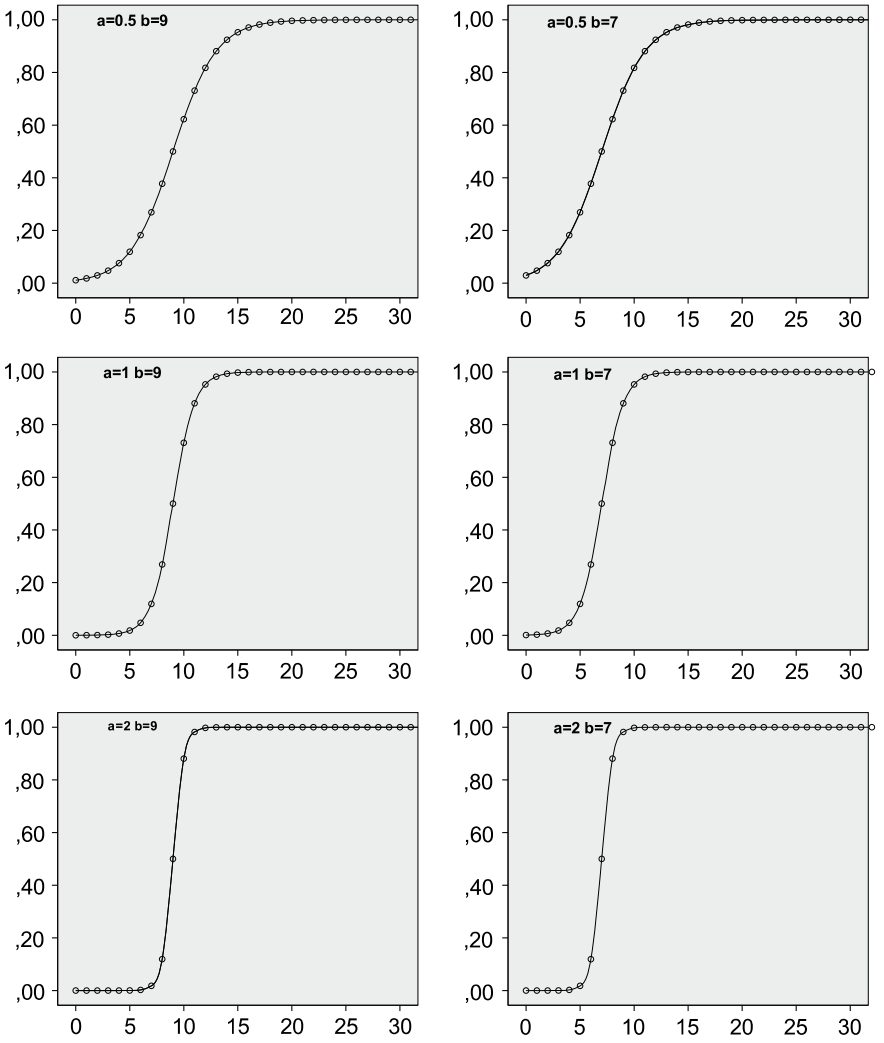


Fig. 2 Graphical representations of membership function (6) with different sets of parameters

5 Concluding Remarks

This paper presents a methodology to build fuzzy composite indicators with the aim of considering both speed and accuracy of reading in the early diagnosis of dyslexia and with the aim of going beyond the rigid unrealistic partition between dyslexic and not dyslexic students. Indeed, the limit between a bad and a pathological performance in psychometric reading tests is somehow fuzzy. The application shows that the proposed indices work well in identify the level of impairment of the students

Table 2 Frequency distributions of the values of the fuzzy composite indicators

Classes	F_{11}	F_{12}	F_{21}	F_{22}
0.0–0.4	0.648	0.643	0.895	0.894
0.4–0.6	0.208	0.206	0.041	0.047
0.6–0.7	0.059	0.062	0.028	0.024
0.7–0.8	0.038	0.039	0.018	0.016
0.8–0.9	0.029	0.029	0.009	0.009
0.9–1.0	0.019	0.021	0.009	0.010
Total	1	1	1	1

and the results are in agreement with the percentages of dyslexic students identified with the traditional diagnostic criterion but give more insights. As a matter of fact, the fuzzy approach also allows us to model the degree of membership to the set of dyslexic students and the related degree of uncertainty of a student to be impaired.

The methodology proposed can be applied to build fuzzy composite indicators for the diagnosis of different learning disabilities in academic learning like, for example, dyscalculia and dysgraphia, and can be used in any area of psychometrics. In [14] a more general approach to build fuzzy composite indicators in psychometrics is presented, with particular attention to the membership function for discrete variables. In this paper we have discussed the use of two well-known membership functions for the specific application of dyslexia and we have proposed a new weighting system. The application has shown that the composite indicator is sensitive to the choice of the membership functions but is robust to the choice of the two weighting systems carefully selected for a fuzzy index for dyslexia. Future works are needed in order to further analyse the robustness of the index to the proposed m.f. and to weights, with different data sets or with simulations studies. Future works are also needed in order to estimate the dyslexia prevalence rate in school age population in Italy, which is still an open question. The indexes proposed in the paper can be more suitably used to estimate this rate than the traditional approach, that does not consider all variables measured by means of reading tests.

References

1. Baliamoune-Lutz, M.: On the Measurement of Human Well-Being: Fuzzy Set Theory and Sen's Capability Approach, UNU-WIDER, Helsinki (2004)
2. Baliamoune-Lutz, M., McGillivray, M.: Fuzzy well-being achievement in Pacific Asia. J. Asia Pac. Econ. **11**, 168–177 (2006)
3. Cerioli, A., Zani, S.: A fuzzy approach to the measurement of poverty. In: Dagum, C., Zenga, M. (eds.) Income and Wealth Distribution, Inequality and Poverty, pp. 272–284. Springer, Berlin (1990)
4. Cheli, B., Lemmi, A.: A totally fuzzy and relative approach to the multidimensional analysis of poverty. Econ. Notes **24**(1), 115–134 (1995)

5. Chiappero Martinetti, E.: A multidimensional assessment of well-being based on sen's functioning approach. Rivista Internazionale di Scienze Sociali **108**(2), 207–239 (2000)
6. Chien, C.-J., Tsai, H.-H.: Using fuzzy numbers to evaluate perceived service quality. Fuzzy Sets Syst. **116**, 289–300 (2000)
7. Darestani, A.Y., Jahromi, A.E.: Measuring customer satisfaction using a fuzzy inference system. J. Appl. Sci. **9**(3), 469–478 (2009)
8. Klir, G.J., Folger, T.A.: Fuzzy Sets, Uncertainity and Information. Prentice-Hall Int, London (1988)
9. Kwong, C.K., Bai, H.: A fuzzy AHP approach to the determination of importance weights of customer requirements in quality function deployment. J. Intell. Manuf. **13**, 367–377 (2002)
10. Lazim, M.A., Osman, M.T.A.: A new Malaysian quality of life index based on fuzzy sets and hierarchical needs. Soc. Indic. Res. **94**(3), 499–508 (2009)
11. Lorusso, M.L., Vernice, M., Dieterich, M., Brizzolara, D., Mariani, E., De Masi, S., D'Angelo, F., Lacorte, E., Mele, A.: The process and criteria for diagnosing specific learning disorderes: indications from the consensus conference promoted by the Italian national institute of health. Ann. Ist. Super. Sanita **50**(1), 77–89 (2014)
12. Morlini, I., Stella, G., Scorza, M.: A new procedure to measure children reading speed and accuracy in Italian. Dyslexia **48**(2), 176–195 (2014)
13. Morlini, I., Stella, G., Scorza, M.: Assessing decoding ability: the role of speed and accuracy and a new composite indicator to measure decoding skill in elementary grades. J. Learn. Disabil. **20**, 54–73 (2015)
14. Morlini, I.: New fuzzy methods for psychometric data. In: Greselin, F., Mola, F., Zenga, F. (eds.) Cladag 2017 Book of Short Papers, Universitas Studiorum. Mantova, Italy (2017)
15. Sartori, G., Job, R., Tressoldi, P.E.: DDE-2 battery for the evaluation of dyslexia and disorthographya. Giunti, Florence (2007)
16. Smithson, M., Verkuilen, J.: Fuzzy Sets Theory: Applications in the Social Sciences. Sage Publications, London (2006)
17. World Health Organization: ICD-10: International Statistical Classification of Diseases and Related Health Problems 10th Revsion. World Health Organization, Geneve (2008)
18. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
19. Zani, S., Berzieri, L.: Measuring customer satisfaction using ordinal variables: an application in a survey on a contact center. Stat. Appl. Italian J. Appl. Stat. **20**(3–4), 331–351 (2008)
20. Zani, S., Milioli, M.A., Morlini, I.: Fuzzy methods and satisfaction indices. In: Kenett, R.S., Salini, S. (eds.) Modern Analysis of Customer Surveys, pp. 439–455. Wiley, Chichester (2012)
21. Zani, S., Milioli, M.A., Morlini, I.: Fuzzy composite indicators: an application for measuring customer satisfaction. In: Torelli, N., Pesarin, F., Bar-Hen, A. (eds.) Advances in Theoretical and Applied Statistics, pp. 241–251. Springer, Berlin (2013)
22. Zimmermann, H.J.: Fuzzy Sets Theory and its Applications, 2nd edn. Kluwer, Boston (1993)

Who Tweets in Italian? Demographic Characteristics of Twitter Users

Righi Alessandra, Mauro M. Gentile and Domenico M. Bianco

Abstract In this paper we try for the first time to shed light on the use of Twitter by the Italian speaking users quantifying the total audience and some relevant characteristics: in particular, gender and location. The attempt is based on publicly available APIs data referring both to profile documents and tweets. Through real-time calculation is possible to infer the gender mainly using the *name* field of the users' profile, while the geo-location is deduced using the *location* field and the geotagged tweets.

Keywords Twitter · Italian users · Social media · Big data · Machine learning

1 Introduction and Motivation

In recent years social media have become an important data source about the opinions and the sentiment of their users because they allow to capture in real-time and in a spontaneous what the users think about a certain topic. In Italy, Facebook, Twitter and more recently Instagram appear to be the most used media; Twitter has a greater accessibility and allows a more readily text analysis [10].

Twitter is a microblogging service which lets users post 280 characters long messages or tweets. Created in 2006, it has today 328 million monthly active users, according to Statista Website, and, according to Alexa Website, Twitter is the twelfth

R. Alessandra (✉) · M. M. Gentile · D. M. Bianco
Milan, Italy
e-mail: righi@istat.it

M. M. Gentile
e-mail: mauro.gentile@iese.net

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_25

most visited site in the world[1] and the 15th in Italy. Some estimates at the national level, e.g. [13], quantify in more than 6.9 million the Italian Twitter users on a population of 61 million.[2] Other estimates calculate that the users represent around 11% of the total population and 24% of 14–29 year olds [4].

Nevertheless, the use of any social media as a data source entails some challenges concerning the representativeness and the time stability of the source and the need to infer or define the socio-economic characteristics of the users. The latter, in fact, would allow to correct the strong selectivity of the social media users [4, 9]. Unfortunately, official information about who Twitter users are is not available. Twitter does not require users to self-report demographic data in the profiles information (metadata), and this scarcity of information influences researchers attempt to explore how social phenomena manifest online according to gender, age, location, occupation, and social class. Nevertheless, some demographics on the users can be inferred from publicly available information. The main purpose of this article is to take a snapshot of the Twitter Italian network and in particular to study some demographic characteristics associated to the Italian twitterers making use only of publicly available information. The presented data were collected in May 2017.

The first original contribution of the paper is to quantify the total Italian speaking Twitter audience, since, as far as we know, no other authors have previously studied this matter. Other findings refer to the quantification of some relevant demographic characteristics: we have tried to derive from the information supplied by the user both the place of residence (or presence) and the gender, considering that the latter does not figure in the users' profiles. This paper is structured as follows. Section 2 presents a brief review on related works. The Twitter data description and the technological and statistical approaches used are described in Sect. 3. Main results are discussed in Sect. 4. Section 5 discusses some open questions and concludes.

2 Related Works

Obtaining information on Twitter users attributes is challenging but it is becoming crucial because the number of researchers using Twitter information to predict financial tangibles as well as intangible assets (such as reputation and demographics for marketing scope) is rapidly increasing [2].

Thus, there have already been attempts to profile the demographic characteristics of Twitter users, especially in the U.S. Ito et al. [16] found in the extant literature different ways to estimate Twitter user attributes: according the first authors estimate the demographics of users through the contents of the texts (tweets and/or metadata) using a text-based method; in the second authors analyse the followers/followees whose tweets contain plentiful text features through a community-based method; thus exploiting the attributes of neighbors on social graph. There is also a third, hybrid, method including both tweets as text information and followers/followees

[1] According to Alexa.

[2] National demographic estimate, January 2016.

as community information [15]. These induced demographic proxies (on location, gender, language use, occupation and even social class) have in turn been used to understand differences in behaviour, such as the tendency to enable location services and geotag tweets [30, 31].

More specifically, Ikeda et al. [14] estimated user attributes (gender, age, and location) using Support Vector Machines (SVM). Their experiments tackled three attributes, and the results showed 88% accuracy with regard to gender.

Ikeda et al. [15] proposed demographic estimation algorithms for profiling Twitter users, based on their tweet and community relationships. The method is applicable to various user demographics and is suitable even for users who only tweet infrequently.

Burger et al. [3] estimated the gender of users by using a supervised learning method that employs both words and character n-grams as features, achieving 92% accuracy using the feature set of tweets, profile documents, screen name, and name.

Pennacchiotti et al. [23] estimated three users attributes (political orientation, race, and affinity for Starbucks Coffee) using the profile documents, tweeting tendency, and characteristic words in tweets as features. They update attribute-class label information by using the social graph and estimate user attributes by the Gradient Boosted Decision Trees. Chu et al. [7] used the tweeting tendency, the content of tweets, and the profile as features to distinguish human from bots by Linear Discriminant Analysis.

In the second stream is the work of Zamal et al. [33] where latent attributes of Twitter users are inferred from neighbors, but many more studies followed an hybrid method. Rao et al. [25] used n-grams or sociolinguistic features and estimated four user attributes (gender, age, location, and political orientation) by SVM. Their proposed method achieved 70-80% accuracy in estimating the attributes, and they reported that the Twitter-specific features (number of followers, number of friends, friends/followers ratio, reply ratio, number of tweets, and number of retweets) are not useful. Cheng et al. [6] working on the idea that there exists a strong correlation between a word and a particular region proposed two methods for estimating the city-level location of users: a probability model based on the correlation between a location and each word in the tweets and a grid-based neighborhood smoothing model to adjust the estimation of the user location. Moreover, Culotta et al. [8] created a distantly labeled dataset by collecting audience measurement data for 1500 websites and fit a regression model to predict the demographics using information about the followers of each website on Twitter. Huang et al. [12] built a classification model (Gradient Boosted Trees) to identify nationalities of Twitter users and trained a classifier to detect the nationality of Twitter users based on a number of features. Sakaki et al. [27] proposed a method for combining text processing and image processing to infer user's gender.

Coming to the specific characteristics investigated, some authors focused on age [21], others on gender [3, 18, 25], or on race/ethnicity [20, 23, 26], or on well-being [28] and on income [24]. The majority of these approaches rely on hand-annotated training data, require explicit self-identification by the user, or are limited to very coarse attribute values. A related lightly supervised approach includes Chang et al. [5], who infer user-level ethnicity using name/ethnicity distributions provided by the

Census; Mohammady and Culotta [20] trained an ethnicity model for Twitter using county-level supervision.

Comparisons between the demographics derived from Twitter and those coming from traditional sources (Census, surveys or others) are also attempted: Mislove et al. [19] compared the Twitter user distribution (by gender, race, and location) with the actual population distribution showing that the Twitter distribution is biased. Sloan [29] compared the UK Twitter population as estimated by recent work on demographic proxies [31] with data from the British Social Attitudes Survey 2015, then studying the relationship between demographic characteristics and the use of geoservices and geotagging [30]. Daas et al. [9] shows ways of profiling the gender of the users and the results of the combining the sample of Twitter users with their accompanying publicly available LinkedIn profiles.

Despite the increasing interest of the topic, there are still few attempts to perform large-scale demographic studies on Twitter users, due to difficulties in improving the effectiveness of methods and, consequently, the accuracy of the estimates.

A final remark should mention the problem of the possible presence of fake accounts and bots. Fake accounts are used for different purposes, such as to manipulate real users, to distort the actual statistics and to steal social network information. Bots' profiles are usually realistic and have names that seem true or that are taken from other accounts. They are identified only through their social behaviour and the contents of their tweets.[3] Gurajala et al. [11] used a combination of a pattern-matching algorithm on screen names and the analysis of update times; Lee et al. [17] relied on some behavioural statistics (posting patterns, friend information and user demographics) to train a classifier distinguishing real accounts from fake ones. A precise estimation of this phenomenon still does not exist; Varol et al. [32] estimated that 9–15% of the Twitter accounts are bots. They built different indicators and performed an exhaustive study analyzing 14 million active English speaking users (with at least 200 tweets).

3 Methodology and Data

3.1 Users' Base

Every Twitter account, besides other information (such as date of creation, description made by the user, location, name, screen name, id, whether the account has the geotagging feature enabled, number of tweets posted, followers and followings), has a field named *lang* which is "the user's self-declared user interface language" according to the official documentation.[4] It usually corresponds to the user interface language detected when the user created the Twitter account. This information cannot obviously be considered an evidence of the user's nationality. However, an user

[3] For further information see [22] and the references therein.

[4] https://dev.twitter.com/overview/api/users

interacting with Twitter using the Italian interface is most likely to be an Italian (resident, native or mother-tongue). The Italian language, in fact, is less widespread and used as a second language in respect of other languages, i.e. English or Spanish.[5]

In the following analysis we will consider only the Twitter users whose language is set to Italian. Even though in this way we will not consider all the Italians tweeting in other languages.

Thus, by means of Twitter APIs we searched for users with *lang* set to 'it', we obtained their user profile information, we analyse them and then we discard them (due to Twitter policy[6]). Consequently, the analysis process has been performed in real time and this raises several technical and methodological issues.

As a field allowing to distinguish between the accounts belonging to individuals and these referring to (public and private) enterprises or associations is absent from the user profile, before proceeding in the calculation of the demographics of the users we have to separate the accounts referring to enterprises from those of individuals. In order to search for the small enterprises and the artwork activities, we first detected all the accounts presenting in the *name* field one of more than 420 terms related to subcategories of economic activities according to NACE rev. 2 classification or other terms related to economic activities as they are currently presented on social media (terms such as restaurant, hotel and association; the complete list is available on request). Furthermore, we build a whitelist of terms and expressions in order not to remove legitimate user accounts (e.g. "fond of", "lover" and "interested of").

In addition to this, the denominations of approximately 43.5 thousand Italian enterprises (with 10 employees and over) having a website that according to a previous study present links to social media (including Twitter) [1][7] were searched in our database. After a cleansing phase (non-significant parts such as "srl", "spa" or "industry" were removed from the company names), these names were searched in the accounts' names. Enterprises are expected to write their names without mistakes into the *username* field, hence we searched for nearly exact matches between the cleaned names and the usernames. Company names and usernames were compared using the Levensthein distance between strings choosing a sufficiently high similarity threshold (indeed, 5 thresholds depending on usernames' length) in order to reduce the number of false positives. This second approach is prone to errors since many companies are named with the founder's surname; in order to reduce the misclassification errors we excluded enterprises with short names (less than 5 letters) and used a white and a blacklist to further filter the matched usernames.[8] As we were quite

[5] According to wikipedia, there are 64 million native Italian speakers in the EU and 85 million in the world when in Italy there are 61 million inhabitants. Regarding English, there are 360–400 million native speakers and 600–700 million people that speaks English as a second language.

[6] https://dev.twitter.com/overview/terms/policy.html

[7] The enterprises whose websites were scraped in the cited study, were the majority (64%) of the enterprises (with 10 employees and over) having a website, but only the half of these enterprises presented links to social media.

[8] For example, consider a company named "rossi" and the username "alexRossi". The username contains the company name but the remaining letters can be interpreted as a male proper name and hence the username is not labelled as a company.

restrictive in our procedure in order to avoid false positives, using this approach we identified only 7 thousand accounts.

At the end, we identified 209,830 accounts belonging to enterprises or associations (97% of them were identified using the first approach). The list of economic activities was eliminated from our database, and the following analysis will focus on the individuals' accounts since the main purpose of this work is to study the demographic characteristics of the users.

3.2 Gender

In order to determine the gender of the Italian speaking users we followed two approaches. In the first one, based on the gender-specific nature of names, we inferred the gender comparing their names with a gendered list of first names. In the second one, we trained a machine learning classifier to make the choice.

The procedure is applied only to the *name* field of each account and not to the *screen name* field because empirical results demonstrate that, for the gender determination, the screen names are quite useless since they contain most likely nicknames rather than the real names of the users.

This gender determination process (GDP from now on) consists in an iterative inclusion search: we looked for inclusion of each gendered name from the gendered list within the name field of each user account. Once a match is found, the gender of the matched name is associated to the user.

We collected a list of the most common Italian and European female and male first names using the statistics on the national most frequent ones (by gender) available on the National Statistics Institutes web pages of the largest European countries (Italy, Germany, UK, France, Spain, Portugal). Then we studied possible ambiguities determined by different gendered use of the same names in languages other than Italian (i.e. Andrea), and we decided to consider these names only in the gendered meaning used in Italian.

In order to reduce possible errors, the gendered list of names was divided into separated trunks: double names, short names (at most 4 characters long), long names (at least 5 characters long).

Both the users names (deriving from Twitter *name* field) and the names in the gendered lists are processed: they were transformed into lower case, the accents were removed, and punctuation and numbers were substituted with blank spaces. Ambiguous account names and those containing "&" or "e" (meaning "and" in Italian) were filtered out.

Only names from the double name list are initially processed. This expedient allowed to prevent potential classification errors due to compounded names containing both a masculine name and a feminine name (possible cases in Italian), even if our double names list was quite limited.

Then, we looked for *all* the exact matches between the tokenized account names and the gendered list of long and short names. When the match returned two or more

possibilities, we classified the account according to the gender of the first match found. This expedient helped to prevent errors due to surnames with a possible match in the gendered lists.

We also treated *names* written using camel case convention by splitting them on capital letters and then searching for an exact match for the first name which we have separated from the capitalized surname.

All the accounts not yet classified in previous matching steps are looked for through an inclusion match. The procedure verifies if a name in the gendered lists is included in the account name of which we want to infer the gender. This is the most prone to errors phase of the whole procedure since is the least restrictive. In order to limit the number of introduced errors, the inclusion is checked only for names of at least 5 characters. Nevertheless, as the Italian names are quite long (in our gendered list there are (20% of names with more than 4 letters), this threshold does not significantly reduce the risk of misclassifications.

In order to label the gender of around 20% of total counted users names not gendered by the iterative search described above, we used other profile's information. Thus, we trained a machine learning classifier on the non-empty *bio* fields. We thought, in fact, that the Italian language would help to this task since, differently from the English language, words and verbs are not gender-neutral. As train set we used all accounts with a non-empty *bio* and whose gender had been previously determined with a high level of confidence. Through a GridSearch we have fine-tuned both a Logistic Regression and a Support vector machine (SVM). Consequently, we infer the gender applying to the accounts the best classier found through the GridSearch.

3.3 Geographic Location

In order to determining the geographical distribution of accounts, we mainly relied on the profile field *location* after opportune cleansing and normalisation. Twitter provides a specific field, at user profile level, to store where the user is located. This is a non-mandatory open field that each user is requested to fill during the registration (but it can be also changed later).

We applied an inclusive match algorithm similar to what we did for the gender identification, while for those users not geolocalizated in this way, we analysed the geo-tagged tweets.

Even if the specific field to set the location is considered, the difficulties faced in treating this information relate to the issue that only 21.5% of the users filled their location field and that the concept of location in many cases it is expressed in playful terms or the same places can be referred to in a variety of ways. The location can be expressed in terms of state, region, province, city or city fraction, with a specific address, and even with longitudes and latitudes. Furthermore, typos and abbreviations are frequent.

We limited the search to the Italian speaking twitter users living in Italy, in order to get a sub-group of resident/present people in some way comparable with the population official statistics.

Similarly to what already done for gender determination, the location of users has been determined searching for the inclusion match between a list of localities[9] and the content of the *location* field of each account after some initial text cleansing (lowering cases, removals of numbers or punctuations or accents). Once matched, the location reported in the profile was enriched with the correct municipality denomination (if needed) and with the geographical area, the region, the province the matched municipality belongs to. All these information enabled us to represent users geographical distribution at different granularity.

We iterated this process in order to search for the names of the province, of the region, of the geographical area for the matching. Nevertheless, to limit typos and ambiguities due to abbreviations, we should manually amend the content of the *location* field. This has been done for the most frequent 10 thousand occurrences. After having recognised an occurrence as a location, we associated the correct province, region and geographical area to all users with that location. Although this step was quite time consuming, it allowed to increase the number of geolocalizated users by only 2%.

For users whose location could not be defined through the *location* field, we analyzed the timeline of her tweets. More in details, we built an occurrence table with the locations from which her recent tweets originated. The most frequent location was elected as the place of residence/presence of the user.

Unfortunately, this approach was not always feasible, since not every user authorized Twitter to track the geographical origin of the tweets: only 12% of Italian speaking users satisfied this condition.

4 Results

Our method allowed to identify 14,232,154 distinct users having set the field *language* to 'it' (Italian) in their accounts.

The search for the economic activities among the identified accounts allowed to find around 210 thousand accounts not referred to individuals, mainly by using the keywords related to economic activities. Consequently, these accounts were not considered in further analysis.

The total number of accounts is 23% of total resident population at January 2016 according Official statistics. This share is higher than the 11% calculated according a recent traditional survey by Censis [4], but our figure refers to the Italian speaking users that could be a wider aggregate than the resident users (normally considered in surveys).

[9] i.e. the Italian National Institute of Statistics list of municipalities, containing 7978 Italian municipalities.

Considering the degree of activity on the social network of the counted users, our findings showed that 38% of the users have never written a tweet, 13% have written only one tweet since the registration and 202 thousands do not even follow any user. It emerges a rather passive use of the social media and this finding seems confirmed by the fact that only 28% of the twitterers have tweeted in one-year period. Moreover, if we define active an account having posted just one tweet in a 4-month period, we found that only 10% of total users and 16% of those having ever tweeted can be defined active user.

In order to check our results, we followed the tweets related to the actual Italian trending topics during a month. In this way we could count how many of these twitterers have not been identified with our search method and the result is that only 0.7% of these users were not included in our previous set.

As this new subset consisted of usually active users (i.e., one out of three users tweeting about trending topics has written more than 30 tweets during the listening period) having a higher probability to be catch in our search, in order to have a correct estimate of the missing cases in the total audience we should weight the observed share keeping in account the different tweet behaviour. In this way, around 170 thousand resulted as missing cases.

4.1 Main Characteristics

We determined the gender of more than 11 million users (11,170,875), that is 78% of the total users. We consider this an important result considering the simplicity of the method. In Table 1, we report the gender distribution of accounts by phase of the process: 86% of the accounts were identified simply thanks to a single match and this means that most of users writes correctly the name and it is well separated from the surname (if existing).

In the gender distribution of the Italian Twitter users males resulted to be over-represented: 56% vs. 48% observed in the total resident population, whereas considering only the male share in the resident population aged 16–79 years (a target closer to the social media users), namely 52%, the two quotas are closer.

As no biases seems to be in the list of names, there should be an effective overrepresentation of males in the users even if we cannot completely reject the hypothesis that females are more reluctant to write explicitly their names.

Upon completion of gender determination procedure, a validation phase and accuracy estimation phase has been applied. We calculated the number of misclassifications in a randomly selected sample by GDP phase and in Table 1 the error rates are reported.

As easily expected, *double names* phase was the least prone to error, given the large length of names belonging to this category and the fact that we looked for an exact match. Thanks to these, gender inference can be done with a nearly 100% confidence in this phase. The *Exact match* seems to be a very reliable method also: only 3 errors on 1000 names were checked. *Multiple matches* has still a very good

Table 1 Gender determination accuracy by phase of the search

Phase	Classified names	Error rate (%)
Double names	70,801	0.0
One match	9,648,036	0.3
Multiple matches	714,672	2.5
Camel case names	216,722	7.0
Simple search	520,644	18.7
Total	11,170,875	4.2

scoring (2.5%), whereas the split method on capital letters yields to a 7% error rate and the *simple search* phase to a nearly 19% error rate.

Moreover, an attempt to train a Machine Learning classifier on the *bio* field to infer the gender for the 20% of the users not covered by the list of names approach reached only 75% accuracy, that is, a lower level than the overall accuracy obtained through the described GDP.

Through the methods described in Sect. 3.3, we managed to geolocate at least the region for 2,330,881 users, that is, 16.6% of the identified users and 4.1% of the resident population. In particular, 1,997,114 locations were found through the *location* field, whereas the remaining 333,767 cases were inferred from the text of the tweets. In order to validate these results, we sampled 1500 users and we found a 2.4% error rate.

Moreover, the finer the geographical granularity the lower the geolocated users' number; this is due to the fact that some users indicated only the name of the region or of the area of the region in the *location*field: consequently, the localized users by region summed up to 2,330,881, while the users by province were 2,225,641. It is worthwhile to remember that in our processing people indicating (in the field location) a place situated in a certain province have been summed up to create the total accounts of that province.

The highest share of geolocated users (28.6%) was in the North-West, followed by the Centre (24.3%) which instead presented the highest incidence of Twitter users on resident population (8.6%).

Despite the limited number of localized cases, the territorial distribution of the twitterers was similar to the resident population distribution, both at regional and at province level (at this level the dissimilarity index was 0.03).

However, the share of twitterers on total geolocated users was much higher than the share of local residents compared to the total population in the provinces of Rome (11.7% vs. 7.2%) and Milan (9.5% vs. 5.3%). This also happened in the provinces of other major cities (Florence, Naples Venice, Turin, Bologna and Genoa), but to a lesser extent (see Fig. 1). Conversely, in some medium-sized provinces of the South (Caserta and Salerno), of Lombardy (Bergamo, Monza, Varese) and of Trentino Alto Adige (Bolzano) the incidence of twitterers was far below the share of the local population on the total resident population.

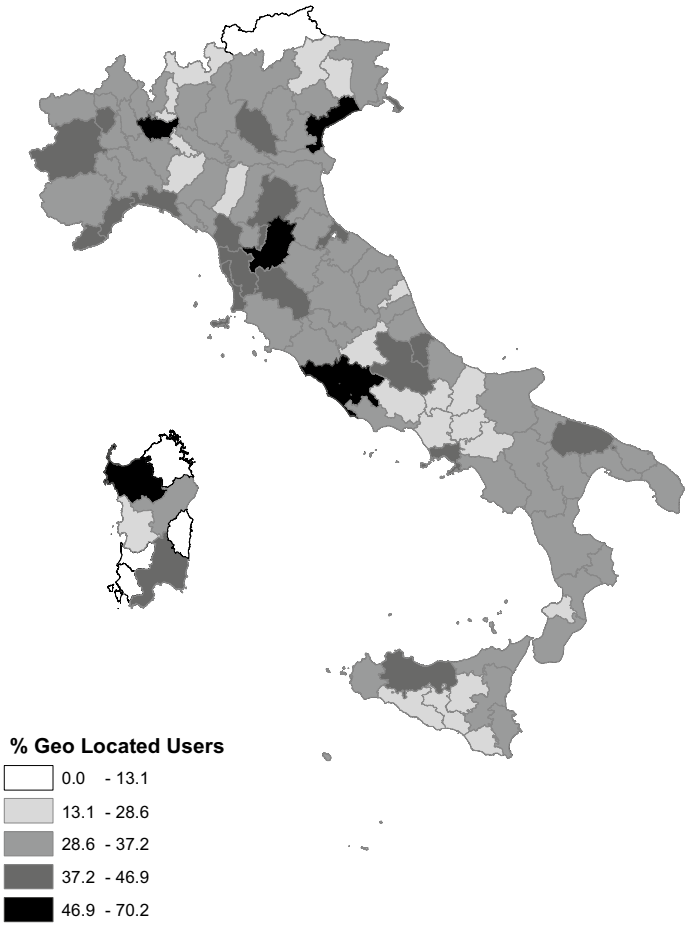


Fig. 1 Distribution of Twitter users in Italy by province

This higher penetration rate of Twitter in larger cities compared to other areas could be due to the tourist nature of these cities, where Italian-speaking tourists often join local twitterers sending messages through the platform.

Tables 2 and 3 report the most general results regarding gender and localization. We succeeded in determining both gender and residence of approximately 1.8 million users. Conversely, we could not determine any information of nearly 2.4 million users, that is the 17% of the users.

Gender was quite easier characteristic to determine than location, firstly, because users tend not to disguise their name and, secondly, because the location is an optional field. No appreciable differences between sexes in their attitude to share the location in the social media were detected. Results by geographical area (Table 4) show that the share of males, varying from 55.2% in Lazio to around 60% in Molise and

Table 2 Gendered users with determined location. Percentages are calculated with respect to the total number of users, i.e. 14,232,154

Gendered	Geolocalized					
	Y	N	Total	Y (%)	N (%)	Total (%)
Y	1,825,091	9,345,784	11,170,875	13	66	78
N	505,790	2,555,489	3,061,279	4	18	22
Total	2,330,881	11,901,273	14,232,154	17	83	100

Table 3 Geolocalized users with gender specification. Percentages are calculated with respect to the total number of males, females and users with unknown gender

Geolocalized	Gender					
	M	F	Unknown	M (%)	F (%)	Unknown (%)
Y	1,040,855	784,236	505,790	17	16	17
N	5,218,975	4,126,809	2,555,489	83	84	83
Total	6,259,830	4,911,045	3,061,279	100	100	100

Campania, is higher than the female one in every region. The Southern area of the country shows the wider differences between the share of males among Twitter users and in the resident population. In the provinces where the larger cities are located, the rate of masculinity slightly diminishes, while in the other provinces it grows up to 65% and over.

The category unknown by gender seems to be higher in the most inhabited provinces and is not uniformly territorially distributed: it varies from 13% for Olbia-Tempio to over 30% for Cuneo and Bolzano.

Furthermore, the problem of incompleteness of a list of the most popular gendered names is amplified in bilingual regions (Valle d'Aosta and Trentino Alto Adige), where in fact the shares of the category unknown by gender reached 26%.

5 Discussion and Conclusion

This study is the first attempt to verify the demographics of Italian speaking Twitter users with the goal of understanding of the differences between the twitterers and the resident population in order to better value the possible distortion of the results of analysis made using Twitter data.

We tested the feasibility of this kind of analysis by downloadable public API data and, after having eliminated from the analysis around 210 thousand accounts referred to economic activities, the results showed that is possible to obtain the total number of the over 14 million users with only a small degree of uncertainty.

Table 4 Geolocalized users and total resident population by region (absolute and percentage values)

		Geolocalized users				Total resident population		
Area	Region	Absolute	%	% Males	% Unk.	Absolute	%	% Males
Nord-West	Piedmont	163,748	7.0	57.4	23.4	4,404,246	7.3	48.4
	Aosta Valley	3793	0.2	59.3	25.7	127,329	0.2	48.8
	Lombardy	420,438	18.0	56.4	24.3	10,008,349	16.5	48.8
	Liguria	67,397	2.9	55.8	22.9	1,571,053	2.6	47.6
	Total	655,376	28.1	56.6	24.0	16,110,977	26.6	48.6
Nord-East	Trentino-Alto Adige	24,010	1.0	59.6	26.7	1,059,114	1.7	49.1
	Veneto	178,216	7.6	57.0	22.7	4,915,123	8.1	48.8
	Friuli-Venezia Giulia	36,507	1.6	57.1	23.5	1,221,218	2.0	48.4
	Emilia-Romagna	149,175	6.4	55.8	22.1	4,448,146	7.3	48.5
	Total	387,908	16.6	56.7	22.8	11,643,601	19.2	48.6
Centre	Tuscany	164,392	7.1	55.5	21.6	3,744,398	6.2	48.1
	Marches	51,938	2.2	58.9	21.4	1,543,752	2.5	48.4
	Umbria	32,259	1.4	56.6	21.8	891,181	1.5	48.0
	Lazio	307,428	13.2	55.2	23.0	5,888,472	9.7	48.2
	Total	556,017	23.9	55.8	22.4	12,067,803	19.9	48.2
South	Abruzzo	47154	2.0	58.0	19.8	1,326,513	2.2	48.7
	Molise	7977	0.3	60.3	18.3	312,027	0.5	49.1
	Campania	207,309	8.9	59.8	18.1	5,850,850	9.6	48.7
	Apulia	145,840	6.3	58.5	18.3	4,077,166	6.7	48.5
	Basilicata	18,210	0.8	63.5	18.5	573,694	0.9	49.1
	Calabria	60,693	2.6	58.9	18.3	1,970,521	3.2	48.9
	Total	487,183	20.9	59.3	18.4	14,110,771	23.3	48.7
Islands	Sicily	178,328	7.7	57.6	18.2	5,074,261	8.4	48.6
	Sardinia	66,069	2.8	54.9	21.1	1,658,138	2.7	49.0
	Total	244,397	10.5	56.9	19.0	6,732,399	11.1	48.7
Total		2,330,881	100.0	57.0	21.7	60,665,551	100.0	48.6

Unfortunately, the classification by gender was more difficult to obtain, as we were able to determine the gender of 11 million users. Whereas, referring to the geographic location of users, we identified only 17% of users, although the imputation error for this subgroup is very low. This is due to the fact that the *location* field is an open field which can be filled (or not) with abbreviations or fancy names, containing many mistakes, too difficult to correct. The limited number of localized cases is a

particularly negative aspect because the real value of Twitter data is related to their granularity allowing researcher and analysts to develop data analysis at local level.

The main findings of this study are that males are overrepresented among the Twitter users and in larger cities areas the share of males is shrinking whereas the number of people whose gender is unknown is increasing. The territorial distribution seems to be quite similar to that of the total population, with only few over-representations for provinces as Rome and Milan.

Following our approach, based on public API data, it was difficult to localize a significant share of users. The determination may improve by purchasing data from the Provider, as in this case the share of accounts with available indications could be broader (at least according to similar studies on other national contexts).

Thus, our results are not completely encouraging[10] but limitations could be at least partially overcome supplementing our approach with a text-based, image-based or community-based approach.

The text analysis of the terms used by the users or the processing of the user profiles' images (even if the latest requires a lot of commitment in terms of computational time) could complement the results of gender classification made only using a name list.

Furthermore, we inferred information for every single account without exploiting the Twitter social network structure, whereas a community-based approach would allow to infer individual characteristics of a user from her followers/followings (since there are accounts that are most likely to be followed by males instead of females). Nevertheless, as the Italian speaking users have a small number of followers/followings, the adoption of this technique could be applied only to a subset of the Twitter users.

As a final remark, we should point out that our findings could be inaccurate due to the presence of fake accounts and bots used for distort the actual Twitter statistics and manipulate the public opinion. Their presence can partly affect both the total number of real users and the share of gendered and geolocalized users. As shown in Sect. 2 by the cited literature on this topic, fake accounts can be identified only studying their social behaviour and the text of the tweets, namely, through techniques we did not use in this study. Consequently, further specific studies are needed to evaluate the real impact of bots on the Italian speaking Twitter network.

References

1. Barcaroli, G., Bianchi, G., Nurra, A.: Internet as a data source: Ict use of enterprises: web ordering, job advertising and presence on social media. In: Big Data Committee Annual Report 2017, ISTAT, CIKM '10. https://www.istat.it/it/files//2018/09/Big-data-committee.pdf (2018)

[10] We tried also to determine the users profession using the *bio* field, through a list of roughly 1000 professions. Results were absolutely not satisfactory maybe because the *bio* field is an open field that each user interprets in her own way.

2. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. http://arxiv.org/abs/1010.3003 (2010)
3. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pp. 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics. ISBN 978-1-937284-11-4. http://dl.acm.org/citation.cfm?id=2145432.2145568 (2011)
4. Censis. 13° rapporto censis-ucsi sulla comunicazione i media tra élite e popolo. http://www.censis.it/17?shadow_pubblicazione=120570 (2016)
5. Chang, J., Rosenn, I., Backstrom, L., Marlow,C.: Epluribus: Ethnicity on social networks. In: ICWSM (2010)
6. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, New York, NY, USA, pp. 759–768. ACM. ISBN 978-1-4503-0099-5. https://doi.org/10.1145/1871437.1871535. http://doi.acm.org/10.1145/1871437.1871535 (2010)
7. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? In: Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC '10, New York, NY, USA, pp. 21–30. ACM. ISBN 978-1-4503-0133-6. https://doi.org/10.1145/1920261.1920265. http://doi.acm.org/10.1145/1920261.1920265 (2010)
8. Culotta, A., Ravi, N.K., Cutler, J: Predicting the demographics of twitter users from website traffic data. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, pp. 72–78. AAAI Press. ISBN 0-262-51129-0. http://dl.acm.org/citation.cfm?id=2887007.2887018 (2015)
9. Daas, P.J., Burger, J., Le, Q., ten Bosch, O., Puts, M.J.: Profiling of Twitter Users: A Big Data Selectivity Study (2016)
10. Della Ratta, F., Pontecorvo, M.E., Vaccari, C., Virgillito, A.: Big data and textual analysis: a corpus selection from twitter. Rome between the fear of terrorism and the jubilee. https://www.researchgate.net/publication/303843023_Big_data_and_textual_analysis_a_corpus_selection_from_Twitter_Rome_between_the_fear_of_terrorism_and_the_Jubilee (2016)
11. Gurajala, S., White, J.S., Hudson, B., Matthews, J.N.: Fake twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. In: Proceedings of the 2015 International Conference on Social Media & Society, SMSociety '15, New York, NY, USA, pp. 9:1–9:7. ACM. ISBN 978-1-4503-3923-0. https://doi.org/10.1145/2789187.2789206. http://doi.acm.org/10.1145/2789187.2789206 (2015)
12. Huang, W., Weber, I., Vieweg, S.: Inferring nationalities of twitter users and studying international linking. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, HT '14, New York, NY, USA, pp. 237–242. ACM. ISBN 978-1-4503-2954-5. https://doi.org/10.1145/2631775.2631825. http://doi.acm.org/10.1145/2631775.2631825 (2014)
13. ICTGlobus. Social media in italia: analisi dei flussi di utilizzo del 2016. https://www.ictglobus.com/social-media-in-italia-analisi-dei-flussi-di-utilizzo-del-2016/ (2017)
14. Ikeda, K., Hattori, G., Matsumoto, K., Ono, C., Higashino, T.: Demographic estimation of twitter users for marketing analysis. IPSJ Trans. Consum. Devices Syst. **2**(1), 82–93 (2012)
15. Ikeda, K., Hattori, G., Ono, C., Asoh, H., Higashino, T.: Twitter user profiling based on text and community mining for market analysis. Knowl.-Based Syst. **51**(1), 35–47. ISSN 0950-7051. https://doi.org/10.1016/j.knosys.2013.06.020. https://doi.org/10.1016/j.knosys.2013.06.020 (2013)
16. Ito, J., Nishida, K., Hoshide, T., Toda, H., Uchiyama, T.: Demographic and psychographic estimation of twitter users using social structures, pp. 27–46. Springer International Publishing, Cham (2014). ISBN 978-3-319-13590-8. https://doi.org/10.1007/978-3-319-13590-8_2. https://doi.org/10.1007/978-3-319-13590-8_2
17. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pp. 435–442, New York, NY, USA. ACM (2010). ISBN 978-1-4503-0153-4. https://doi.org/10.1145/1835449.1835522. http://doi.acm.org/10.1145/1835449.1835522

18. Liu, W., Ruths, D.: What's in a name? using first names as features for gender inference in twitter. In: AAAI spring symposium: Analyzing microtext, vol. 13, p. 01 (2013)
19. Mislove, A., Jørgensen, S., Ahn, Y.-Y., Onnela, J.-P., Rosenquist, J.: Understanding the demographics of twitter users, pp. 554–557. AAAI Press (2011). ISBN 978-1-57735-505-2
20. Mohammady, E., Culotta, A.: Using county demographics to infer attributes of twitter users. ACL **2014**, 7 (2014)
21. Nguyen, D., Smith, N.A., Rosé, C.P.: Author age prediction from text using linear regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH '11, pp. 115–123, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284046. http://dl.acm.org/citation.cfm?id=2107636.2107651
22. Paquet-Clouston, M., Bilodeau, O., Décary-Hétu, D.: Can we trust social media data?: Social network manipulation by an iot botnet. In: Proceedings of the 8th International Conference on Social Media & Society, #SMSociety17, pp. 15:1–15:9, New York, NY, USA. ACM. ISBN 978-1-4503-4847-8. https://doi.org/10.1145/3097286.3097301. http://doi.acm.org/10.1145/3097286.3097301 (2017)
23. Pennacchiotti, M., Popescu, A.-M.: A machine learning approach to twitter user classification. In: ICWSM (2011)
24. Preotiuc-Pietro, D., Volkova, S., Lampos,V., Bachrach, Y., Aletras, N.: Studying user income through language, behaviour and affect in social media. PLOS One 10(9), 1–17 (2015). https://doi.org/10.1371/journal.pone.0138717. https://doi.org/10.1371/journal.pone.0138717
25. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents, SMUC '10, pp. 37–44, New York, NY, USA. ACM. ISBN 978-1-4503-0386-6. https://doi.org/10.1145/1871985.1871993. http://doi.acm.org/10.1145/1871985.1871993 (2010)
26. Rao, D., Paul, M.J., Fink, C., Yarowsky, D., Oates, T., Coppersmith, G.: Hierarchical bayesian models for latent attribute detection in social media. In: Adamic, L.A., Baeza-Yates, R.A., Counts, S. (eds.) ICWSM. The AAAI Press. http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#RaoPFYOC11 (2011)
27. Sakaki, S., Miura, Y., Ma, X., Hattori, K., Ohkuma, T.: Twitter user gender inference using combined analysis of text and image processing. V&L Net **2014**, 54 (2014)
28. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Lucas, R.E., Agrawal, M., Park, G.J., Lakshmikanth, S.K., Jha, S., Seligman, M.E. et al.: Characterizing geographic variation in well-being using tweets. In: ICWSM (2013)
29. Sloan, L.: Who tweets in the united kingdom? Profiling the twitter population using the british social attitudes survey 2015. Social Media + Society, **3**(1), 2056305117698981 (2017). https://doi.org/10.1177/2056305117698981. http://dx.doi.org/10.1177/2056305117698981
30. Sloan, L., Morgan, J.: Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. PLOS One **10**(11), 1–15 (2015). https://doi.org/10.1371/journal.pone.0142209. https://doi.org/10.1371/journal.pone.0142209
31. Sloan, L., Morgan, J., Burnap, P., Williams, M.: Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PLOS One **10**(3), 1–20 (2015). https://doi.org/10.1371/journal.pone.0115545. https://doi.org/10.1371/journal.pone.0115545
32. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. CoRR abs/1703.03107, http://arxiv.org/abs/1703.03107 (2017)
33. Zamal, F.A., Liu, W., Ruths, D.: Homophily and latent attribute inference: inferring latent attributes of twitter users from neighbors. In: Breslin, J.G., Ellison, N.B., Shanahan, J.G., Tufekci, Z. (eds.) ICWSM. The AAAI Press. http://dblp.uni-trier.de/db/conf/icwsm/icwsm2012.html#ZamalLR12 (2012)

Applying Data Science in Economics and Labour Market

An Approach to Developing a Scoring System for Peer-to-Peer (p2p) Lending Platform

Alexander Agapitov, Irina Lakman, Zoya Maksimenko and Natalia Efimenko

Abstract The paper reviews the possibilities of using survival analysis tools to configure scoring systems for p2p lending platform. Along with the Cox model, the models of log-logistic regression, accelerated failure time (AFT) model and Weibull regression were considered in this study. To test the stability of the factor influence the models were built when discretizing the observation period (12 months, 24 months and 36 months). The sample consisted of 887,379 observations for the period of 2007–2016. The study examined loans issued for the period of 36 months. Proportional hazard models were also analyzed taking into account the grouping feature of borrowers reditworthiness. The best model describing the state duration before the default was chosen. As a result of the analysis the factors affecting the probability of the borrower default during the considered period of time were revealed. It was determined that the greatest influence on the default risk was exerted by the purpose of loan and the interest rate regardless of the considered dynamics. The borrower's income also had a significant impact on the default risk.

Keywords Crowdinvesting · Peer-to-peer (p2p) lending · Credit scoring · Survival analysis · Cox proportional hazard model

A. Agapitov (✉) · I. Lakman · Z. Maksimenko · N. Efimenko
Ufa State Aviation Technical University, Ufa, Russia
e-mail: aleks6321@yandex.ru

I. Lakman
e-mail: lackmania@mail.ru

Z. Maksimenko
e-mail: zubazzz@mail.ru

N. Efimenko
e-mail: efimenko@ufanet.ru

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_26

1 Introduction

Over the last 2–3 years the number of services that allows using peer-to-peer (p2p) lending system has grown in Russia. According to the report of the Central Bank of Russia by means of crowdinvesting platforms, 52,2 million rubles [7] or slightly less than 1 million U.S. dollars have been for 2015 and it is significantly less than the volumes of p2p lending in such countries as, for example, Great Britain and the USA. Moreover, the dynamics of similar service development has explosive nature. According to publicly available sources [23] about 2 billion dollars in the form of p2p loans have been issued for 2012 in the USA, and more than 25 billion in 2015. One of the reasons for poor development of the institution of peer-to-peer lending in Russia, besides the lack of legal clarity [12], is the impossibility of providing creditors with high quality services of the potential borrower scoring assessment that allows establishing a floating interest rate depending on the probability of the certain borrower default. The majority of successful western platforms of p2p-lending, in particular the world's largest Lending Club [13] platform, have the built-in recommendatory online services of the loan interest rate based on the estimated level of the potential borrower reliability. However these services have some shortcomings:

- the lack of possibility to vary the borrower's rate during the whole term of the loan, i.e. if the borrower passes the time fence of a possible default, then a lower interest rate may be established for him/her;
- the interest rate is not personalized for each specific borrower, but it is segmented by certain groups of borrowers.

This is largely due to the fact that currently credit scoring systems are created only for a simple classification of borrowers according to the principle of classifying it as "bad" or "good." Not the most popular, but used method of teaching the scoring model is a method based on a survival analysis that allows not only classifying borrowers, but also assessing the factors that affect the duration of a state before the borrower defaults, i.e. determining the possible time of default.

The survival analysis methodology to credit scoring was first introduced by Narain [16]. Narain showed the advantages of using survival analysis tools over standard approaches when training the scoring models. A period of 24 months was used as the state duration for borrower default. Subsequently, there were numerous studies showing the comparative advantages of survival analysis over standard methods of credit scoring.

Stepanova and Thomas [20] identified the advantages of survival analysis compared to logistic regression, and also showed that one of the significant factors influencing the state duration before the borrower default was the purpose of the personal loan. These authors proposed using ROC-analysis as a measure of assessing the quality of the model.

Sarlija et al. [19] also used ROC-analysis to assess the quality of scoring models identifying the significant factors built with the use of logistic regression, survival analysis, and neural networks. The authors showed the superiority of neural networks

in solving this problem; however the study considered the standard Cox model with no set of distribution, methods of coefficient assessment ratios were not varied, the selection between models estimated by various methods was not carried out.

Bellotti and Crook [2–4] suggested to add macroeconomic influence factors to the scoring model trained with the use of the survival model. Okumu et al. [17] paid particular attention to the gender perspective when estimating the Cox model.

The comparison of results from the application of the Cox proportional hazard model and the AFT model in credit scoring can be seen in the publication of Pazdera et al. [18].

The comprehensive work of Man [14] is devoted to the practical application of survival analysis tools and logistic regression with a change in the period of the state duration before the default. He pointed out that both models showed similar results, but survival models required less data cleaning. The same results from the comparison of the two models were adduced by Marimo [15]. He notes that survival models provide more important information the survival function, and not just the probability of the default.

Watkins et al. [22] as well as Man [14] recorded various periods of the state duration before the default in assessing the Cox model in order to show the stability of influence factors.

Dirick et al. [8] in their work varied not only a type of risk function and the periods of the state duration before the default at the same time, but also the sample size.

Authors of previous research considered questions of the scoring system creation earlier using Hekman's models [6]. Present research attempts to build scoring models using tools of the survival analysis.

2 Methodology

In order to assess the state duration before the default the following components required to build survival models: object (borrower); event (default occurrence); and duration variable (period from the loan granting up to the default occurrence in months) were used in the present study. The main objective of the analysis is to reveal the factors affecting the probability of occurrence or nonoccurrence of the borrower default during the considered period of time $P(T \leq t)$.

At the first stage of the model training it is necessary to build graphs of the survival function estimated by the Kaplan–Meier method when grouping research objects according to the attributes of any alternatives [11]. This approach allows making a conclusion about the difference in survival functions for the different categories of objects. Therefore, the approach forms an inference about what factors can potentially be predictors for the probability of the borrower default. Furthermore, analysis of the survival function graphs allows making an assumption about the distribution functions.

At the second stage for the accurate assessment of differences in the survival functions when grouped according to the attributes of any alternatives the log-rank

criterion of Mantel–Haenszel and the criterion of Gehan–Wilcoxon are used [10]. In all tests the null hypothesis is the assumption that there are no differences in survival functions for various attributes of any alternatives.

At the third stage the nonparametric Cox proportional hazards model [6] is constructed where the factors defined at the previous two stages are regressors:

$$\lambda_i(t|x_i) = \lambda_0(t) * \exp^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k} \quad (1)$$

Here $\lambda_i(t|x_i)$ is the default risk in the period t under the condition of different values of the influence factors x_i, $\lambda_0(t)$ is the basic (average) ruin risk in the period t, the exponent plays the role of the multiplicative risk effect.

The assessments of coefficient β_j of the Cox model (1) are determined using the partial likelihood method according to Efron or Breslow techniques [5, 9]. The choice of the best evaluation methodology is based on a minimum of information criteria of Akaike and Schwartz.

Along with the Cox model, the models of log-logistic regression, accelerated failure time (AFT) model, and Weibull regression will also be considered in this study. To test the stability of the factor influence the models will be built when discretizing the observation period (12, 24 and 36 months).

Thus, in total, 15 models will be built and the choice of the best one will be based on the analysis of ROC-curves as well as on the calculation of the derived AUC indicator (area under the curve). As a result an adequate model for building scoring systems with the highest AUC value will be recommended.

In order to build all the models the software R was used [21].

3 Data

The study uses loan data of the Californian company Lending Club [13], focusing on peer-to-peer lending. The sample consisted of 887,379 observations for the period of 2007–2016. The study examines loans issued for a period of 36 months. After the required credit period remained and all missing data deleted, there were 602,871 observations.

As stated earlier the object (observation) is a borrower. The risk of an event (default) occurrence in a certain period is predicted for the borrower. This object was under observation and therefore was at risk: at any period of time an event may occur when he is eliminated from the risk group. The following periods of observation were considered:

- up to 12 months the number of recorded defaults is 20,721;
- up to 24 months the number of recorded defaults is 34,342;
- up to 36 months the number of recorded defaults is 38,277.

In survival analysis some of the observations are always censored. In the present study, borrowers who continued to service the loan during the observation period, as

well as borrowers who repaid the loan ahead of time were considered as censored data. Defaults occurred during the observation period were considered as full observations. Predictors of the default in the study were the interest rate on the loan, the length of employment, the annual income and the region of the borrower's residence, the housing ownership, the credit history, the size of the loan and its purpose and financial reliability of the borrower calculated by Lending Club on the scale from A to D where A is the best possible grade and D the worst.

The period from the moment when the object (the customer) borrowed money up to the date when the object defaulted was considered as a duration variable. The paper examines borrowers with a loan repayment period of 36 months.

4 Experimental Results

The analysis of the survival functions graphs obtained with the help of the Kaplan–Meier estimates showed that the majority of the default predictor had significant differences in survival functions between the alternatives. For example, Fig. 1 shows a graph of the Kaplan–Meier function for the Funded Amount predictor which clearly demonstrates the difference in survival functions for the various alternative attributes.

Table 1 shows the results of Mantel–Haenszel log-rank test and Gehan–Wilcoxon test for each borrower determinant. The test results showed a statistically significant difference in the groups for each variable. It was concluded that in order to build models for determining the probability of the default occurrence in the considered period of time all the borrower predictors should be used.

Coefficient assessments of the Cox proportional hazards models with discrete observation time were obtained by the partial likelihood method using Efron and Breslow techniques. The least values of Akaike (AIC) and Schwartz (BIC) information criteria (Table 2) are found for the Cox models assessed according to Efron method, hence such assessments are more reliable.

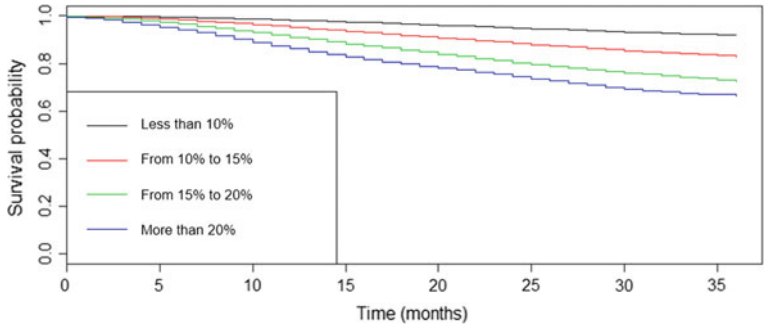


Fig. 1 A graph of the Kaplan–Meier function for the Funded Amount predictor

Table 1 Survival analysis: tests

Variable	Log-rank test		Gehan's generalized Wilcoxon	
	χ^2 statistic	Degrees of freedom	χ^2 statistic	Degrees of freedom
Home ownership	968***	2	987***	2
Earliest credit line	742***	2	749***	2
Interest rate	10887***	3	11036***	3
Annual income	209***	6	2121***	6
Funded amount	183***	4	196***	4
Employment length	499***	5	512***	5
Region of the US	130***	8	131***	8
Credit purpose	1439***	8	1395***	8

*** indicate the χ^2 statistic is significant at the 1% level

Table 2 Survival analysis: values of AIC and BIC

Method of assessment	Period of observation	AIC	BIC
Efron	12 months	1140	1144
	24 months	1178	1181
	36 months	1196	1199
Breslow	12 months	1148	1149
	24 months	1180	1181
	36 months	1199	1201

Along with nonparametric models for estimating the state duration before the borrower default, parametric models such as accelerated failure time (AFT) models were also evaluated. The models were built on the assumption of the correlation of the state duration function to the log-logistic distribution and the Weibull distribution. After carrying out the ROC-analysis and determining the area under the ROC curve obtained for each of the built models it was revealed that there was the highest AUC value for the Cox proportional hazards models.

Table 3 presents the results of the assessment of the Cox proportional hazard model (exponents of the model coefficient assessments) performed by Efron partial likelihood methods with the time discretization (12, 24, 36 months).

The proportional hazard model was also evaluated taking into account the bank's customer groups of "reliable" and "unreliable" clients received by Lending Club. Table 4 shows the results of calculations by multipliers compared to the basis risk calculated using the Cox model for "reliable" and "unreliable" clients respectively [1].

As a result of the analysis the following conclusions can be made:

Table 3 Survival analysis: Cox models

Variable	Level	12 months	24 months	36 months
Home ownership	Mortgage	0.861***	0.885***	0.895***
	Own	0.941***	0.932***	0.936***
Earliest credit line	After 2000	1.113***	1.099***	1.094***
Interest rate	>10%	2.864***	2.302***	2.250***
	From 15% to 20%	4.440***	4.016***	3.873***
	>20%	6.567***	5.509***	5.324***
Annual income	From 15 to 30	0.912	0.940	0.948
	From 30 to 50	0.837*	0.856**	0.869*
	From 50 to 75	0.741***	0.742***	0.751***
	From 75 to 100	0.652***	0.633***	0.642***
	From 100 to 150	0.586***	0.574***	0.578***
	>150	0.559***	0.538***	0.545***
Funded amount	From 5000 to 10,000	1.168***	1.165***	1.174***
	From 10,000 to 15,000	1.176***	1.222***	1.241***
	From 15,000 to 25,000	1.251***	1.305***	1.321***
	>25,000	1.409***	1.430***	1.446***
Employment length	Less than 1 year	0.937	0.902***	0.888***
	1 year	0.841***	0.821***	0.810***
	From 2 to 5 years	0.785***	0.803***	0.800***
	From 6 to 9 years	0.786***	0.816***	0.817***
	10 and more	0.738***	0.768***	0.770***
Region of the US	Mountain	1.009	1.022	1.022
	West North Central	0.966	0.964	0.964
	East North Central	0.911***	0.931***	0.930***
	West South Central	0.929**	0.945*	0.941*
	East South Central	1.099*	1.107***	1.101***
	South Atlantic	0.968	1.008	1.009
	Mid-Atlantic	0.999	0.999	0.994
	New England	0.884***	0.904***	0.899***
Purpose	Credit card	0.727***	0.807***	0.809***
	Major purchase	0.948	0.893*	0.868*
	Other	0.912*	0.978	0.976
	Car	0.857*	0.841*	0.843*
	Medical	1.233***	1.209***	1.181***
	Small business	1.473***	1.496***	1.462***
	House	1.005	1.108	1.106
	Home improvement	0.949	0.984	0.985

***, ** and * indicate the parameter estimates are significant at the 1%, 5% and 10% levels, respectively

Table 4 Survival analysis: Cox models

Variable	Level	Good	Bad
Home ownership	Mortgage	0.899***	0.905***
	Own	0.922***	0.946
Earliest credit line	From 1990 to 2000	1.115***	1.041***
	After 2000	1.150***	1.145***
Interest rate	>10%	1.864***	—
	From 15% to 20%	—	1.416***
	>20%	—	2.009***
Annual income	From 15 to 30	—	0.998***
	From 30 to 50	0.868**	0.934***
	From 50 to 75	0.706***	0.845***
	From 75 to 100	0.597***	0.741***
	From 100 to 150	0.548***	0.665**
	>150	0.514***	0.636***
Funded Amount	From 5000 to 10,000	1.049***	1.228***
	From 10,000 to 15,000	1.084***	1.332***
	From 15,000 to 25,000	1.160***	1.431***
	>25,000	1.204***	1.487***
Employment length	Less than 1 year	0.766*	0.943**
	1 year	0.690***	0.868**
	From 2 to 5 years	0.691***	0.852***
	From 6 to 9 years	0.717***	0.864***
	10 and more	0.699	0.798
Region of the US	Mountain	1.063	1.000
	West North Central	0.992	0.946
	East North Central	0.940***	0.924***
	West South Central	0.966***	0.926***
	East South Central	1.128***	1.076***
	South Atlantic	1.050	0.983
	Mid-Atlantic	1.062*	0.957
	New England	0.946***	0.870***
Purpose	Credit card	0.867***	0.843***
	Major purchase	0.888	0.874
	Other	1.216	0.955
	Car	0.807***	0.899***
	Medical	1.369*	1.136
	Small business	1.980***	1.322***
	House	1.044**	1.134*
	Home improvement	1.043	0.978

***, ** and * indicate the parameter estimates are significant at the 1%, 5% and 10% levels, respectively

1. The availability of own housing, or housing purchased in a mortgage helps reduce the risk of the borrower default during the considered periods (12, 24 and 36 months) approximately 0.87 and 0.94 times respectively. It should be noted that in contrast to the standard assumption, the availability of a borrowers mortgage is a factor that significantly reduces the risk of reaching a loan delinquency in the considered periods. For "unreliable" borrowers, the following risks were identified. Borrowers who live in owner-occupied dwelling bear the risk by 10% greater compared to borrowers living in rented accommodation.
2. The most significant impact on the default risk by a certain period in comparison with other factors is provided by the interest rate factor. So if the annual loan rate is from 10% to 15%, the default risk at any period of the borrower observation increases by an average of 2.4 times. If the interest rate is between 15% and 20%, the risk of the borrower default in the first year of the crediting period increases 4.4 times compared to the base risk. In the case when a loan was issued to a borrower at a rate of more than 20%, it is expected that the loan will be defaulted in the first year of the crediting period 6.5 times more often, in the second and third years 5.4 times more often than the average for all borrowers.
3. The increase in the annual income of an equal partnership bank customer significantly reduces the default risk. For example, for customers with incomes of more than $150,000 U.S., the risk is reduced almost 2 times compared to the basis risk. Moreover, such a customer determinant equally reduces the default risk in the first, second and third years of the observation from the date of the loan receipt.
4. The risk multipliers of the loan default increase with a higher loan amount Moreover, these indicators practically do not depend on the duration periods of the borrower observation.
5. There is a decrease in the default risk during the observation period to 36 months for borrowers living in the areas of East North Central, East South Central and New England, and conversely, an increase in the probability of default is 1.1 times compared with the basis risk for borrowers living in East South Central. "Unreliable" borrowers living on the East Coast of the USA, on average, carry lower risk of debt compared to the inhabitants of the West Coast and mountain states.
6. As it was shown in the work of Stepanova and Thomas [20] the purpose of credit has a significant impact on the probability of default by a certain date. The debt risk of a borrower who has a loan for a small business is 1.5 times higher for the whole observation period than for customers with a basic risk. The risk is also increased 1.2 times for borrowers with a loan for medical services. It is an interesting fact that the purpose of a loan related to the purchase or repair of a house does not change the probability of the basis default risk at any of the observation periods (12, 24 and 36 months). The purpose of a loan related to the purchase of a car or the use of a credit card on the contrary reduces the default risk at any period of the considered crediting time.
7. "Reliable" clients have a lower risk of debt with high socio-economic indicators compared to the conventional model. At the same time, borrowers who took credit for small business have much higher risks.

Thus, the developed model describes well not only the risk of the borrower default, but also its dynamics. The obtained estimates can form the basis for the development of comprehensive recommendations on the establishment of a floating interest rate for each particular borrower.

5 Conclusion

The study using the example of Lending Club data showed that survival analysis tools, in particular the Cox proportional hazard model with a partial likelihood function estimated by the Efron approximation, can be used to configure scoring systems for p2p lending.

The scientific novelty of the approach proposed by authors is the possibility to determine and quote a personalized floating rate on the credit depending on the risks and time elapsed from the date of the credit granting. In contrast, the standard methods involve differentiation of rates only by customers and the total period of lending.

The practical significance of the research results is the implementation of this scoring model for assessing the borrowing capacity of a customer for the peer-to-peer lending platform will increase the attractiveness of appropriate technology for customers.

The proposed approach is planned to be implemented in one of the Russian banks specializing in working with small businesses and developing a platform for p2p-lending.

References

1. Agapitov, A., Lackman, I., Maksimenko, Z.: Determination of basis risk multiplier of a borrower default using survival analysis. In: Proceedings of the Conference of the Italian Statistical Society (Statistics and Data Science: new challenges, new generations SIS 2017), pp. 1–6. Firenze University Press, Firenze (2017)
2. Bellotti, T., Crook, J.: Credit scoring with macroeconomic variables using survival analysis. J. Oper. Res. Soc. **60**(12), 1699–1707 (2009)
3. Bellotti, T., Crook, J.: Forecasting and stress testing credit card default using dynamic models. Int. J. Forecast. **29**(4), 563–574 (2013)
4. Bellotti, T., Crook, J.: Retail credit stress testing using a discrete hazard model with macroeconomic factors. J. Oper. Res. Soc. **65**(3), 340–350 (2014)
5. Breslow, N.E.: Covariance analysis of censored survival data. Biometrics **30**, 89–99 (1974)
6. Cox, D.: Partial Likelihood. Biometrika **62**, 269–276 (1975)
7. Crowdfunding. Crowdinvesting in Russia. The official website. 22 April 2016. http://www.cbr.ru/press/event/?id=287. Accessed 20 June 2017
8. Dirick, L., Claeskens, G., Baesens, B.: Time to default in credit scoring using survival analysis: a benchmark study. J. Oper. Res. Soc. **68**, 652–665 (2017)
9. Efron, B.: The efficiency of cox's likelihood function for censored data. J. Am. Stat. Assoc. **72**(359), 557–565 (1977)

10. Gehan, E.A.: A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika **52**(1–2), 203–223 (1965)
11. Kaplan, E., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**(282), 457–481 (1958)
12. Kuznetsov, V.A.: Crowdfanding: actual issues of regulation. Money and credit, 1, pp. 65–73 (2017)
13. Lending Club Corporation. Official site. https://www.lendingclub.com. Accessed 30 June 2017
14. Man, R.: Survival analysis in credit scoring: a framework for PD estimation. University of Twente, Netherlands. http://essay.utwente.nl/65049/1/ThesisRamonMan.pdf (2014). Accessed 03 July 2017
15. Marimo, M.: Survival analysis of bank loans and credit risk prognosis. University of the Witwatersrand, Johannesburg. https://goo.gl/UVtQif (2015). Accessed 03 July 2017
16. Narain, B.: Survival analysis and the credit granting decision. In: Thomas, L.C., Crook, J.N., Edelman, D.B. (eds) Credit Scoring and Credit Control, pp. 109–122. Oxford University Press, Oxford (1992)
17. Okumu, A., Wekesa, M.S., Mwita, P.: Modelling credit risk for personal loans using product-limit estimator. Int. J. Financ. Res. **3**(1), 22–32 (2012)
18. Pazdera, J., Rychnovsky, M., Zahradník P. Survival analysis in credit scoringin credit scoring. In: Seminar on Modelling in Economics. Charles University, Prague. http://artax.karlin.mff.cuni.cz/~rychm5am/Project.pdf (2009). Accessed 03 July 2017
19. Sarlija, A., Bensic, M., Zekic-susac, M.: Modeling customer revolving credit scoring using logistic regression, survival analysis and neural networks. In: Proceedings of the 7th WSEAS International Conference on Neural Networks, pp. 164–169. Croatia, Cavtat (2006)
20. Stepanova, M., Thomas, L.: Survival analysis methods for personal loan data. Oper. Res. **50**(2), 277–289 (2002)
21. Therneau, T.: A package for survival analysis. R package version 2.41-3. https://cran.r-project.org/web/packages/survival/index.html (2017). Accessed 20 July 2017
22. Watkins, J., Vasnev, A., Gerlach, R.: Survival analysis for credit scoring: incidence and latency. In: OME Working Paper, The University of Sydney, Sydney. https://ses.library.usyd.edu.au/bitstream/2123/8161/1/OMWP_2009_03.pdf (2009). Accessed 03 July 2017
23. Zeldin, M.: Crowdinvesting in Russia: to be or not to be? Rusbase is an independent publication about technology and business, event organizer and creator of services for entrepreneurs, investors and corporations. https://rb.ru/opinion/zac/ (2016). Accessed 20 June 2017

What Do Employers Look for When Hiring New Graduates? Answers from the Electus Survey

Paolo Mariani, Andrea Marletta and Mariangela Zenga

Abstract This paper presents the main results obtained from Electus survey targeting 471 Lombardy companies with at least 15 employees. The project wants to acquire the knowledge about criteria for entrepreneurs in the choice for graduates demanding a job vacancy. This study, also, aims to evaluate the features of a graduate's profile employers for potential candidates in five job positions (Administration clerk; Human Resource assistant; ICT professional; Marketing assistant; CRM assistant). In order to estimate the entrepreneurs' preferences about skills and competencies for the new hirings, Conjoint Analysis is adopted. Finally, using a new definition of the relative importance of attributes, the analysis finds out the monetary value for skills owned by the candidates.

Keywords Labour market · Conjoint analysis · Monetary evaluation

1 Introduction

The relationship between the requested competencies by entrepreneurs and the skills owned by the new hirings has to be considered an essential step to understand the labour market dynamics. This knowledge represents a crucial point to make policy in general on employability, but in particular for youth employability. In a context of high unemployment, the productive world is averse in investing in the human

P. Mariani · A. Marletta (✉)
Department of Economics, Management and Statistics, University of Milano-Bicocca, Milano, Italy
e-mail: andrea.marletta@unimib.it

P. Mariani
e-mail: paolo.mariani@unimib.it

M. Zenga
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy
e-mail: mariangela.zenga@unimib.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_27

capital. In such tendency, the competition searching for a job, push the individuals to spend resources and time in education. Specially during the financial crisis period, the reduction of the mismatch between the demand and the offer in the labour market could refrain the unemployment youth rate and optimize the efficiency of educational resources. In relation to the labour market, it seems to be important the analysis of the companies and their expectations about a new hiring. It appears useful to understand the dynamic of the recruitment, in particular in this work the focus is on the importance of the competencies requested by the entrepreneurs to a new graduates. For this reason, the aim of this paper is to carry out an analysis of the employers' preferences for graduates' profiles evaluated as candidates in a job position by using a Conjoint Analysis [8]. Specifically, the analysis intends to detect how some characteristics of the new graduates can affect a possible future recruitment and retribution. Moreover, the paper would like to define toward some across the-board skills, universally recognized as "best practices" for a graduate. Finally, the analysis allows to achieve differences and valuations between wage and competencies for new graduates. From a methodological point of view, in the context of Conjoint Analysis, it is introduced the use of an index of relative importance of the attributes. The study is based on the multi-centre research, ELECTUS (Education-for-Labour Elicitation from Companies' Attitudes towards University Studies [5]) a research project involving several Italian universities.

The paper is organized as follows. Section 2 presents the ELECTUS research, Sect. 3 introduces the Conjoint analysis methodology and the new index of the relative importance for the Attributes, the results of the analysis are reported in Sect. 4. Section 5 is reserved to discussion and final remarks.

2 The Electus Project

The Electus project wants to acquire structured knowledge about criteria for entrepreneurs in the choice for graduates demanding a job vacancy. This aim is coherent with the European Commission criteria useful to define a contact point between the world of formation and the job market.

The results of the research aim to give a concrete help to the stakeholders operating in the job market, in particular:

- for graduates, they could knowingly address their ambitions in a professional field in relationship to the business market, searching for a correspondence between thier skills and what the companies are looking for;
- for universities, they could adopt educational methodologies and instructional activities for the definition of professional figures reducing the mismatch between the demand and the supply of the job market;
- for entrepreneurs, they have the opportunity to think about impartial criteria for the recruitment processes so that the candidates for a job vacancy have already all the requirements to be hired;

- for the policies of youth labour market, the market should become more fluid and recover more quickly from the economical crisis coming back to a high youth employment rate.

Data were collected using a software program called Sawtooth [12]. To make easier the participation of the entrepreneurs to the survey, the survey was conducted using CAWI technique. The survey consists in a brief questionnaire contained two sections: the conjoint experiment and general information about the company (demographic questions). In the conjoint experiment respondents have to mark 4 profiles for 5 different job positions from 1 over to 10. These profiles were built as a combination of 6 attributes. The combinations for all the alternatives provided by a full factorial fashion were numerous that it was necessary to reduce the possibilities using an ad-hoc fractional factorial design. At the end, the experiment design was both orthogonal and balanced. The experimental design was realized by the Sawtooth program itself.

The 5 job positions under observation are:

- Administration clerk;
- Human Resource assistant;
- ICT professional;
- Marketing assistant;
- CRM assistant.

To specify the candidates' profile, 6 attributes were used:

- *Field of Study* with 10 levels (Philosophy and literature, Educational sciences, Political science/ Sociology, Economics, Law, Statistics, Industrial engineering, Mathematics/ Computer sciences, Psychology, Foreign languages);
- *Degree Mark* with 3 levels (Low, Medium, High);
- *Degree Level* with 2 levels (Bachelor, Master);
- *English Knowledge* with 2 levels (Suitable for communication with foreigners, Inadequate for communication with foreigners);
- *Relevant work experience* with 4 levels (No experience at all, Internship during or after completion of university studies, Discontinuous or occasional work during university studies, One year or more of regular work);
- *Willigness to Travel on Business* with 3 levels (Unwilling to travel on business, Willing to travel on business only for short periods, Willing to travel on business even for long periods).

After having rated the selected profile and chosen the best one, the entrepreneurs had to propose a Gross Annual Salary for the chosen profile in order to measure the so-called 'willingness to pay' [2].

3 Methodology

In order to estimate the entrepreneurs' preferences about skills and competencies for the new hirings, in this work, Conjoint Analysis was adopted.

To define the aims and the ways of a business strategy, a company have to evaluate features, needs and the expected behaviors of the potential competitors. This implies a market segmentation of consumers in homogeneous groups, usually this information is used to address marketing policies taking into account their necessities. For our purposes, the object of the analysis is represented by graduates, so they will be grouped on the basis of some competencies.

Conjoint Analysis is a flexible segmentation technique starting from the expression of the preferences of the statistical units. Our statistical units are the entrepreneurs and the product that they are going to evaluate are the graduates.

It was introduced in 1964 by Luce and Tukey [9] and revised by Green and Srinivasan [7] in 1978. In a first step they defined the utility as the value related to a level of satisfaction obtained by a consumer using a product with some features. The Utility function assigns a level of satisfaction to each product. Conjoint Analysis has 3 peculiar features:

- decompositive nature: starting from a preference judgement about a product, it is possible to obtain values about single attributes of the good;
- individual estimates: Conjoint Analysis allows to derive a predictive model for each respondents;
- flexibility in the functional form: the relationship between dependent and independent variables is not established a-priori.

Working with Conjoint Analysis implies the definition of attributes, levels and profiles. The attributes or factors are the product's features, in this case the competencies of the graduates. The levels of the attributes represents all possible ways of expressing the attributes. The profiles are possible combinations of the levels of the attributes.

Conjoint Analysis creates a direct correspondence between the definition of utility and preference: if a product is preferred, its utility function will be higher. Starting from the preference, partial utilities are computed as the associated importance for each level of the attributes. Total utility is defined as the sum of partial utilities given by a combination of level of attributes.

The utility function U_k is defined as follows:

$$U_k = \sum_{i=0}^{n} \beta_i x_{ik} \tag{1}$$

where x_0 is equal to 1 and n is the number of all levels of the attributes which define the combination of a given good. Each variable x_{ij} is a dichotomous variable that refers to a specific attribute level; it equals 1 if the corresponding attribute level is present in the combination of attributes that describes the alternative k; otherwise,

that variable is 0. As a result, the utility associated with alternative k (U_k) is obtained by summing the terms $\beta_i \; x_{ik}$ over all attribute levels, where β_i is the partial change in U_k for the presence of attribute level i, holding all other variables constant. In this paper, it refers to this piece-wise linear function as a part-worth function model that gives a specific utility value for each level of the considered attributes, usually referred to as part-worth utility.

Conjoint Analysis also allows to evaluate the relative importance of the single attributes in the consumers' choice. For any attribute j, the relative importance can be computed by dividing its utility range by the sum of all utility ranges as follows:

$$I_j = \frac{\max\left(W_j\right) - \min\left(W_j\right)}{\sum_{j=1}^{J}\left[\max\left(W_j\right) - \min\left(W_j\right)\right]}, \tag{2}$$

where J is the number of attributes and W_j is the set of part-worth utilities referred to the various levels of attribute j. Usually, importance values are represented as percentages and have the property of summing to one hundred.

Part-worth utilities and importance indexes represent the starting point to obtain an economic valuation of the attributes [10]. This monetary valuation is obtained comparing all possible profiles with a baseline profile b and the related utility U_b.

The way to obtain this economic coefficient can be synthesized in 3 steps:

- computation of the variation in terms of utilities M_i;
- derivation of economic coefficient $MI_{(p),ij}$ using the importance indexes I_j;
- valorization of the coefficient V_{ij} using a the total revenue π associated to the baseline profile.

Step 1

The utility variation M_i is computed by replacing one attribute level of the baseline profile b with attribute level i using this formula:

$$M_i = \frac{U_i - U_b}{U_b} \tag{3}$$

where U_i denotes the sum of the utility scores associated with alternative profile i and U_b (assumed to be different from 0) denotes the sum of the part-worth utilities associated with the baseline profile b of the job. Equation (3) indicates whether the baseline profile b modification gives a loss or a gain. If $M_i = 0$, there is any loss or gain in terms of total utility. However, the utility change arising from an attribute-level modification can be considered more or less important by respondents. Hence, this change can have a more important economic impact respect to a utility modification, which has a similar intensity but involves a less relevant attribute. As a solution, it is used as weight the relative importance of the modified attribute [6].

Step 2

Not all variations in terms of utilities are equal, for this reason it is necessary to weigh M_i with the importance indexes I_j obtained using part-worth utilities. The coefficient formulation becomes the following:

$$MI_{ij} = M_i * I_j. \tag{4}$$

When the number of the levels varies widely among the attributes, it seems to be useful to take into account this variability directly in the computation of the Relative Importance of the Attribute.

The proposed approach could be intended as an extension of the coefficient of economic valuation already defined (see [10]). The extension consist in the use of the number of levels as possible factor to reduce this bias.

As the best of our knowledge, only few authors (see, for example [3, 11]) proposed a solution for this problematic issue. The philosophy of this indicator is based on the set of the part-worth utilities referred to the various levels of attribute j for each respondents t. For the tth units, the importance of the jth attribute with J levels could be defined in terms of the average range of the part-worths across the levels of that attribute:

$$Imp_{tj} = \frac{\max\left(W_{tj}\right) - \min\left(W_{tj}\right)}{J}, \tag{5}$$

In the Eq. 5, the effect of the number of the levels for the attribute is mitigated dividing the importance of the attribute by the number of the levels. At the end, for the tth respondent, the relative importance the jth attribute is given by:

$$I_{tj} = \frac{Imp_{tj}}{\sum_{j=1}^{J} Imp_{tj}}. \tag{6}$$

From Eq. 6, it is possible to give the sample distribution of the relative importance of the jth attribute and sample quantile of order p, $I_{(p),j}$. The use of the order statistics can increase the robustness of the analysis [4].

Otherwise, it is possible to express these importance values entering the sample quantile of order p for importance of the modified attribute:

$$MI_{(p),ij} = M_i * I_{(p),j}. \tag{7}$$

Step 3

Assuming a change in the baseline profile, the formula (7) is used to estimate the variation of the total revenue generated. Given the Gross Annual Salary (GAS) associated with the baseline profile, the coefficient of economic valuation is expressed as follows:

$$V_{(p),ij} = MI_{(p),ij} * GAS \tag{8}$$

where $V_{(p),ij}$ denotes the amount of the salary variation. The variation $V_{(p),ij}$ is obtained by supposing that the monetary attribute referred to the job varies in proportion to the change in total utility. This assumption may seem restrictive. However, it is possible to argue that the monetary amount asked for an employer concerning a job reflects how that user values the combination of attributes of the job in terms of utility.

Under this hypothesis, it is credible to assess the economic value of a change in the combination of attributes as a function of the utility and importance of the modified attribute. In addition, CA serves the scope of approximating the real structure of preferences, given that only a partial knowledge of preferences can be known. Therefore it is possible to use this coefficient as a monetary indicator that approximates the impact of a given utility change in monetary terms.

4 Application and Results

Conjoint Analysis is achieved in order to measure entrepreneurs' preferences. Data manipulation and Conjoint Analysis were obtained using *R* software and *Conjoint* package [1].

As far as the Milano-Bicocca research unit is concerned, interviewees were representatives of companies registered on the Portal of Almalaurea for recruitment and linkage, limited to the university site. The population of companies targeted was composed by 4.183 potential recruiters. Companies received a first e-mail inviting to complete the survey. If they did not answer after the first attempt, they were solicited to fill in the questionnaire for three times, once a week. After these attempts, final respondents were 471. Companies profile shows that they were in prevalence sized with 15–49 employers (52%), followed by sized, 50–249 employees enterprises (25.6%) and (22.4%) by sized at least 250 companies. The most represented activity sectors were services to the industry (62.1%), services to the person or the family (16.2%) and manufacturing (14.9%). The majority of companies (89.4%) operated fully or partially within the domestic market. Moreover, they were mainly under the management of the entrepreneur (64.2%). About the attitude towards a new hiring, 55.2% of the firms kept the same number of employees during last 3 years, while 33.3% increased their workforce, about the future more than 70% of the companies predict to hire a new resource.

As it is possible to note from Table 1, the Major preferred by respondents is Economics for Administrative Clerk, Marketing Assistant and Customer Relationship Management. A degree in Psychology is desirable for an Human Resource assistant, while for ICT professional the field of study with the biggest part-worth utility is Computer Sciences/Mathematics.

It is important to remember that, since for definition the sum of utilities for all levels of an attribute equals to 0, less desirable attributes could have negative utilities.

In this paper the definition of cross or specialized competencies is introduced. A competence is defined as a cross competence if part-worth utilities are higher independently from the chosen vacancy. On the other hand, if the level of the attribute changes over the job position, that competence is defined as specialized.

In the application, part-worth utilities seem to be similar for all the attributes, except for the attribute *Field of Study*. For this reason, according the previous definition *Field of study* is a specialized competence. This means that other competencies have some levels that are universally identified as 'best practice' for a graduate.

Table 1 Competencies part-worth utilities for job positions

Competencies	AC	HR	ICT	MKT	CRM
Field of study					
Philosophy and literature	−0.8312	0.1561	−0.6792	−0.1247	−0.5629
Educational sciences	−0.5959	0.8598	−0.0759	−0.2299	−0.2086
Political sciences	0.3031	0.1876	−0.7714	0.0313	0.1996
Economics	1.8811	0.3210	0.2981	1.3350	1.0165
Law	0.0737	0.5498	4.8612	−0.5211	−0.0909
Statistics	0.4506	−0.6956	0.3956	−0.0129	−0.1686
Engineering	−0.5488	−1.5581	0.8889	−0.4019	0.0469
Computer sciences	0.4444	2.9842	2.9842	−0.4163	0.0252
Psychology	−1.0678	1.5375	−1.0325	0.0974	−0.1557
Foreign languages	−0.1091	−0.2371	−1.1121	0.2431	−0.1015
Degree level					
Bachelor	0.0485	0.0251	−0.0483	−0.0092	−0.0586
Master	−0.0485	−0.0251	0.0483	0.0092	0.0586
Degree mark					
Low	−0.3960	−0.2497	−0.1047	−0.1407	−0.2299
Medium	0.2169	0.0950	−0.0431	0.0203	0.1401
High	0.1790	0.1547	0.1478	0.1204	0.0898
English knowledge					
Suitable	0.4608	0.2699	0.0969	0.3145	0.2998
Inadequate	−0.4608	−0.2699	−0.0969	−0.3145	−0.2998
Relevant work experience					
No experience	−0.3169	−0.1666	0.0303	−0.3177	−0.1619
Internship	−0.0045	−0.0019	−0.0182	−0.0464	−0.1313
Occasional	−0.1219	−0.1383	−0.1300	0.1736	0.1014
Regular	0.4433	0.3068	0.1179	0.1905	0.1918
Willigness to travel on business					
Unwilling to travel	−0.0793	−0.3530	−0.0768	−0.0862	−0.4198
Short period	−0.0279	0.0698	−0.0295	0.0610	0.2353
Long period	0.1072	0.2832	0.1063	0.0252	0.1845

Source Electus data (2015)
AC = Administration Clerk
HR = Human Resource assistant
ICT = Information Communication Technology professionals
MKT = Marketing assistant
CRM = Customer Relationship Management

Table 2 Competencies attributes and ideal levels for job vacancies

Competencies	AC	HR	ICT	MKT	CRM
Field of study	Economic	Psychology	Comp.Sci	Economic	Economic
Degree level	Bachelor	Bachelor	Master	Master	Master
Degree mark	Medium	High	High	High	Medium
English knowledge	Suitable	Suitable	Suitable	Suitable	Suitable
Relevant work experience	Regular	Regular	Regular	Regular	Regular
Willingness to travel	Long	Long	Long	Short	Short

Source Electus data (2015)

Utility scores for variable *Degree level* are very close to 0 for each position. This means that there is no significant difference between a bachelor and a master degree for the respondents. This is due to the fact that all analyzed position are very basic and they do not require specialized skills. *Degree Mark* is a skill where best two levels are preferred, so a medium-high marked degree is preferable among candidates. *English Knowledge* shows the highest utility for candidates with capability to develop a fluent communication with foreigners. The attributes named *Relevant work experience* shows a positive score only for graduates with one or more years of regular work. Finally, the *Willingness to Travel on Business* to short or long period leads is a very appreciated quality for candidates.

In Table 2, ideal profiles for each job vacancy are shown. As it is possible to note, ideal profiles are similar each other except for *Field of Study*. This confirms the theory of the existence of some cross or specialized competencies.

The attributes for *Relevant work experience* and *English Knowledge* shows that a best level does not depend from the task they are going to face, so they could be considered as cross competencies. After all, it is easy to imagine that companies prefer to employ a candidate with one year or more of regular work and suitable for communication with foreigners.

Since for attributes *Degree Mark* and *Willingness to travel on business* two levels are recognized as 'best practices', they could be defined as quasi-cross competencies.

Finally, since part-worth utilities for variable *Degree level* are very close to 0 and there is no difference between the levels, this could be defined as a not-binding attribute.

The use of a weighted matrix for individual scores allowed to obtain individual contribution for the index of importance. Since it is well-known and empirically proved by this research a entrenched relationship between *Field of study* and the entrepreneurs' choice for a job vacancy, it seems plausible the application of the modified version of the index presented in Eq. 6. The distribution of the individual contribution has been used to build a non-parametric confidence interval for index of importance. In Table 3, it is shown the comparison between two methods for the computation of the index of importance. Before using the average range Imp_{ij}, the

Table 3 Competencies attributes and ideal levels for job vacancies

Job position	AC		HR		ICT		MKT		CRM	
Competencies	I_j	$I_{0.5,j}$	I_j	$I_{0.5,j}$	I_j	$I_{0.5,j}$	I_j	$I_{0.5,j}$	I_j	$I_{0.5,j}$
Field of study	53%(1)	14%(3)	59%(1)	20%(3)	81%(1)	37%(1)	54%(1)	11%(4)	43%(1)	8%(5)
Degree level	2%(6)	4%(6)	1%(6)	3%(6)	2%(6)	9%(6)	1%(6)	1%(6)	3%(6)	7%(6)
Degree mark	11%(5)	14%(4)	8%(5)	14%(4)	5%(3)	14%(3)	8%(4)	13%(3)	10%(4)	14%(3)
English knowledge	17%(2)	41%(1)	11%(3)	29%(1)	4%(4)	18%(2)	18%(2)	48%(1)	16%(3)	35%(1)
Relevant work experience	14%(3)	16%(2)	9%(4)	12%(4)	5%(2)	9%(5)	15%(3)	18%(2)	10%(5)	11%(4)
Willingness to travel	13%(4)	5%(5)	12%(5)	22%(2)	4%(5)	10%(4)	4%(5)	7%(5)	18%(2)	25%(2)

Source Electus data (2015)

influence of Field of Study was prevalent for all job positions. For each vacancy, in the second column, it is reported the median $I_{0.5,j}$ of the distribution of individual indexes of importance. According to this method, the predominant attribute is now the English Knowledge. This is because of the fact that this competence has only two levels.

In relation to the other skills, there is not a so big difference between the two methods, so Degree Mark, Relevant work experience and Willingness to travel are in intermediate position with values a little bit over the 10% and finally the Degree level is still the least relevant competence.

Here, the considered GAS is assigned to the best profile, for this reason all monetary variation will be negative. This amount is the result of a specific question in the survey in which the respondents should assign a Gross Annual Salary for the new hired profile. This is an average value corresponding only to respondents that selected a best candidate for each position.

New monetary variations are still proportional to part-worth utilities, therefore attributes with low utility scores correspond lower monetary variations (Tables 4, 5, 6, 7, 8 and 9).

About Field of Study, new monetary variations are reduced in comparison with the first approach due to the dramatic decrease of the index of importance due to the new indicator taking into account the number of levels for an attribute. For this reason, *Field of Study* appears as the most penalized attribute and the biggest monetary decreasing amounts to 4.791,65€ for a graduate in Foreign Languages for Information and Communication Technologies professional when compared with a degree in Computer Sciences. This figure is the one needs more specialization and this is confirmed but the highest importance indexes for Field of Study (see Table 3). Since it was defined Field of Study as a specialized competence, the economic variations change over the job position.

As said before, *Degree Level* is the least relevant quality for the respondents, so monetary variations are very low varying over the job vacancies from no difference between a Bachelor and a Master degree for Marketing assistant over to 31.20€ for a Bachelor graduate looking for a job as Customer Relationship Management.

Table 4 Monetary varations $V_{(p),ij}$ for Field of Study

Job position	AC	HR	ICT	MKT	CRM
Field of study	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$
Philosophy and literature	−1.152,00	−959,40	−4.284,05	−650,00	−422,40
Educational sciences	−1.051,20	−457,60	−3.579,05	−696,80	−328,80
Political sciences	−669,60	−941,20	−4.392,15	−582,40	−218,40
Economics	0,00	−837,20	−3.141,95	0,00	0,00
Law	−768,00	−686,40	−4.537,85	−826,80	−297,60
Statistics	−607,20	−1.557,40	−3.026,80	−600,60	−316,80
Engineering	−1.032,00	−2.152,80	−2.451,05	−774,80	−259,20
Computer sciences	−609,60	−1.856,40	0,00	−780,00	−266,40
Psychology	−1.252,80	0,00	−4.697,65	−551,20	−314,40
Foreign languages	−844,80	−1.237,60	−4.791,65	−486,20	−300,00

Source Electus data (2015)

Table 5 Monetary variations $V_{(p),ij}$ for degree level

Job position	AC	HR	ICT	MKT	CRM
Degree level	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$
Bachelor	0,00	0,00	−28,20	0,00	−31,20
Master	−14,00	−5,20	0,00	0,00	0,00

Source Electus data (2015)

Table 6 Monetary varation for degree mark

Job position	AC	HR	ICT	MKT	CRM
Degree Mark	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$
Low	−389,20	−195,00	−112,80	−137,80	−199,20
Medium	0,00	−31,20	−84,60	−52,00	0,00
High	−25,20	0,00	0,00	0,00	−26,40

Source Electus data (2015)

Table 6 shows the difference for levels of *Degree Mark*, to have a Medium or High Mark appears to be not significantly different from 0. The variation is relevant when the comparison is with a low mark graduate and its value lies in the interval from 112,80€ for an ICT Professional over to 389,20€ for an Administration Clerk.

Table 7 Monetary varations $V_{(p),ij}$ for english knowledge

Job position	AC	HR	ICT	MKT	CRM
English knowledge	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$
Suitable	0,00	0,00	0,00	0,00	0,00
Inadequate	−1.355,60	−553,80	−110,45	−1.224,60	−804,00

Source Electus data (2015)

Table 8 Monetary variations $V_{(p),ij}$ for work experience

Job position	AC	HR	ICT	MKT	CRM
Work experience	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$
No experience	−431,20	−197,60	−25,85	−371,80	−134,40
Internship	−254,80	−124,80	−37,60	−171,60	−124,80
Occasional	−319,20	−184,60	−70,50	−13,00	−33,60
Regular	0,00	0,00	0,00	0,00	0,00

Source Electus data (2015)

Table 9 Monetary varations $V_{(p),ij}$ for willingness to travel

Job Position	AC	HR	ICT	MKT	CRM
Willingness to travel	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$	$V_{(0.5),ij}$
Unwilling to travel	−33,60	−488,80	−58,75	−41,60	−626,40
Short period	−25,20	−166,40	−42,30	0,00	0,00
Long period	0,00	0,00	0,00	−10,40	−48,00

Source Electus data (2015)

An interesting value is assumed by variations about *English Knowledge*, so since the new method was introduced, it became the more requested skill and its interval varies from 110,45€ for an ICT Professional over to 1.355,60€ for an Administration Clerk for a graduate with no capability to communicate with foreign people.

About *Work experience*, the $V_{(p),ij}$ coefficients varies in the interval from 13.00€ for a graduate with occasional experience for a job as Marketing Assistant over to 431,20€ for a graduate with no regular work experience in Administration Clerk.

As already occurred for the *Degree Mark*, even for *Willingness to Travel on Business*, there is only a level significantly different from the baseline level, so the $V_{(p),ij}$ coefficients are significant differently from 0 only for graduates unwilling to travel, varying from 33,60€ for an Administration Clerk over to 626,40€ for the CRM assistant. This means that respondents required the willingness to travel, it does not matter if for short or long periods.

5 Conclusions

The analysis of the importance of the competencies requested by the entrepreneurs to a new graduates could be considered a crucial point to understand and try to reduce the mismatch between the Higher Education and the Labour Market. This work presents an analysis of the preferences for new graduates' profiles five positions, reporting differences and valuations between wage and competencies. The focus is on the Labour market for new graduates. The study is based on the multi-centre research ELECTUS. From a methodological point of view, the paper uses a new index of relative importance of the attributes in the context of Conjoint Analysis. This index is based on the average range between the levels of the attributes and results very useful in all those cases in which the range takes values in a spread interval.

The results lead to define the best profile of a graduate. The analysis underlines the presence of some cross competencies common for the five positions, in fact companies seem to prefer a candidate with one year or more of regular work and suitable for communication with foreigners. In general a medium-high marked degree and the willingness to travel on Business to short or long period leads are appreciated quality for candidates. These could be defined as quasi-cross competencies. Obviously, the field of study competence is typical of the job position and it can be considered a specialized competence. If Economics is recognized as the preferred attribute for position of Administration clerk, Marketing assistant and Customer Relationship Management, Psychology results the best for Human Resource assistant and Computer Sciences is the most suitable for Information Communication Technology professionals.

The study shows also the differences, in terms of wage, among the several profiles of new graduates considering the levels of attributes less eligible for the job positions.

Future research will focus the attention on the results coming from a stratification based on socio-demographic features of companies, using also the relative importance of the attributes for the five job positions proposed in the survey.

References

1. Bak, A., Bartlomowicz, T.: Conjoint analysis method and its implementation in conjoint R package. Data Analysis Methods and its Applications, pp. 239–248 (2012)
2. Breidert, C., Hahsler, M., Reutterer, T.: A review of methods for measuring willingness-to-pay. Innov. Mark. **2**(4), 8–32 (2006)
3. Danaher, P.J.: Using conjoint analysis to determine the relative importance of service attributes measured in customer satisfaction surveys. J. Retail. **73**(2), 235–260 (1997)
4. David H.A. and Nagaraja H.N.: Order Statistics, 3rd edn. Wiley, Hoboken
5. Fabbris, L., Scioni, M.: Dimensionality of scores obtained with a paired-comparison tournament system of questionnaire item. In: Meerman, A., Kliewe, T. (eds.) Academic Proceedings of the 2015 University-Industry Interaction Conference: Challenges and Solutions for Fostering Entrepreneurial Universities and Collaborative Innovation (2015)

6. Garavaglia, C., Mariani, P.: How much do consumers value protected designation of origin certifications? Estimates of willingness to pay for pdo dry-cured ham in Italy. Agribusiness **33**(3), 403–423 (2017)
7. Green, P.E., Srinivasan, V.: Conjoint analysis in consumer research: issues and outlook. J. Consum. Res. **5**(2), 103–123 (1978)
8. Lancaster, K.J.: A new approach to consumer theory. J. Polit. Econ. **74**(2), 132–157 (1966)
9. Luce, R.D., Tukey, J.W.: Simultaneous conjoint measurement: a new type of fundamental measurement. J. Math. Psychol. **1**(1), 1–27 (1964)
10. Mariani, P., Mussini, M.: A new coefficient of economic valuation based on utility scores. Argum. Oecon. **30**(1), 33–46 (2013)
11. Mezbahur, R., Lorica, B.G.: Attribute relative importance computation in conjoint analysis. J. Inf. Optim. Sci. **20**(1), 113–120 (1999)
12. Sawtooth, S.: http://www.sawtoothsoftware.com (2017)

Modeling Household Income with Contaminated Unimodal Distributions

Angelo Mazza and Antonio Punzo

Abstract In many countries, income inequality has reached its highest level over the past half century. In the labor market, the technological progress has widened the earnings gap between high- and low-skilled workers. Changes in the structure of households, with a growing percentage of single-headed households, and in family formation, with an increased earnings correlation among partners in couples, is contributing in increasing inequality. A key step in measuring income inequality is the estimation of the income distribution, due to the sensitivity of usual inequality measures to extreme values. To deal with this issue, we propose the use of contaminated lognormal and gamma models and we derive the formulations for computing the Gini index based on the model parameters. An application to 101 empirical income distributions that include countries at different development stages is presented.

Keywords Heavy-tailed distributions · Income distribution · Gini index

1 Introduction

The distribution of households across income categories is a significant demographic and economic characteristic. Income distributions provide information on the inequality of an area's economic well-being that is not reported by summary measures, such as the mean and median [17]. Income distributions are used to assess changes in inequality and poverty over time, to compare measures across countries, over time and before and after specific policy changes, designed, for example, to alleviate poverty [10].

Income inequality has been increasing, almost without interruption, after the late 1960s in most OECD countries, and it is at its highest level for the past half century

A. Mazza · A. Punzo (✉)
Department of Economics and Business, University of Catania, Catania, Italy
e-mail: antonio.punzo@unict.it

A. Mazza
e-mail: a.mazza@unict.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_28

[18]. One reason for the increasing inequality is in the difference between the demand for and supply of skills in the job market, with a consistent growth of the earnings gap between high- and low-skilled workers. Technological progress has been more beneficial for workers with higher skills, who have enjoyed significant income gains, while workers with lower skills have been left behind [18].

Income inequality is also closely related to the changes in household structures and family formation. A growing percentage of individuals live in families with only a single adult head; in OECD countries, single-headed households have risen from an average of 15% in the late 1980s to 20% in the mid-2000s [9]. Whereas some of them (e.g., single parents) are more likely to be poorer than they would be in families containing two adult heads [21], others may belong among high earners (prime-age singles). Therefore, an increase in the proportion of single-headed households may contribute to widening the household income inequality [9].

The rise in female labor income would reduce household inequality if growth in female earnings were concentrated among families that otherwise would have received low incomes. However, when assortative mating patterns are in place (i.e., tendency to choose one's spouse in groups of similar earnings and/or educational levels), earnings gains concentrate among families that would have been well off even without a woman's earnings [21]. In OECD countries, as reported by [18], in 40% of couples both partners belonged to the same or close earnings deciles, compared with 33% about 20 years before. Similarly, in [9] is reported that the correlation coefficients between husbands and wives earnings have increased notably over time in 20 out of 23 OECD countries, suggesting that there is a general trend toward stronger marital sorting by earnings.

The estimation of the income distribution plays a major role in measuring income inequality, and both parametric and nonparametric approaches have been proposed [10]. Within this context, parametric estimation has the advantage of facilitating subsequent inferences about inequality and poverty measures, based on the estimated distribution parameters.

A large number of alternative parametric models have been suggested in the literature (for a survey, see [22]). As well documented in [14], a convenient parametric model should be: defined on a strictly positive support, unimodal, and positively skewed; moreover, all the parameters of the specified model should have a well-defined economic meaning and, following a principle of parsimony, the model should make use of the smallest possible number of parameters for adequate and meaningful representation. Under these conditions, two of the models most frequently applied are the 2-parameter lognormal distribution [19] and the 2-parameter gamma distribution [3], with empirical evidence in favor of the gamma over the lognormal distribution, judging by goodness of fit criteria, as shown by [39] for the USA and [5] for the Netherlands.

Unfortunately, as emphasized by [13, 16, 45], real income data are often "contaminated" by outliers (referred to as outlying incomes herein, in analogy with [1])—at one or both ends of the distribution—that affect the estimation of the parameters for the chosen model. This in turn will affect the inequality measure computed from the estimated parameters. Thus, the detection of outlying incomes, and the develop-

ment of robust methods of parameter estimation insensitive to their presence, is an important problem. As suggested by [15], outlying incomes should be defined with respect to a reference distribution; that is, the shape of the "normal" incomes has to be assumed to define what a outlying income is, and the region of outlying incomes can be defined, e.g., as a region where the density of the reference distribution is low. By choosing, for parsimony sake, a 2-parameter unimodal model as reference distribution, and parameterizing it with respect to the mode λ and to another parameter ν that is closely related to the distribution variability, we consider the simple family of 4-parameter contaminated unimodal models, introduced by [44], in order to accommodate all the available incomes (see also [30]). The model is a 2-component mixture in which one of the components, with a large prior probability, represents the normal incomes (reference distribution), and the other, with a small prior probability, the same mode, and an inflated ν-parameter, represents the outlying incomes. It represents a simple theoretical model for the occurrence of outlying incomes and the two additional parameters, with respect to the parameters of the reference distribution, have a direct interpretation in terms of proportion of normal incomes and degree of contamination (a sort of measure of how different outlying incomes are from the bulk of the normal incomes). Advantageously, these contaminated models also allow for automatic detection of outlying incomes via a simple and natural procedure based on maximum *a posteriori* probabilities. As examples of mode-parameterized unimodal reference distributions we will consider the gamma and the lognormal densities. This choice is also justified by the fact that gamma and lognormal densities are known to be nice for modelling mid range incomes [12]; thus, their contamination allows to fit better the tails, which is a fundamental aspect that typically yields the definition of more complicated (less parsimonious) distributions for the whole income range.

2 A General Framework for Contaminated Unimodal Densities Definite on a Positive Support

Let X be the positive random variable denoting the income. Requiring, as usual, that the density $p(x)$ of X should be unimodal and positively skewed (cf. [14], p. 10), we can use for $p(x)$ the general (4-parameter) contaminated unimodal model of [44]; see also [29]. According to this model, the density function is written as

$$p(x;\boldsymbol{\vartheta}) = \alpha f(x;\lambda,\nu) + (1-\alpha) f(x;\lambda,\eta\nu), \quad x > 0, \tag{1}$$

where $\boldsymbol{\vartheta} = (\alpha, \lambda, \nu, \eta)'$ and

- $f(x;\lambda,\nu)$ is the unimodal density chosen as reference distribution for the income, with $\lambda > 0$ denoting the mode and $\nu > 0$ governing the concentration of f around the mode.

- $\alpha \in (0.5, 1)$ can be seen as the proportion of normal incomes. Note that α is constrained to be greater than 0.5 because, in robust statistics, it is usually assumed that at least half of the observations are normal.
- $\eta > 1$ denotes the degree of contamination and, because of the assumption $\eta > 1$, it can be interpreted as the increase in variability due to the outlying incomes with respect to the reference distribution $f(x; \lambda, \nu)$; hence, it is an inflation parameter.

Of course, because both the reference distribution $f(x; \lambda, \nu)$ and the inflated distribution $f(x; \lambda, \eta\nu)$ have their maximum in λ, this also guarantees that p will produce a unimodal density with mode λ. As a limiting case, when $\alpha \to 1^-$ and $\eta \to 1^+$, the reference distribution $f(x; \lambda, \nu)$ is obtained.

More specifically, among the existing 2-parameter distributions that can be used for f, we have chosen to adopt unimodal gamma and lognormal densities parametrized with respect to the mode. In the case of unimodal gamma distributions, the adjective "unimodal" is useful to highlight the subclass of gamma densities on which attention is focused on. However, other distributions defined on a positive support may be used if they can be mode-parametrized; an example could be represented by the Weibull distribution [5]. Examples of contaminated skewed distributions, applied for mixture modelling, are given in [27, 32].

An advantage of model (1) is that, once $\boldsymbol{\vartheta}$ is estimated, say $\widehat{\boldsymbol{\vartheta}}$, we can establish whether a generic income, say x^*, is either normal or outlying via the empirical posterior probability

$$P\left(x^* \text{ is normal} \,\middle|\, \widehat{\boldsymbol{\vartheta}}\right) = \frac{\widehat{\alpha} f\left(x^*; \widehat{\lambda}, \widehat{\nu}\right)}{p\left(x^*; \widehat{\boldsymbol{\vartheta}}\right)}. \tag{2}$$

Based on (2), x^* will be considered normal if $P(x^* \text{ is normal}|\widehat{\boldsymbol{\vartheta}}) > 1/2$, while it will be considered outlier otherwise. The resulting information, if desired, can be used to eliminate the outlying incomes [6]; however, we do not pursue this trimming approach because outliers are automatically down-weighted in the maximum likelihood estimation of the parameters (see [24, 25, 31, 33–35] for a discussion about down-weighting with respect to the contaminated normal distribution).

In the following, formulation and properties of the adopted mode-parametrized unimodal gamma and lognormal densities are outlined.

2.1 Mode-Parametrized Unimodal Gamma Distribution

In order to define a reference distribution for the income, to be inserted in (1), the following subclass of mode-parametrized unimodal gamma distributions is here considered:

$$f(x; \lambda, \nu) = \frac{x^{\frac{\lambda}{\nu}} e^{-\frac{x}{\nu}}}{\nu^{\frac{\lambda}{\nu}+1} \Gamma\left(\frac{\lambda}{\nu} + 1\right)}, \quad x > 0, \tag{3}$$

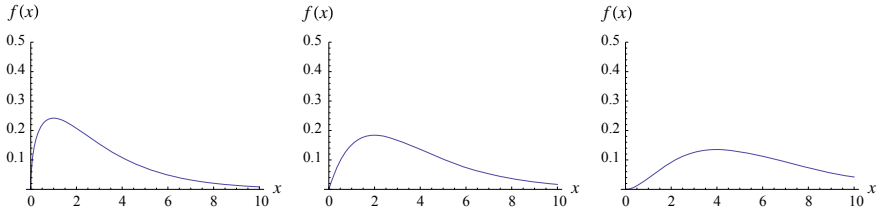


Fig. 1 Mode-parameterized unimodal gamma densities (3) with $\nu = 2$

with $\lambda > 0$ and $\nu > 0$.

Although the standard parameterization, given by

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)}, \quad x > 0, \tag{4}$$

with $\alpha, \beta > 0$, differs from (3), these parameterizations are related by

$$\begin{cases} \alpha = \dfrac{\lambda}{\nu} + 1 \\ \beta = \nu \end{cases} \quad \Rightarrow \quad \begin{cases} \lambda = \beta(\alpha - 1) \\ \nu = \beta \end{cases}. \tag{5}$$

Because $\lambda > 0$ and $\nu > 0$, (3) coincides with (4) under the constraints $\beta > 0$ and $\alpha > 1$. From the standard theory on (4), we know that if $\alpha \geq 1$, $f(x)$ has a single mode at $\beta(\alpha - 1)$, while if $\alpha \in (0, 1)$, $f(x)$ tends to infinity as $x \to 0^+$ (see [20], p. 168). To summarize, we are focusing only on the subclass of unimodal gamma densities, omitting all the (unlimited) reverse J-shaped cases that have a vertical asymptote in $x = 0$.

The shape of the unimodal gamma densities in (3) changes according to the value of λ; this is shown by a set of gamma densities displayed in Fig. 1. The variance of a random variable with density function (3) is

$$\nu^2 + \lambda\nu. \tag{6}$$

The last expression, analyzed as a function of λ, is a straight line with a positive slope ν; consequently, for fixed ν, the variability increases in line with the value of λ. Conversely, fixing λ in (6), the variance increases if ν increases, confirming that ν governs the spread of the distribution. The effect of varying ν, for fixed λ, is illustrated in Fig. 2. Further details about the parameterization of the gamma density given in (3) can be found in [4, 8].

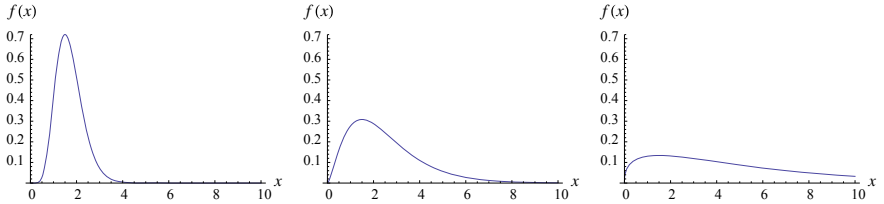


Fig. 2 Mode-parameterized unimodal gamma densities (3) with $\lambda = 1.5$

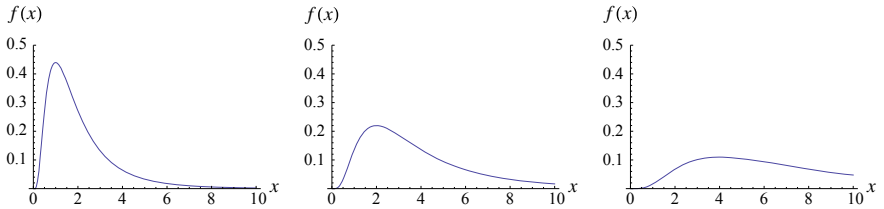


Fig. 3 Mode-parameterized lognormal densities (8) with $\nu = 0.5$

2.2 Mode-Parametrized Lognormal Distribution

The lognormal distribution given by

$$f(x;\mu,\sigma) = \frac{e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma x}, \quad x > 0, \tag{7}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$, is (already) unimodal with mode at $e^{\mu-\sigma^2}$ (see [20], p. 117). In order to consider model (7) in (1), as a reference distribution for the income, we consider the following mode-parametrized lognormal distribution

$$f(x;\lambda,\nu) = \frac{e^{-\frac{(\ln x-\ln\lambda-\nu)^2}{2\nu}}}{\sqrt{2\pi\nu}x}, \quad x > 0, \tag{8}$$

with $\lambda > 0$ and $\nu > 0$. The parameterizations (7) and (8) are directly related by (Figs. 3 and 4)

$$\begin{cases} \mu = \ln\lambda + \nu \\ \sigma^2 = \nu \end{cases} \quad \Rightarrow \quad \begin{cases} \lambda = e^{\mu-\sigma^2} \\ \nu = \sigma^2 \end{cases}. \tag{9}$$

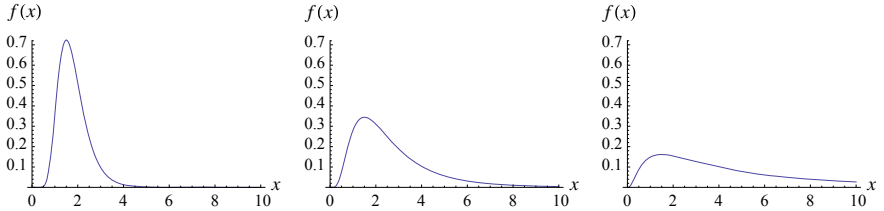


Fig. 4 Mode-parameterized lognormal densities (8) with $\lambda = 1.5$

3 Maximum Weighted Likelihood Estimation

Household income data often come from surveys; in such a case, denoting with n the sample size, a sample weight, say w_i, is assigned to each household income, say x_i, $i = 1, \ldots, n$, to correct for imperfections in the sample that might lead to bias and other departures between the sample and the reference population. Such imperfections include the selection of units with unequal probabilities, non-coverage of the population, and non-response. Therefore, the vector of parameters $\boldsymbol{\vartheta}$ of the contaminated density $p(\cdot; \boldsymbol{\vartheta})$ in (1) needs to be estimated by maximizing the weighted log-likelihood function (see e.g. [41], Sect. 3.4.4)

$$l(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} w_i \ln [p(x_i; \boldsymbol{\vartheta})]. \tag{10}$$

Operationally, maximization of (10) with respect to $\boldsymbol{\vartheta}$ is obtained by the general-purpose optimizer `optim()` for R, included in the **stats** package. The BFGS algorithm, passed to `optim()` via the argument `method`, is used for maximization.

Naturally, the choice of the starting values for the BFGS algorithm constitutes an important issue. The standard initialization consists of selecting a value for $\boldsymbol{\vartheta}^{(1)}$, value of $\boldsymbol{\vartheta}$ at the first iteration of the algorithm. Instead of selecting $\boldsymbol{\vartheta}^{(1)}$ randomly, we use the following technique. As already said in Sect. 2, when $\alpha \to 1^-$ and $\eta \to 1^+$, the contaminated density $p(x; \boldsymbol{\vartheta})$ in (1) tends to the reference distribution $f(x; \lambda, \nu)$. Then, the maximum weighted likelihood estimates of the parameters λ and ν for the reference distribution, along with the constraints $\alpha = \widetilde{\alpha}$ (with $\widetilde{\alpha} \to 1^-$) and $\eta = \widetilde{\eta}$ (with $\widetilde{\eta} \to 1^+$), can be used to initialize the contaminated model; in the analyses of Sect. 6, we use $\widetilde{\alpha} = 0.999$ and $\widetilde{\eta} = 1.001$. From an operational point of view, thanks to the monotonicity property of the BFGS algorithm, this also guarantees that the pseudo log-likelihood of the contaminated model will be always greater than, or equal to, the pseudo log-likelihood of the reference model. This is a fundamental consideration for the use of likelihood-based model selection criteria for choosing between the reference model and its corresponding contaminated version.

4 The Gini Coefficient

As mentioned in Sect. 1, one of the most important uses of the estimated income distribution is the evaluation of income inequality. Among all the inequality measures, the Gini coefficient is perhaps the most useful—and certainly the most widely used—measure of changes in inequality [7].

With respect to the contaminated density $p(x;\boldsymbol{\vartheta})$ in (1), the Gini coefficient, simply denoted as G, can be computed as

$$G(\boldsymbol{\vartheta}) = 1 - \frac{1}{E(X;\boldsymbol{\vartheta})}\int_0^{\infty}[1 - F(x;\boldsymbol{\vartheta})]^2\,dx, \tag{11}$$

where

$$F(x;\boldsymbol{\vartheta}) = \alpha H(x;\lambda,\nu) + (1-\alpha)\,H(x;\lambda,\eta\nu), \quad x > 0, \tag{12}$$

denotes the contaminated cumulative distribution function (c.d.f.), being $H(x;\lambda,\nu)$ the c.d.f. related to the reference distribution $f(x;\lambda,\nu)$, while

$$E(X;\boldsymbol{\vartheta}) = \alpha E(X;\lambda,\nu) + (1-\alpha)\,E(X;\lambda,\eta\nu), \tag{13}$$

denotes the expectation of the contaminated density, being $E(X;\lambda,\nu)$ the expectation of $f(x;\lambda,\nu)$. Details about $F(x;\boldsymbol{\vartheta})$ and $E(X;\boldsymbol{\vartheta})$ are given in Appendix 7 for the contaminated gamma and lognormal densities. However, regardless from the considered distribution, the integral in (11) has to be calculated numerically, as often happens in the literature for several parametric models for the income distribution [26]. For the calculation of this integral we use the R function `integrate()`.

5 Model Selection

In comparing nested/non-nested models which can/cannot differ in the number of parameters, the need arises to find an automatic way to select the best one. One way (the usual way) to perform model selection is via computation of a convenient (likelihood-based) model selection criterion across all competing models, and then choosing the model associated with the best value of the adopted criterion [38]. In the standard inferential context, famous examples are: the Akaike information criterion (AIC; [2]), the Takeuchi information criterion (TIC; [43]), and the Bayesian information criterion (BIC; [40]). However, they are not valid under the complex survey scenario described in Sect. 3.

The dAIC has been recently introduced by [23] as a model selection criterion to be used under complex sampling schemes. According to the notation introduced in Sect. 3, the dAIC can be written as

$$\mathrm{dAIC} = -2l\left(\widehat{\boldsymbol{\vartheta}}\right) + 2\,\mathrm{tr}\left(\widehat{V}^{-1}\widehat{U}\right), \tag{14}$$

where

$$\widehat{V} = -\sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} v\left(x_i; \widehat{\boldsymbol{\vartheta}}\right)$$

and

$$\widehat{U} = \sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} \left[u\left(x_i; \widehat{\boldsymbol{\vartheta}}\right) u\left(x_i; \widehat{\boldsymbol{\vartheta}}\right)'\right].$$

The dAIC in (14) coincides with the TIC when there are not sample weights [23].

6 Applications to Real Income Data

6.1 Description of the Income Data

In this section we fit the lognormal and gamma distributions and their contaminated counterparts to a diverse set of countries and over several years. This allows to test the strength of the competing models over countries with very different income distributions, notably formerly Communist or in earlier stages of economic development.

Household income data are from the Luxembourg Income Study (LIS) database (http://www.lisdatacenter.org/) for 31 upper- and middle-income countries; see Table 1. Data are grouped into 5-year waves and, for some countries, go back as far as the 1978; in total, we employed 101 different datasets. Additional information on each dataset can be accessed at http://www.lisdatacenter.org/our-data/lis-database/documentation/list-of-datasets/.

All computations were performed using custom programs for the R computing environment [37]. The R code was executed on LISSY, a remote-execution system that allows researchers to access from remote location the LIS microdata while respecting privacy restrictions required by the countries providing the data (see http://www.lisdatacenter.org/data-access/lissy/).

6.2 Results

Table 1 shows, for each empirical income distribution considered, values of the dAIC obtained fitting the four competing models and the corresponding estimated Gini

Table 1 Income distribution models fitted. For each dataset, the model with the best dAIC is in boldface

Country	Year	Log-Normal		ContLog-Normal				Gamma		ContGamma				Usual Gini	
		dAIC	Gini	dAIC	α	η	Gini	dAIC	Gini	dAIC	α	η	Gini	Unwins.	Winsorized
Australia	2008	174,842	0.46	174,721	0.99	4.50	0.53	173,692	0.39	**173,452**	0.96	4.66	0.39	0.37	0.36
	2010	171,787	0.46	171,787	1.00	1.00	0.46	170,441	0.38	**170,323**	0.99	5.32	0.38	0.37	0.36
Austria	1994	175,224	0.43	175,224	1.00	1.00	0.43	174,072	0.36	**173,749**	0.97	6.54	0.36	0.34	0.33
	1997	179,381	0.42	179,247	0.99	4.72	0.45	177,976	0.35	**177,878**	0.97	4.26	0.35	0.33	0.32
	2000	173,667	0.38	173,591	0.99	3.59	0.39	172,578	0.33	**172,474**	0.72	2.67	0.32	0.31	0.31
Belgium	1985	181,792	0.29	181,310	0.99	6.79	0.32	180,965	0.27	**180,779**	0.89	3.55	0.26	0.26	0.25
	1988	178,367	0.29	177,995	0.97	5.21	0.32	177,846	0.26	**177,546**	0.94	4.97	0.27	0.26	0.25
	1992	186,526	0.33	186,401	0.99	4.98	0.33	185,879	0.29	**185,854**	0.91	2.31	0.29	0.29	0.29
	1995	156,394	0.39	156,302	0.97	3.39	0.43	155,574	0.34	**155,231**	0.97	6.79	0.35	0.32	0.31
	1997	177,244	0.38	177,191	0.98	3.31	0.38	176,262	0.33	**176,227**	0.52	2.20	0.33	0.31	0.31
	2000	135,977	0.38	135,553	0.91	5.18	0.44	137,095	0.38	**134,782**	0.99	67.55	0.40	0.36	0.29
Colombia	2004	213,438	0.53	213,423	0.96	1.84	0.55	214,722	0.49	**213,285**	0.84	7.21	0.54	0.54	0.51
	2007	280,442	0.61	280,346	1.00	1.00	0.84	281,003	0.50	**279,508**	0.84	6.29	0.57	0.57	0.55
	2010	274,674	0.57	274,674	1.00	1.00	0.57	275,604	0.50	**274,069**	0.89	6.92	0.55	0.55	0.52
	2013	265,734	0.59	265,734	1.00	1.00	0.59	265,913	0.50	**264,961**	0.87	5.38	0.54	0.54	0.52
Czech republic	1996	217,629	0.74	217,629	1.00	1.00	0.73	**213,488**	0.50	213,488	1.00	1.00	0.51	0.43	0.41
	2007	209,492	0.69	209,492	1.00	1.00	0.69	**205,314**	0.48	205,314	1.00	1.00	0.48	0.39	0.38
	2010	213,149	0.72	213,149	1.00	1.00	0.71	**208,548**	0.49	208,548	1.00	1.00	0.49	0.40	0.39
Denmark	1987	193,443	0.58	193,443	1.00	1.00	0.59	190,429	0.50	**189,690**	1.00	1.00	0.42	0.34	0.34
	2000	197,696	0.63	197,696	1.00	1.00	0.63	194,134	0.50	**194,134**	1.00	1.00	0.50	0.35	0.35
	2004	196,602	0.61	196,602	1.00	1.00	0.61	193,348	0.50	**193,348**	1.00	1.00	0.50	0.36	0.35
	2007	200,008	0.60	200,008	1.00	1.00	0.59	196,871	0.50	**196,871**	1.00	1.00	0.50	0.35	0.34
	2010	193,151	0.62	193,151	1.00	1.00	0.62	189,990	0.50	**189,893**	1.00	1.00	0.50	0.37	0.36

(continued)

Table 1 (continued)

Country	Year	Log-Normal		ContLog-Normal				Gamma		ContGamma				Usual Gini	
		dAIC	Gini	dAIC	α	η	Gini	dAIC	Gini	dAIC	α	η	Gini	Unwins.	Winsorized
Dominican republic	2007	223,869	0.58	223,869	1.00	1.00	0.58	224,465	0.50	**223,151**	0.86	5.83	0.55	0.54	0.52
Finland	2000	192,263	0.68	192,263	1.00	1.00	0.69	188,166	0.47	**188,124**	1.00	1.00	0.47	0.39	0.38
France	1978	146,301	0.37	146,104	0.96	3.52	0.40	146,214	0.34	**145,572**	0.95	6.60	0.35	0.34	0.33
	2000	170,056	0.46	170,056	1.00	1.00	0.45	168,833	0.38	**168,753**	0.97	3.60	0.39	0.37	0.36
Germany	1989	156,646	0.42	156,489	0.97	3.60	0.46	155,327	0.35	**154,926**	0.99	9.44	0.35	0.32	0.31
	2000	165,797	0.51	165,797	1.00	1.00	0.51	163,757	0.40	**163,661**	1.00	1.00	0.40	0.37	0.36
Greece	1995	205,821	0.41	205,821	1.00	1.00	0.41	205,399	0.37	**205,198**	0.93	3.83	0.37	0.37	0.36
	2000	208,194	0.42	208,194	1.00	1.00	0.42	207,552	0.36	**207,342**	0.96	4.55	0.36	0.36	0.35
Hungary	1991	216,629	0.63	216,629	1.00	1.00	0.62	214,684	0.49	**214,627**	0.98	4.16	0.49	0.46	0.45
	1994	201,885	0.70	201,885	1.00	1.00	0.70	199,968	0.50	**199,856**	1.00	1.00	0.51	0.50	0.49
	2005	207,526	0.70	207,526	1.00	1.00	0.71	204,528	0.50	**204,417**	1.00	1.00	0.50	0.45	0.44
	2007	122,453	0.59	122,453	1.00	1.00	0.60	121,213	0.46	**121,102**	0.98	5.82	0.47	0.41	0.40
	2009	123,459	0.55	123,459	1.00	1.00	0.55	122,356	0.43	**122,327**	0.99	4.72	0.44	0.42	0.41
	2012	128,489	0.61	128,489	1.00	1.00	0.60	127,174	0.46	**127,126**	1.00	1.00	0.47	0.44	0.43
Iceland	2004	276,877	0.54	276,877	1.00	1.00	0.54	**272,772**	0.39	272,772	1.00	1.00	0.40	0.34	0.33
	2007	280,660	0.51	280,016	0.98	5.49	0.70	277,035	0.38	**276,971**	0.99	5.01	0.38	0.34	0.33
	2010	274,337	0.59	274,337	1.00	1.00	0.59	**270,309**	0.43	270,309	1.00	1.00	0.42	0.37	0.36
Israel	2010	193,229	0.55	193,229	1.00	1.00	0.54	192,408	0.46	**191,844**	0.99	10.66	0.47	0.46	0.44
	2012	201,983	0.53	201,983	1.00	1.00	0.53	201,008	0.44	**200,792**	0.96	3.95	0.45	0.45	0.43
Italy	1986	248,748	0.35	248,681	0.98	3.36	0.37	248,273	0.32	**247,923**	0.98	7.41	0.32	0.31	0.30
	1987	237,540	0.35	237,455	0.96	2.86	0.35	237,524	0.32	**237,228**	0.50	4.20	0.33	0.33	0.32
	1989	241,199	0.32	241,109	0.94	2.62	0.32	241,327	0.30	**240,958**	0.89	4.18	0.30	0.31	0.30

(continued)

Table 1 (continued)

Country	Year	Log-Normal		ContLog-Normal				Gamma		ContGamma				Usual Gini	
		dAIC	Gini	dAIC	α	η	Gini	dAIC	Gini	dAIC	α	η	Gini	Unwins.	Winsorized
	1991	242,362	0.31	242,295	0.99	3.75	0.32	242,559	0.29	**242,098**	0.98	8.92	0.30	0.30	0.29
	1993	236,497	0.39	236,497	1.00	1.00	0.40	235,923	0.35	**235,792**	0.95	3.62	0.35	0.34	0.34
	1995	233,918	0.40	233,918	1.00	1.00	0.40	233,355	0.35	**233,136**	0.97	5.40	0.35	0.34	0.33
	1998	227,808	0.40	227,738	0.98	3.26	0.43	227,569	0.36	**227,034**	0.96	6.76	0.36	0.35	0.33
	2000	232,659	0.38	232,612	0.98	3.01	0.39	232,347	0.34	**232,006**	0.94	4.94	0.34	0.33	0.32
Japan	2008	258,682	0.44	258,682	1.00	1.00	0.44	257,416	0.37	**257,249**	0.99	5.80	0.37	0.36	0.35
Luxembourg	1985	200,293	0.36	198,128	0.98	12.15	0.56	197,853	0.29	**197,232**	0.85	5.32	0.28	0.26	0.26
	1991	209,589	0.32	209,191	0.97	4.96	0.34	208,734	0.29	**208,494**	0.69	3.23	0.28	0.28	0.27
	1994	209,988	0.36	209,453	0.97	5.55	0.39	208,652	0.31	**208,454**	0.59	3.21	0.31	0.29	0.29
	1997	209,938	0.43	209,451	0.98	5.54	0.50	207,785	0.35	**207,754**	0.72	2.14	0.35	0.32	0.32
	2000	216,834	0.45	216,406	0.97	5.22	0.55	214,513	0.36	**214,506**	0.77	1.79	0.36	0.33	0.33
Mexico	1984	252,758	0.51	252,758	1.00	1.00	0.51	252,624	0.44	**251,917**	0.88	4.37	0.46	0.46	0.45
	1989	308,886	0.50	308,811	0.99	2.67	0.52	310,402	0.46	**308,610**	0.89	6.93	0.50	0.50	0.47
	1992	323,570	0.56	323,140	1.00	1.00	0.98	324,454	0.49	**322,459**	0.87	7.25	0.54	0.53	0.50
Netherlands	1999	142,792	0.44	142,421	0.98	5.36	0.53	140,734	0.35	**140,660**	0.97	3.63	0.35	0.31	0.30
Norway	1986	204,698	0.56	204,698	1.00	1.00	0.55	**200,623**	0.40	200,623	1.00	1.00	0.40	0.32	0.32
	1991	204,297	0.61	204,297	1.00	1.00	0.61	**200,164**	0.43	200,164	1.00	1.00	0.43	0.35	0.34
	2000	207,165	0.57	207,165	1.00	1.00	0.57	204,521	0.50	**204,521**	1.00	1.00	0.50	0.35	0.34
	2004	212,068	0.62	212,068	1.00	1.00	0.63	208,962	0.50	**208,962**	1.00	1.00	0.50	0.39	0.38
	2007	216,163	0.60	216,163	1.00	1.00	0.61	212,995	0.45	**212,919**	1.00	1.00	0.45	0.38	0.37
	2010	217,252	0.61	217,252	1.00	1.00	0.62	214,370	0.50	**214,370**	1.00	1.00	0.50	0.39	0.38
Poland	1992	296,922	0.61	296,922	1.00	1.00	0.61	**293,195**	0.44	293,195	1.00	1.00	0.44	0.38	0.37

(continued)

Table 1 (continued)

Country	Year	Log-Normal		ContLog-Normal				Gamma		ContGamma				Usual Gini	
		dAIC	Gini	dAIC	α	η	Gini	dAIC	Gini	dAIC	α	η	Gini	Unwins.	Winsorized
Romania	1995	313,234	0.55	313,234	1.00	1.00	0.55	311,487	0.45	**311,109**	1.00	1.00	0.45	0.43	0.41
	1997	335,822	0.54	335,822	1.00	1.00	0.54	333,786	0.44	**333,603**	1.00	1.00	0.44	0.42	0.41
Russia	2007	193,966	0.66	193,966	1.00	1.00	0.66	**191,307**	0.49	191,307	1.00	1.00	0.49	0.45	0.44
	2010	210,506	0.64	210,506	1.00	1.00	0.63	**207,819**	0.48	207,819	1.00	1.00	0.48	0.45	0.44
	2013	206,504	0.68	206,504	1.00	1.00	0.69	**203,481**	0.50	203,481	1.00	1.00	0.50	0.45	0.44
Serbia	2006	201,243	0.62	201,243	1.00	1.00	0.63	**198,634**	0.47	198,634	1.00	1.00	0.47	0.44	0.43
	2010	191,649	0.67	191,649	1.00	1.00	0.67	**189,239**	0.50	189,239	1.00	1.00	0.50	0.46	0.45
	2013	192,608	0.69	192,608	1.00	1.00	0.68	190,187	0.50	**190,054**	1.00	1.00	0.52	0.48	0.46
Slovak republic	1992	164,230	0.33	164,066	0.98	4.10	0.35	163,729	0.30	**163,477**	0.96	4.86	0.29	0.29	0.28
Slovenia	1997	268,552	0.65	268,552	1.00	1.00	0.65	**264,878**	0.50	264,878	1.00	1.00	0.50	0.39	0.39
	1999	263,488	0.65	263,488	1.00	1.00	0.65	**259,649**	0.50	259,649	1.00	1.00	0.50	0.38	0.38
	2004	264,562	0.69	264,562	1.00	1.00	0.69	**260,750**	0.50	260,750	1.00	1.00	0.50	0.41	0.40
South Korea	2006	330,212	0.67	330,212	1.00	1.00	0.68	**324,441**	0.46	324,441	1.00	1.00	0.45	0.38	0.37
Spain	1980	232,641	0.73	232,641	1.00	1.00	0.74	**225,810**	0.46	225,810	1.00	1.00	0.47	0.37	0.36
	1985	243,427	0.61	243,427	1.00	1.00	0.60	**240,087**	0.45	240,087	1.00	1.00	0.45	0.41	0.40
	1990	227,455	0.50	227,455	1.00	1.00	0.50	225,001	0.39	**224,853**	1.00	1.00	0.39	0.36	0.35
	1995	207,951	0.48	207,951	1.00	1.00	0.48	206,888	0.40	**206,833**	0.97	3.16	0.41	0.39	0.38
	2000	212,516	0.44	212,516	1.00	1.00	0.43	211,804	0.38	**211,565**	0.96	4.52	0.38	0.38	0.36

(continued)

Table 1 (continued)

Country	Year	Log-Normal		ContLog-Normal				Gamma		ContGamma				Usual Gini	
		dAIC	Gini	dAIC	α	η	Gini	dAIC	Gini	dAIC	α	η	Gini	Unwins.	Winsorized
Sweden	1992	192,502	0.68	192,502	1.00	1.00	0.68	**188,690**	0.50	188,690	1.00	1.00	0.50	0.39	0.39
	1995	186,763	0.69	186,763	1.00	1.00	0.71	**183,207**	0.48	183,207	1.00	1.00	0.48	0.40	0.40
	2005	198,841	0.67	198,841	1.00	1.00	0.66	**194,892**	0.46	194,892	1.00	1.00	0.47	0.38	0.37
Taiwan	1981	242,130	0.29	241,722	0.97	4.34	0.30	242,163	0.28	**241,411**	0.62	4.66	0.28	0.28	0.27
	1986	245,807	0.31	245,154	0.97	5.15	0.33	245,572	0.29	**244,608**	0.74	4.97	0.30	0.29	0.28
	1991	255,814	0.33	255,309	0.97	4.65	0.35	255,498	0.30	**254,866**	0.87	4.71	0.31	0.30	0.29
	1995	255,178	0.39	254,888	0.94	3.75	0.42	253,386	0.33	**253,072**	0.50	4.13	0.33	0.31	0.31
	1997	254,951	0.40	254,772	0.93	3.36	0.41	253,055	0.34	**252,770**	0.50	4.20	0.33	0.32	0.31
	2000	249,211	0.40	249,089	0.99	3.93	0.43	247,526	0.34	**247,324**	0.50	3.85	0.34	0.32	0.32
	2005	244,392	0.43	244,392	1.00	1.00	0.43	242,577	0.36	**242,525**	0.95	2.63	0.36	0.34	0.33
	2007	245,419	0.45	245,419	1.00	1.00	0.44	243,657	0.37	**243,595**	0.97	3.21	0.37	0.35	0.35
	2010	240,171	0.44	240,171	1.00	1.00	0.43	238,768	0.37	**238,577**	0.98	5.14	0.38	0.35	0.34
	2013	238,259	0.50	238,259	1.00	1.00	0.50	235,208	0.39	**235,092**	0.98	4.16	0.39	0.35	0.34
Uruguay	2007	195,730	0.54	195,730	1.00	1.00	0.54	195,852	0.47	**195,264**	0.84	4.13	0.49	0.50	0.48
	2010	202,700	0.51	202,700	1.00	1.00	0.51	202,732	0.45	**202,172**	0.87	4.17	0.47	0.46	0.45
	2013	211,869	0.48	211,869	1.00	1.00	0.47	211,394	0.41	**211,109**	0.92	3.75	0.42	0.42	0.41

coefficient. The last two columns show the usual Gini index computed with and without winsorizing 1% of upper and lower extreme values.

Among uncontaminated models, the gamma provides the best fit in 87% of the cases. However, the log-normal model provides better fits for the South-American countries considered, which are Colombia, Dominican Republic, Mexico and Uruguay. Note that, within a given country, the ordering of the models according to the dAIC usually holds over all, or most part of, the years considered.

The contaminated gamma model outperforms all the other models in 78% of the cases. In the remaining cases, which include countries that were part of the former Eastern Bloc (Czech Republic, Poland, Russia, Serbia, Slovenia) plus Iceland and South Korea, the gamma and the contaminated gamma models always obtain the same or very similar likelihoods, with slightly better dAIC values for the uncontaminated gamma because of its lower number of parameters.

The estimated level of inequality is strongly affected by the model chosen, due to differences in goodness of fit, particularly in the tails of the distribution. In this study, the Gini coefficients estimated by the lognormal model are always higher than those estimated by the gamma model. However, contaminated distributions provide estimates that are often slightly higher than their uncontaminated counterparts.

It is also worth noting that in most countries the Gini coefficients has increased over time, confirming the general tendency towards greater inequality described in the introduction. However, we must mention that we are dealing with gross income, and before public transfers. Inequality would have been lower if net incomes were analyzed, especially for countries with higher progressive tax structures and larger social programs.

7 Conclusions

The rising disparities in household earnings and their effects on economic growth [42], social cohesion [36], health and life expectancy [28] are a main concern, "the most important problem that we are facing now today" according to noted economist Robert Shiller [11].

One problem with measuring variations in income distribution inequality, over time and among different countries, is the sensitivity of usual inequality measures to extreme values. To deal with this issue, we proposed the use of contaminated lognormal and gamma models and we derived the formulations for computing the Gini index based on the model parameters.

An application to 101 different empirical income distributions, which encompassed 31 upper- and middle-income countries at different years, has been presented. According to the dAIC selection criterion, the contaminated gamma model outperformed all the other models in 78% of the cases. In the remaining 22%, mostly countries that were part of the former Eastern Bloc, the uncontaminated gamma distribution obtained slightly better dAIC values because of its lower number of parameters. The Gini coefficients, computed using the estimated distribution param-

eters resulted strongly affected by the model chosen, due to differences in goodness of fit, particularly in the tails of the distribution. The Gini coefficients estimated confirmed the general tendency towards greater inequality for almost all the countries considered.

Appendix

Expectation and Cumulative Distribution Function

Here, we explicit the expectation and the c.d.f. for the contaminated gamma and lognormal densities.

Contaminated Gamma Distribution

According to (5), as well as based on the standard results for the gamma distribution given in (4), the expectation for the unimodal gamma distribution in (3) is given by

$$E\left(X;\lambda,\nu\right)=\lambda+\nu,$$

while the corresponding c.d.f. is

$$H\left(x;\lambda,\nu\right)=\frac{\gamma\left(1+\frac{\lambda}{\nu},\frac{x}{\nu}\right)}{\Gamma\left(1+\frac{\lambda}{\nu}\right)},\quad x>0,$$

where $\gamma\left(\cdot,\cdot\right)$ is the lower incomplete gamma function. Therefore, based on (13), the expectation for the contaminated gamma density is

$$E\left(X;\boldsymbol{\vartheta}\right)=\lambda+\nu\left[\alpha+\left(1-\alpha\right)\eta\right],$$

while, based on (12), the c.d.f. is

$$F\left(x;\boldsymbol{\vartheta}\right)=\alpha\frac{\gamma\left(1+\frac{\lambda}{\nu},\frac{x}{\nu}\right)}{\Gamma\left(1+\frac{\lambda}{\nu}\right)}+\left(1-\alpha\right)\frac{\gamma\left(1+\frac{\lambda}{\eta\nu},\frac{x}{\eta\nu}\right)}{\Gamma\left(1+\frac{\lambda}{\eta\nu}\right)},\quad x>0.$$

Contaminated Lognormal Distribution

According to (9), as well as based on the standard results for the lognormal distribution given in (8), the expectation for the unimodal lognormal distribution in (7) is given by

$$E\left(X;\lambda,\nu\right)=\lambda e^{\frac{3}{2}\nu},$$

while the corresponding c.d.f. is

$$H\left(x;\lambda,\nu\right)=\frac{1}{2}+\frac{1}{2}\mathrm{erf}\left(\frac{\ln x-\ln\lambda-\nu}{\sqrt{2\nu}}\right),\quad x>0,$$

where $\mathrm{erf}\left(\cdot\right)$ is the error function. Therefore, based on (13), the expectation for the contaminated lognormal density is

$$E\left(X;\boldsymbol{\vartheta}\right)=\lambda\left(\alpha e^{\frac{3}{2}\nu}+\left(1-\alpha\right)e^{\frac{3}{2}\eta\nu}\right),$$

while, based on (12), the c.d.f. is

$$F\left(x;\boldsymbol{\vartheta}\right)=\frac{1}{2}+\frac{1}{2}\left[\alpha\mathrm{erf}\left(\frac{\ln x-\ln\lambda-\nu}{\sqrt{2\nu}}\right)+\left(1-\alpha\right)\mathrm{erf}\left(\frac{\ln x-\ln\lambda-\eta\nu}{\sqrt{2\eta\nu}}\right)\right],\ x>0.$$

References

1. Aitkin, M., Wilson, G.T.: Mixture models, outliers, and the EM algorithm. Technometrics **22**(3), 325–331 (1980)
2. Akaike, H.: Information theory and an extension of maximum likelihood principle. In: Petrov, B.N., Csaki, F. (eds.) Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)
3. Ammon, O.: Die Gesellschaftsordnung und ihre Natürlichen Grundlagen. Jena (1895)
4. Bagnato, L., Punzo, A.: Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. Comput. Stat. **28**(4), 1571–1597 (2013)
5. Bartels, C.P.A., van Metelen, H.: Alternative probability density functions of income. Research Memorandum 29, Vrije University Amsterdam (1975)
6. Berkane, M., Bentler, P.M.: Estimation of contamination parameters and identification of outliers in multivariate data. Sociol. Methods Res. **17**(1), 55–64 (1988)
7. Budd, E.C.: Distribution issues: trends and policies. Am. Econ. Rev. **60**(2), 247–260 (1970). Papers and Proceedings of the Eighty-second Annual Meeting of the American Economic Association
8. Chen, S.: Probability density function estimation using gamma kernels. Ann. Inst. Stat. Math. **52**(3), 471–480 (2000)
9. Chen, W.H., Förster, M., Llena-Nozal, A.: Demographic or labour market trends: what determines the distribution of household earnings in OECD countries? OECD J.: Econ. Stud. **2013**(1), 179–207 (2014)

10. Chotikapanich, D., Griffiths, W.E.: Estimating income distributions using a mixture of gamma densities. In: Chotikapanich, D. (ed.) Modeling Income Distributions and Lorenz Curves, Economic Studies in Inequality, Social Exclusion and Well-Being, chap. 16, pp. 285–302. Springer, New York (2008)
11. Christoffersen, J.: Rising inequality 'most important problem', says nobel-winning economist. St, Louis Post-Dispatch (2013)
12. Cowell, F.: Measuring Inequality. London School of Economics Perspectives in Economic Analysis, OUP Oxford (2011). https://books.google.it/books?id=0-V4wIGDxhIC
13. Cowell, F.A., Victoria-Feser, M.P.: Robustness properties of inequality measures. Econometrica **64**(1), 77–101 (1996)
14. Dagum, C.: A new model of personal income distribution: Specification and estimation. In: Chotikapanich, D. (ed.) Modeling Income Distributions and Lorenz Curves, Economic Studies in Equality, Social Exclusion and Well-Being, vol. 5, chap. 1, pp. 3–25. Springer, New York (2008)
15. Davies, L., Gather, U.: The identification of multiple outliers. J. Am. Stat. Assoc. **88**(423), 782–792 (1993)
16. Feser, M.P.V.: Robust estimation of personal income distribution models. Research Paper DARP/4, London School of Economics and Political Science (1993)
17. Fonseca, L., Tayman, J.: Postcensal estimates of household income distributions. Demography **26**(1), 149–159 (1989)
18. Forster, M., Chen, W., Llenanozal, A.: Divided We Stand: Why Inequality Keeps Rising. OECD (2011)
19. Gibrat, R.: Les Inégalités Économiques. Librairie du Recueil Sirey, Paris (1931)
20. Johnson, N.L., Kotz, S.: Continuous Univariate Distributions, vol. 1. Wiley, New York (1970)
21. Karoly, L.A., Burtless, G.: Demographic change, rising earnings inequality, and the distribution of personal well-being, 1959–1989. Demography **32**(3), 379–405 (1995)
22. Kleiber, C., Kotz, S.: Statistical Size Distributions in Economics and Actuarial Sciences, Wiley Series in Probability and Statistics, vol. 470. Wiley, New York (2003)
23. Lumley, T., Scott, A.: AIC and BIC for modeling with complex survey data. J. Surv. Stat. Methodol. **3**(1), 1–18 (2015)
24. Maruotti, A., Punzo, A.: Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. Comput. Stat. Data Anal. **113**, 475–496 (2017)
25. Mazza, A., Punzo, A.: Mixtures of multivariate contaminated normal regression models. Stat. Pap. (2017). https://doi.org/10.1007/s00362-017-0964-y
26. McDonald, J.B., Ransom, M.: The generalized beta distribution as a model for the distribution of income: Estimation of related measures of inequality. In: Chotikapanich, D. (ed.) Modeling Income Distributions and Lorenz Curves, Economic Studies in Equality, Social Exclusion and Well-Being, vol. 5, chap. 8, pp. 147–166. Springer, New York (2008)
27. Morris, K., Punzo, A., McNicholas, P.D., Browne, R.P.: Asymmetric clusters and outliers: mixtures of multivariate contaminated shifted asymmetric laplace distributions. Comput. Stat. Data Anal. **132**, 145–166 (2019)
28. Pickett, K.E., Wilkinson, R.G.: Income inequality and health: a causal review. Soc. Sci. Med. **128**, 316–326 (2015)
29. Punzo, A.: A new look at the inverse Gaussian distribution with applications to insurance and economic data. J. Appl. Stat. **46**(7), 1260–1287 (2019)
30. Punzo, A., Bagnato, L., Maruotti, A.: Compound unimodal distributions for insurance losses. Insur.: Math. Econ. **81**, 95–107 (2018)
31. Punzo, A., Maruotti, A.: Clustering multivariate longitudinal observations: the contaminated Gaussian hidden Markov model. J. Comput. Graph. Stat. **25**(4), 1097–1116 (2016)
32. Punzo, A., Mazza, A., Maruotti, A.: Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions. J. Appl. Stat. **45**(14), 2563–2584 (2018)
33. Punzo, A., Mazza, A., McNicholas, P.D.: ContaminatedMixt: an R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. J. Stat. Softw. **85**(10), 1–25 (2018)

34. Punzo, A., McNicholas, P.D.: Parsimonious mixtures of multivariate contaminated normal distributions. Biom. J. **58**(6), 1506–1537 (2016)
35. Punzo, A., McNicholas, P.D.: Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. J. Classif. **34**(2), 249–293 (2017)
36. Putnam, R.D.: Bowling alone: the collapse and revival of American community. Simon and Schuster (2001)
37. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). https://www.R-project.org/
38. Raftery, A.E.: Bayesian model selection in social research. Sociol. Methodol. **25**, 111–164 (1995)
39. Salem, A.B.Z., Mount, T.D.: A convenient descriptive model of income distribution: the gamma density. Econometrica **42**(6), 1115–1127 (1974)
40. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
41. Skinner, C.J., Holt, D., Smith, T.M.F.: Analysis of Complex Surveys. Wiley Series in Probability and Mathematical Statistics. Wiley (1989)
42. Stiglitz, J.: The global crisis, social protection and jobs. Int. Labour Rev. **152**(s1), 93–106 (2013)
43. Takeuchi, K.: Distribution of informational statistics and a criterion of model fitting. Suri-Kagaku (Mathematical Sciences) **153**(1), 12–18 (1976)
44. Tomarchio, S.D., Punzo, A.: Heavy-tailed unimodal dichotomous compound models for the distribution of insurance losses. J. Appl. Stat. (2019). https://doi.org/10.1111/rssa.12466
45. Van Praag, B., Hagenaars, A., Van Eck, W.: The influence of classification and observation errors on the measurement of income inequality. Econometrica **51**(4), 1093–1108 (1983)

Endowments and Rewards in the Labour Market: Their Role in Changing Wage Inequality in Europe

Gennaro Punzo, Mariateresa Ciommi, Gaetano Musella and Rosalia Castellano

Abstract This paper proposes a comparative analysis on how the recent structural changes in the workforce composition affect wage inequality in a set of European countries. By performing RIF regression on the EU-SILC data, we assess how much of the overall Gini gap between 2005 and 2013 is due to employees' characteristics rather than the capability of each country's labour market to capitalise skills. The outright deterioration of all jobs, irrespective of skill levels required, and the lack of a well-defined structure of labour market may jeopardise wage distribution, and the wage structure plays a leading role in this process.

Keywords Wage inequality · Employment structure · Job polarisation · Upgrading of occupations · RIF regression · Europe

1 Introduction

A basic prerequisite of the Kuznets theory holds that inequality tends to decline with the economic progress [22]. Substantial changes in global macroeconomic environment might create a general inequality climate for both developed and developing

G. Punzo (✉) · G. Musella
Department of Economic and Legal Studies, University of Naples Parthenope, Naples, Italy
e-mail: gennaro.punzo@uniparthenope.it

G. Musella
e-mail: gaetano.musella@uniparthenope.it

M. Ciommi
Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy
e-mail: m.ciommi@univpm.it

R. Castellano
Department of Management and Quantitative Studies, University of Naples Parthenope, Naples, Italy
e-mail: lia.castellano@uniparthenope.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_29

countries [14]. For instance, the US income distribution suffered a hard shock during the Great Depression of the 1930s and the Second World War (1939−1945) with permanent fallouts in the years ahead. The US income inequality was still comparatively high in the 1970s and continued to grow until the USA reached the top of the rich country inequality pyramid [21].

The ongoing global crisis—the worst since 1930—has produced painful effects for most Europe, especially for countries with weaker economies. As detailed by Eurostat (on-line database), the Eurozone unemployment increased from 7.5% to 11.3% in 2007−2013, and Mediterranean and Central/Eastern European countries were affected by unemployment more severely [25].

To address these emergencies, at least three of the five general goals of the Europe 2020 strategy for smart, sustainable and inclusive growth directly relate to employment, productivity, and inequality. With the purpose of reaching the employment rate of 75% for 20−64-year-olds, increasing at least 40% of 30−34-year-olds completing tertiary education and lifting 20 million people out of poverty by 2020, the strategy focuses on the target of *new skills for new jobs*, taking the headline idea of *more and better jobs* from the earlier Lisbon agenda.

However, within the same country, workers with varying levels of skills suffered at different extent and intensity. As argued by Eurofound [8], the relatively recent trends identified major declines in the demand for jobs in the middle of the skills hierarchy. This resulted in structural shifts in the composition of labour force that give rise to varying labour market outcomes and income inequality trajectories [4]. Therefore, changes in income inequality might be contextualised in the structure of the countries' labour markets in terms of job polarisation, upgrading, and more rarely, downgrading of occupations [4, 16].

Specifically, job polarisation consists of a relative expansion in the demand of jobs occupying the top and bottom of the skills hierarchy and shrinking in jobs in the middle. The upgrading of occupations favours high-qualified activities with respect to low- and middle-skill jobs [1, 17], whereas the downgrading occurs if low-skill jobs grow faster than the rest [18].

On this ground, this work aims at identifying regularities in the structural shifts in the labour market comparatively for ten European countries and their potential relationships with the changes in wage distribution. Borrowing the geographical classification by Nolan et al. [25], which approximately corresponds to the standard welfare regimes typology [7], the following countries are selected:

1. The *Big Three* of Europe: France, Germany, and the United Kingdom.
2. The four Mediterranean countries: Italy, Greece, Portugal, and Spain.
3. Three Central/Eastern countries: the Czech Republic, Hungary, and Poland.

Using the favours Influence Function (RIF) regression [10, 11] we: (i) explore the gaps in wage inequality between 2005 and 2013 for each country covered and decompose them into the composition and wage structure effects (aggregate decomposition); (ii) evaluate the contribution that each primary force of wage inequality gives to both components (detailed decomposition). The choice of 2005 and 2013 as the reference years allows us to obtain clues about the socio-economic scenarios

that foreshadowed the global crisis and their role in affecting the structure of labour markets and patterns of wage inequality.

The paper is structured as follows. Section 2 addresses the methodology of RIF regression and the data used to perform the analysis. In Sect. 3, the two components of Gini gaps are discussed in light of the main country's labour market transformations. Concluding remarks are presented in Sect. 4.

2 Methodology and Data

RIF regression of Gini on (log of) gross individual wage replaces the log-wage as the dependent variable with the recentered influence function of the Gini coefficient $v(F)$ and directly estimates the impact of covariates on Gini [10, 11]. Therefore, the RIF method includes a preliminary step in which a set of covariates are tested as potential determinants of the observed wage inequality by country. Explanatory variables are grouped in individual characteristics (gender, couple, health), human capital (experience, education), job background (type of contract, economic status), and occupation type variables. In doing so, RIF regression allows the evaluation of those factors that are quantitatively more significant to make inequality gaps over time as well as their contribution in shaping the two components (composition vs. wage structure effects) in which the overall Gini change in 2005−2013 of each country is decomposed.

Data are from the European Union-Survey on Income and Living Conditions (EU-SILC), which is the primary reference source for comparable socio-economic statistics in Europe. Moving from the assumption that inequality starts in the labour market, changes in wage distributions become the key factors behind inequality trends. For this reason, our analysis focuses on employees, aged 16−64, irrespective of their activity sector, excluding those employed in military occupations. They are classified in the three distinct groups of high-, middle- and low-skilled employees based on the level of expertise required to perform their specific job. Given the strong correlation between the current average education level and skills required to perform that job [9], the average level of education is selected as a measure of the skills needed.

RIF regression is well suited to the objective of this paper because it can obtain the decomposition of Gini (or also for median, quantile, and variance), whereas the Oaxaca-Blinder (OB) method enables the decomposition to be applied only to the mean [11]. RIF regression overcomes other two limitations of the OB method [3, 26]: (i) the estimations of composition and wage structure effects can be misleading if the linear model is unspecified [2], (ii) the contribution of each covariate to wage structure is highly sensitive to the choice of the base group [15, 27]. The Juhn, Murphy and Pierce method [19, 20] and the quantile-based decomposition by Machado and Mata [23] already removed these disadvantages, but they are unable to trace the contribution provided by each covariate to the composition effect when they are used to decompose various distributional statistics [10].

The observed wage (Y_i) can be written without imposing a specific functional form considering the wage determination function of observed components X_i and some unobserved components ϵ_i :

$$Y_{gi} = f_g(X_i, \epsilon_i) \quad for \quad g = 0, 1 \tag{1}$$

$g = 1$ for workers observed in group 1 and $g = 0$ for those in group 0. In this work, the two groups are composed of employees at time 2005 and 2013.

Let $v(F_y)$ be the generic distributional statistic to study (in this work, Gini), the first-order directional derivative is known as its influence function $F(y, v(F_y))$ so that it measures the relative effect of a small change in the underlying outcome distribution on the statistic of interest. The RIF is:

$$RIF(y; v(F_y)) = IF(y; v(F_y)) + v(F_y) \tag{2}$$

As regards the Gini coefficient, the distributional statistic $v(F_y)$ is defined as:

$$v^{GC}(F_Y) = 1 - 2\mu^{-1}R(F_Y) \tag{3}$$

where $R(F_Y) = \int_0^1 GL(p(y); F_Y)dp$ with $p(y) = F_Y(y)$ and the Generalised Lorenz ordinate of F_Y is given by $GL(p(y); F_Y) = \int_{-\infty}^{F^{-1}(p)} zdF_Y(z)$.

Following Firpo et al. [10], the recentered influence function of Gini becomes:

$$RIF(y; v^{GC}) = 1 + 2\mu^{-2}R(F_y) - 2\mu^{-1}\left[y\left[1 - p(y)\right] + GL(p(y); F_y)\right] \tag{4}$$

The key term for decomposing v^{GC} is the counterfactual distributional statistic v_c^{GC}, which is the distributional statistic that would have prevailed if workers observed in group 1 had the wage structure of period 0. Using the counterfactual distribution, the decomposition of Gini gap between the periods 0 and 1 is:

$$\widehat{\Delta}_0^{v^{GC}} = \widehat{\Delta}_S^{v^{GC}} + \widehat{\Delta}_X^{v^{GC}} = \overline{X}_1\left(\widehat{\gamma}_{1,v^{GC}} - \widehat{\gamma}_{0,v^{GC}}^C\right) + \left(\overline{X}_0^C - \overline{X}_0\right)\widehat{\gamma}_{0,v^{GC}} \tag{5}$$

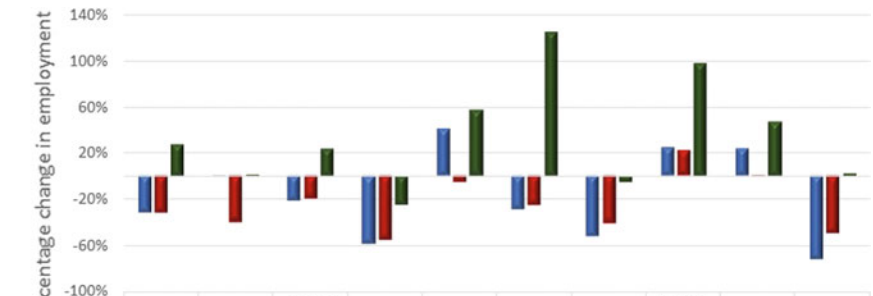
Therefore, the overall inequality gap $\left(\widehat{\Delta}_0^{v^{GC}}\right)$ is decomposed into the wage structure $\left(\widehat{\Delta}_S^{v^{GC}}\right)$ and the composition effects $\left(\widehat{\Delta}_X^{v^{GC}}\right)$. The first term corresponds to the effect on v^{GC} of a change from $f_1(\cdot, \cdot)$ to $f_0(\cdot, \cdot)$ while keeping the distribution of $(X, \epsilon)\,|G = 1$ constant. Conversely, the composition effect keeps the wage structure effect $f_0(\cdot, \cdot)$ constant and measures the effect of changes from $(X, \epsilon)\,|G = 1$ to $(X, \epsilon)\,|G = 0$. The estimation of the coefficients of each group $\left(\widehat{\gamma}_{g,v^{GC}},\ g = 0, 1\right)$ and those of the counterfactual distributions $\left(\widehat{\gamma}_{0,v^{GC}}^C\right)$ requires first estimating the weighting functions $\omega_1(G)$, $\omega_0(G)$ and $\omega_C(G, X)$. Further methodological details

on the estimation of γ parameters, on the weighting procedure and on the contribution of a single covariate to the decomposition can be found in DiNardo et al. [5], Firpo et al. [10], and Fortin et al. [11].

3 Discussing Wage Inequality in Light of the Structural Changes

This section discusses our empirical results focusing on the evolution of wage inequality in 2005−2013—i.e., the magnitude of the Gini gaps, the components in which it can be decomposed, and the factors that mostly contribute in shaping these components—in light of the varying patterns that foreshadowed in each country's labour market. Figure 1 shows the percentage changes in employment shares between 2005 and 2013 for each of the three groups of employees by skill level. The results allow the countries to be classified according to the patterns of the labour market in terms of job polarisation, upgrading of occupations or neither of the two.

Once the RIF regression of Gini on log-wage have been estimated on the above set of covariates by country, the overall Gini differences in 2005−2013 are decomposed into the composition effect and wage structure (Tables 1, 3 and 5). The former assesses the share of Gini changes attributable to personal characteristics. The latter explores the capability of the country's labour market to transform personal skills into job opportunities and earnings and explains why employees are rewarded differently for the same personal endowments. Standard errors of components are computed according to the method detailed in Fortin et al. [11].



	France	Germany	United Kingdom	Italy	Greece	Portugal	Spain	Czech Republic	Hungary	Poland
Low-Skilled	-31.74%	-1.08%	-21.16%	-57.74%	40.48%	-28.55%	-50.81%	24.85%	23.49%	-70.59%
Middle-Skilled	-31.94%	-39.36%	-19.21%	-54.48%	-5.71%	-24.80%	-39.65%	21.46%	-0.50%	-48.14%
High-Skilled	27.17%	0.76%	23.35%	-25.33%	56.96%	124.94%	-5.69%	97.73%	46.78%	1.86%

Fig. 1 Percentage changes in employment shares by skill levels and by country. 2005−2013 *Source* Authors' elaboration on EU-SILC data

3.1 *The* Big Three *of Europe: France, Germany, and the UK*

As shown in Fig. 1, the recent structural changes in the employment composition allow the French and British labour markets to be configured as upgrading of occupations. They both share the growth in professions that demand high skills and the simultaneous contraction in the demand of low- and middle-skill activities. In particular, in France, low- and middle-skill jobs have decreased by 32%, whereas the share of high-skill employees has increased by 27%. The United Kingdom follows similar trends albeit with less intensity with respect to France. Instead, in Germany, middle-skill jobs have declined as a share of employment by about 40% with slightly increasing levels of high-skill occupations. Accordingly, the pattern of the German labour market may be classified as relatively polarised.

Based on our results (Table 1), the overall Gini has declined between 2005 and 2013 in France and Germany in line with the literature [6, 13] that argues how the overall inequality was rather stable in Germany during the 1980s, increased after reunification, especially in 2000−2005, and slightly decreased afterwards [4]. Similarly, Frémeaux and Piketty [12] stress how in France inequality among full-time employees decreased during the 1980s and 1990s and has been stable since then. The high minimum wage, which has continued to increase from 1980 to 2010, has surely helped reduce income inequality in France. Conversely, the United Kingdom shows a rise in the overall inequality in 2005−2013, and a more detailed analysis by Machin [24] demonstrates how the growth inequality has been concentrated in the upper part of the distribution since the 2000s.

In both France and Germany, a great deal of total changes in Gini index is due to the composition effect: up to more than 90% for Germany, where the wage structure is even not significant, and more than one-half for France. In particular, inside the composition effect, being a more skilled employee (e.g., teaching professional, technicians, and small enterprise managers) with a permanent/full time contract contributes in reducing wage inequality over time in both countries (Table 2). Instead, the wage structure plays an exclusive role in the United Kingdom in increasing wage inequality, stressing the low capacity of the country's labour market to transform inputs into less unequal job-related careers and earnings.

Table 1 Aggregate RIF decomposition of Gini on log-wage. The Big Three

Gap 2005−2013	France		Germany		The UK	
Total gap	−0.0041***	–	−0.0043***	–	0.0011*	–
Composition	−0.0021***	51.2%	−0.0039***	90.7%	−0.0004	−36.4%
Wage structure	−0.0020***	48.8%	−0.0004	9.3%	0.0015**	136.4%

*Significant at 10%; **Significant at 5%; ***Significant at 1%.
Source Authors' elaboration on EU-SILC data

Table 2 Detailed RIF decomposition of Gini. Gap 2005–2013. The Big Three

Variables	France		Germany		The UK	
	CE	WS	CE	WS	CE	WS
Personal characteristics						
Gender (1 if male)	−0.00001	0.00028	−0.00017***	−0.00408***	0.00008**	−0.00069
Couple (ref. married)	−0.00024***	0.00034	0.00006	−0.00283***	0.00004	−0.00151**
Health (1 if no suffer)	0.00008	0.00029	0.00001***	−0.00221*	0.00010	−0.0040***
Human capital						
Experience (years)	−0.00067***	−0.01045***	−0.00075	0.0116***	−0.00018	0.00167
Education (ref. high)						
–Medium	−0.0000	0.00038	−0.00023**	−0.00109	0.00039***	−0.0023***
–Low	0.0000	0.00057*	0.00046***	−0.00057*	−0.00018	−0.00003
Job background						
Contract (1 permanent)	−0.00119***	0.00178	−0.00091***	0.00181	−0.00005	0.01468***
Status (1 if full)	−0.00116***	0.00363**	−0.00151***	−0.00106	0.00065***	0.00173
Occupation (elementary)						
High-skill						
Corporate managers	0.00009**	−0.00062**	0.00015***	0.0003***	−0.00018	−0.00045
Professionals	0.00005	−0.00075***	0.00129***	0.00085**	−0.00061***	−0.00054**
Teaching professionals	−0.00022***	−0.00072**	−0.00083***	−0.00035	−0.00016**	−0.00034*
Technicians	−0.00137***	−0.00135**	−0.00436***	0.00168**	−0.00060***	−0.00003
Middle-skill						
Small enterprise managers	−0.00010**	0.00015	−0.00078***	0.00055**	0.00007	0.00018**
Clerks	0.00207***	−0.00122**	0.00587***	0.00122*	0.00161***	−0.00091
Service workers	−0.00013**	−0.00067	−0.00072***	0.0001	0.0000	−0.00089**
Low-skill						
Agricultural workers	0.00004	−0.00003	0.00020***	−0.0000	−0.0000	0.00004
Machine operators	0.00065***	−0.00042*	−0.00164***	0.0016***	−0.00004	−0.0001
Constant	–	−0.01406**	–	−0.00793*	–	−0.00504

CE = Composition Effect; WS = Wage Structure
*Significant at 10%; **Significant at 5%; ***Significant at 1%
Source Authors' elaboration on EU-SILC data

3.2 *Mediterranean Countries: Italy, Greece, Portugal, and Spain*

Based on our results (Fig. 1), the Portuguese labour market is mostly characterised by the upgrading of occupations. To forehead of a reduction in low- and middle-skill jobs of about one-quarter, Portugal shows the largest proliferation in high-skill activities. Greece has seen a large increase in the share of low- and high-skill jobs (+40% and 57%, respectively) and the shrinkage in middle-skill occupations by 6%. Accordingly, if the pattern of the German labour market is relatively polarised, the Greek labour market may be considered as purely polarised.

As regards Italy and Spain, the structural changes in employment composition between 2005 and 2013 do not enable us to define whether one pattern prevails over the other. In fact, the joint contraction in low-, middle- and high-skill jobs cannot be related to polarisation or to upgrading of occupations. The strong deterioration in the employment structures of both countries—which is even more severe for Italy—gives rise to hybrid patterns of their labour markets. However, Italian and Spanish high-skilled employees suffer relatively smaller declines than their low- and middle-skill counterparts.

Italy (that together with Germany and France forms the *Big Four* of Europe) shows a rise in the overall inequality (Table 3), which is far larger than those of the United Kingdom. This is consistent with the literature [25] that classifies Italy as unequal country more than any other European nations with similar economic growth, but relatively less unequal than any other Mediterranean countries. In fact, Greece still keeps harsher levels of inequality despite the Gini index has increased in 2005−2013 less than in Italy. Gini has also largely increased in Spain while its change is not significant for Portugal in line with the literature [25] that shows a reversal of the previous increase in income inequality since 2005, which has not been large enough to compensate for the strong inequality growth in 1989−1994.

Similarly to what was happening in the United Kingdom, a great deal of the increase in wage inequality in Italy and Spain is due to the wage structure: up to more than 90% for Italy (the composition effect is even not significant). In Spain, the wage structure completely captures the increase in Gini index while the composition effect does not help mitigate this growth (Table 3). As shown in Table 4, the role of wage structure in increasing wage inequality in these countries is explained by

Table 3 Aggregate RIF decomposition of Gini on log-wage. Mediterranean countries

Gap 2005−2013	Italy		Greece		Portugal	Spain	
Total gap	0.0064***		0.0037***		−0.0001	0.0101***	
Composition	0.0004	6.3%	0.0019***	52.6%	0.0041	−0.0021***	−20.8%
Wage structure	0.0060***	93.7%	0.0017***	47.4%	−0.0042	0.0122***	120.8%

*Significant at 10%; **Significant at 5%; ***Significant at 1%

Source Authors' elaboration on EU-SILC data

Table 4 Detailed RIF decomposition of Gini. Gap 2005−2013. Mediterranean countries

Variables	Italy		Greece		Spain	
	CE	WS	CE	WS	CE	WS
Personal characteristics						
Gender (1 if male)	−0.00009***	0.00084*	−0.00015**	0.00350***	−0.00015*	0.00294***
Couple (ref. married)	0.00005**	0.00003	0.00012	−0.0027***	0.00022***	−0.00147**
Health (1 if no suffer)	0.00004	−0.00026	0.00000***	0.00029	−0.00014*	−0.00149
Human capital						
Experience (years)	−0.00064***	−0.0090***	−0.00181	−0.0198***	−0.00120***	−0.0088***
Education (ref. high)						
–Medium	−0.00015***	0.00199***	−0.00004	−0.00048	−0.00000	−0.00019
–Low	0.00027**	0.00177***	0.00109***	−0.00098	0.00004	0.0008
Job background						
Contract (1 permanent)	−0.00023**	−0.0098***	0.00087***	−0.0063***	−0.00234***	−0.0178***
Status (1 if full)	0.00108***	0.00279***	0.00191***	−0.00056	0.00165***	0.00842***
Occupation (elementary)						
High-skill						
Corporate managers	−0.00031***	−0.00009	−0.00000	−0.00024	−0.00005	−0.0004***
Professionals	−0.00010**	−0.0004***	−0.00017	−0.00021	−0.00014**	−0.0013***
Teaching professionals	−0.00072***	0.00002	−0.00033**	0.00014	−0.00042***	−0.0006***
Technicians	−0.00057***	−0.00007	−0.00034**	0.00031	−0.00098***	−0.00026**
Middle-skill						
Small enterprise managers	−0.00007*	0.00035***	−0.00025	0.00047***	0.00005*	0.00005*
Clerks	0.00144***	−0.0020***	0.00114***	0.00025	0.00135***	−0.0032***
Service workers	−0.00047***	−0.0005***	−0.00033**	−0.00014	−0.00045***	−0.0019***
Low-skill						
Agricultural workers	0.00000	0.00003	−0.00002	0.00001	−0.00010**	−0.00009
Machine operators	0.00086***	−0.0007***	0.00022**	0.00022	0.00056***	−0.0012***
Constant	–	0.02100***	–	0.02803***	–	0.03862***

CE = Composition Effect; WS = Wage Structure
*Significant at 10%; **Significant at 5%; ***Significant at 1%
Source Authors' elaboration on EU-SILC data

Table 5 Aggregate RIF decomposition of Gini on log-wage. Central/Eastern countries

Gap 2005−2013	Czech Republic		Hungary		Poland	
Total gap	−0.0060***	–	−0.0119***	–	−0.0176***	–
Composition	−0.0043***	71.47%	−0.0077***	64.71%	−0.0040***	22.73%
Wage structure	−0.0017***	28.73%	−0.0042***	35.29%	−0.0136***	77.27%

*Significant at 10%; **Significant at 5%; ***Significant at 1%
Source Authors' elaboration on EU-SILC data

personal characteristics and contract type (i.e., being man and working part-time), while having a high education is crucial to reducing wage inequality in Italy.

3.3 Central/Eastern Countries: Czech Republic, Hungary, Poland

Figure 1 shows that in Poland the drastic reduction in the demand for low- (−71%) and middle-skill (−48%) jobs is opposed only a slow-growing in highly specialised jobs (+2%). One specific point deserves the Czech Republic where there has been the growth in all jobs regardless of the level of skills required, and surprisingly, the demand for high-skill jobs has practically doubled (+98%). These structural changes in the Polish and Czech labour markets provide evidence of two patterns that can potentially evolve in the future but, at present, are relatively upgraded.

While the strong decline in middle-skill jobs in Germany is associated to a slight increase in high-skill activities, in Hungary the small decrease in middle-skill jobs goes together with an important expansion in jobs at the high (+47%) and low (+23%) end of the skill spectrum. Accordingly, the Hungarian labour market is purely polarised in the same manner as the Greek labour market.

The overall Gini has declined over time in each Central/Eastern country covered (Table 5) and the magnitude of the fall has been more pronounced than that of Germany and France. These countries experienced difficult times (began in 1989 with the transition from the *command* economy to a more market-based system), which also generated great divergences in their inequality levels. In fact, while the Gini growth was even higher than 10 points for Hungary, it was less severe for Czech Republic and Poland [25].

The composition effect mostly explains the overall decrease in wage inequality between 2005 and 2013 (just under the three-quarters for Czech Republic and two-thirds for Hungary). Focusing on the contribution of each covariate to the composition effect (Table 6), human capital (work experience), job background (permanent and/or full time contracts) and high-skill jobs are the main driving forces for the reduction

Table 6 Detailed RIF decomposition of Gini. Gap 2005−2013. Central/Eastern countries

Variables	Czech Republic		Hungary		Poland	
	CE	WS	CE	WS	CE	WS
Personal characteristics						
Gender (1 if male)	0.00000	0.00025	0.00009*	−0.00024	0.00000	0.0021***
Couple (ref. married)	−0.00021**	−0.00005	−0.00002	−0.00069	−0.00001	−0.00139***
Health (1 if no suffer)	0.00004	−0.00069	−0.00019	0.00047	−0.00007	0.00129
Human capital						
Experience (years)	−0.00024**	−0.00252	−0.00088***	−0.0031	−0.00028**	0.0101***
Education (ref. high)						
–Medium	0.00039***	0.00439**	−0.00006	0.0072***	0.00033***	0.00003
–Low	0.00011	0.00001	0.00092***	0.00088**	−0.00024***	−0.00052***
Job background						
Contract (1 permanent)	−0.00035***	−0.00519***	−0.00193***	0.00056	0.00030**	0.0064***
Status (1 if full)	−0.00365***	0.0140***	−0.00370***	0.00606**	−0.00426***	0.0085***
Occupation (elementary)						
High-skill						
Corporate managers	−0.00004	−0.00017	−0.00060**	−0.00004	−0.00009***	−0.00022
Professionals	−0.00021**	0.0001	−0.00002	0.00003	0.00002	−0.00063*
Teaching professionals	−0.00068***	−0.00014	−0.00121***	0.00103**	−0.00003	−0.00042
Technicians	−0.00064***	−0.00091	−0.00069***	−0.00056	−0.00004	−0.00069*
Middle-skill						
Small enterprise managers	0.00001	−0.0002	−0.00004	−0.00002	0.00025	−0.00030**
Clerks	0.00118***	−0.00158*	0.00093***	−0.00176**	−0.00000	−0.00258***
Service workers	−0.00015*	−0.00075	−0.00005	−0.00135***	0.00002	−0.00033
Low-skill						
Agricultural workers	0.00005	−0.00009	0.00000	0.00000	−0.00000	−0.00006
Machine operators	0.00012	−0.00100*	−0.00019**	−0.00174***	0.00007	−0.00106***
Constant	–	−0.00715	–	−0.01092**	–	−0.03390***

CE = Composition Effect; WS = Wage Structure
*Significant at 10%; **Significant at 5%; ***Significant at 1%
Source Authors' elaboration on EU-SILC data

in wage inequality. One exception is Poland where the composition effect captures just one-quarter of the overall decrease, and inside the wage structure, a great deal of wage inequality reduction is associated to the occupation type (each profession reduces wage inequality compared to elementary jobs).

4 Concluding Remarks

In countries that experienced a decline in wage inequality, a great deal of the total changes in Gini index is due to the composition effect. In fact, up to more than 90% for Germany (where the wage structure is even not significant), three-quarters for the Czech Republic and two-third for Hungary of the reduction in wage inequality depends on the changes in workers' characteristics. In other words, endowments and potentialities of employees contributed more effectively to decrease (or at least not to increase) wage inequality in these countries.

Instead, the wage structure plays a leading role (Spain)—if not exclusive (the United Kingdom, Italy)—in increasing wage inequality, stressing the low capacity of the countries' labour markets to transform inputs into better job-related careers and higher earnings. Therefore, not only the skill endowments but also the ways in which they are rewarded in the countries' labour markets are crucial in explaining differentials in wage inequality over time. A detailed analysis identified the human capital endowments and the job-related characteristics as the individual resources that mostly contribute in shaping, in one direction or another, wage inequality gaps within the two components of composition and wage structure effects.

Those countries that experienced a decrease (or at least a not increase) in wage inequality—France, Portugal, Poland, the Czech Republic, Hungary and Germany—share shifts in the employment composition between 2005 and 2013 that led to more explicit and clearly defined structures of their labour markets (upgrading or relatively upgrading, polarisation or relatively polarisation). Probably, the employment changes, which led the labour markets towards more upgraded or polarised structures, usually less unequal, discontinued the inequality growth within the country with an equalising effect on the wage distribution.

In Greece, the employment changes towards a more polarised pattern only slowed the growth in inequality within the country, mainly due to the recent crisis that has hit Greece so even harder. Conversely, in Italy and Spain, where the distribution of occupations by skill levels appears to be more ambiguous, the increasing differentials in wage inequality are mostly attributable to the lower efficiency of their labour markets to offer better job opportunities and careers, and thus, better salaries for employees. In other words, the outright deterioration of all jobs, irrespective of skill levels required, and the lack of a clear structure of the Italian and Spanish labour markets have exacerbated disparities among the three sub-groups of employees, increasing the overall wage inequality within countries.

References

1. Autor, D.: Outsourcing at will: the contribution of unjust dismissal doctrine to the growth of employment outsourcing. J. Labor Econ. **21(1)** 1–42 (2003)
2. Barsky, R., Bound, J., Charles, K., Lupton, J.: Accounting for the black-white wealth gap: a nonparametric approach. J. Am. Stat. Assoc. **97**(459), 663–673 (2002)
3. Blinder, A.: Wage discrimination: reduced form and structural estimates. J. Hum. Resour. **8**, 436–455 (1973)
4. Castellano, R., Musella, G., Punzo, G.: Structure of the labour market and wage inequality: evidence from European countries. Qual. Quant. **51**(5), 2191–2218 (2017)
5. DiNardo, J., Fortin, N., Lemieux, T.: Labor Market Institutions and the Distribution of Wages, 1973–1993: A semi-parametric approach. Econometrica **64**, 1001–1045 (1996)
6. Dustmann, C., Ludsteck, J., Schonberg, U.: Revisiting the German wage structure. Q. J. Econ. **124**(2), 843–881 (2009)
7. Esping-Andersen, G.: Social Foundation of Post-Industrial Economies. Oxford University Press, Oxford (1999)
8. Eurofound: Drivers of Recent Job Polarisation and Upgrading in Europe: European Jobs Monitor 2014, Publications Office of the European Union, Luxembourg (2014)
9. Eurostat: Educational Intensity of Employment and Polarization in Europe and the US, Eurostat Methodologies and Working Paper (2010)
10. Firpo, S., Fortin, N., Lemieux, T.: Decomposing wage distributions using recentered influence function regressions, University of British Columbia (2007)
11. Fortin, N., Lemieux, T., Firpo, S.: Decomposition Methods in Economics, Handbook of Labor Economics (2011)
12. Frémeaux, N., Piketty, T.: France: How taxation how increase inequality. In: Nolan, et al., (eds) Changing Inequalities and Societal Impacts in Rich Countries: Thirty Countries' Experiences. Oxford University Press, Oxford (2014)
13. Fuchs-Schundeln, N., Krueger, D., Sommer, M.: Inequality trends for germany in the last two decades: a tale of two countries. Rev. Econ. Dyn. **13**(1), 103–132 (2010)
14. Galbraith, J.K., Kum, H.: Estimating the inequality of household incomes: a statistical approach to the creation of a dense and consistent global data set. Rev. Income Wealth **51**(1), 115–143 (2005)
15. Gardeazabal, J., Ugidos, A.: More on identification in detailed wage decompositions. Rev. Econ. Stat. **86**(4), 1034–1036 (2004)
16. Garofalo, A., Castellano, R., Punzo, G., Musella, G.: Skills and labour incomes: how une-qual is Italy as part of the Southern European countries? Qual. Quant. **52**, 1471–1500 (2018)
17. Goos, M., Manning, A.: Lousy and lovely jobs: the rising polarization of work in Britain. Rev. Econ. Stat. **89**, 118–133 (2007)
18. Hurley, J., Storrie, D., Jungblut, J.M.: Shifts in the Job Structure in Europe During the Great Recession. Publications Office of the European Union, Luxembourg (2015)
19. Juhn, C., Murphy, K.M., Pierce, B.: Accounting for the Slowdown in Black-White Wage Convergence. In: Kosters, M.H. (ed.) Workers and Their Wages: Changing Patterns in the United States, pp. 107–143. AEI Press, Washington D.C. (1991)
20. Juhn, C., Murphy, K., Pierce, B.: Wage inequality and the rise in returns to skill. J. Polit. Econ. **101**, 410–442 (1993)
21. Kenworthy, L., Smeeding, T.: The United States: high and rapidly-rising inequali-ty. Inequal. Its Impacts **2**, 695–717 (2013)
22. Kuznets, S.: Economic growth and income inequality. Am. Econ. Rev. **45**(1), 1–28 (1955)
23. Machado, J.A.F., Mata, J.: Counterfactual decomposition of changes in wage distributions using quantile regression. J. Appl. Econ. **20**, 445–465 (2005)
24. Machin, S.: Changes in wage inequality over the last forty years. In: Gregg, P., Wadsworth, J. (eds.) The Labour Market in Winter: The State of Working Britain. OUP, Oxford (2011)

25. Nolan, B., Salverda, W., Checchi, D., Marx, I., McKnight, A., Tòth, I.G., van de Werfhorst, H.G. (eds.): Changing Inequalities and Societal Impacts in Rich Countries: Thirty Countries' Experiences. OUP Oxford (2014)
26. Oaxaca, R.: Male-female wage differentials in urban labor markets. Int. Econ. Rev. **14**(3), 693–709 (1973)
27. Oaxaca, R., Ransom, M.R.: Identification in detailed wage decompositions. Rev. Econ. Stat. **81**(1), 154–157 (1999)

An Analysis of Wage Distribution Equality Dynamics in Poland Based on Linear Dependencies

Viktoriya Voytsekhovska and Olivier Karl Butzbach

Abstract This work investigates the gross wage distribution dynamics in Poland in different time periods. The study includes several stages and components. We first estimate the linear relationships between wages in adjacent time periods, along with the content analysis of the obtained constant dependency coefficients. We observe differences in the dynamics of wage growth across classes of wage-earners. We also calculate the value of Gini coefficients, as well as the characteristics of wage equality distribution. We then analyze the obtained linear dependencies with the use of dispersion elements analysis. Our findings show that the dynamics of wages distribution in Poland are in line with the government's goals with regard to a fairer wage distribution consistent with the current stage of the country's socio-economic development. We then analyze these findings in light of the dynamics of wage distribution. In particular, we focus on differences in wage growth across classes of wage earners. A logarithmic function of the cost effect is used for quantitative analysis. Overall, this study contributes to the literature on the patterns of wage distribution dynamics, which have important policy implications for Poland and other countries at similar stages of socio-economic development.

Keywords Wages dynamics · Income distribution · Inequality · Gini coefficient · Poland

V. Voytsekhovska (✉)
Department of Economics and Management, Lviv Polytechnic National University, Lviv, Ukraine
e-mail: viktoriia.v.voitsekhovska@lpnu.ua

O. K. Butzbach (✉)
Department of Political Sciences, Università degli Studi della Campania, Luigi Vanvitelli, Naples, Italy
e-mail: OlivierKarl.BUTZBACH@unina2.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_30

1 Introduction

Increased attention has been paid to income inequality over the past decade. Government bodies and international organizations alike have warned against the social and economic consequences of increasing income inequality. As Thomas Piketty, Anthony Atkinson and colleagues have shown in various studies, income inequality in most advanced economies have been increasing for about forty years, after a long decline that started in the 1920s and ended in the 1960s [1, 2]. In its recent update on international income inequality statistics, the OECD notes that while in the 1980s, the top 10% income earners earned 7.1 times the income accrued to the bottom 10%, this ratio has reached 9.5 in recent years [3].

Multiple factors can help explain such trend in the long-term: for instance, increased income inequality may result from higher investment in human capital combined with increasing demand for highly skilled employees [1]; see also [4]; and, for an early presentation of a similar argument [5]. Thomas Piketty, in his book on "Capital in the XXIst century", identifies more structural factors that explain patterns of income inequality since the late XIXth century. Other factors have also been emphasized, such as taxation, educational policies and decision-making structures [6–8].

However, most of the existing approaches to income inequality are empirically rooted in analyses of advanced economies' patterns. Much remains to be known about income and wage inequality patterns in emerging or developing economies. The aim of the present paper is precisely to shed light on such patterns in the case of Poland—a "transition" economy that transformed from a post-socialist economy to a European Union member in the past two decades.

2 A Review of the Literature on Income Inequality and Economic Growth

The renewed interest in understanding the causes of income inequality should not make us forget earlier economic research, such as the seminal works by Domar [9], Kuznets [10]. In these works, Domar and Kuznets were seeking to correlate economic growth with income distribution and unemployment. Champernowne [11] considered income distribution from the point of view of Pareto's law for occupational groups, which were approximated by a linear model.

A more general relationship between income inequality and economic growth was proposed by Kuznets in his 1955 work, which showed that the Gini index, a well-known indicator of inequality, changes along rates of economic growth. In the same work, Kuznets suggested that to fully understand income inequality economists should use findings of other scientific disciplines, especially as they relate to technological and other socio-economic changes.

Wage (in)equal distribution has been specifically analyzed by Sattinger [12], who found that the latter was decisively affected by the nature and cost of unemployment, on the one hand, and broader income inequality, on the other hand. In particular, Sattinger raised a number of important questions concerning wage distribution inequalities in the US labour market. In a study echoing Kuznets' works [13] analyzes the relationship between income inequalities and economic growth and proposes to further identify the different channels through which inequality in different parts of the distribution may influence the growth process.

One such channel may be the minimal wage, which may affect the entire wage distribution. Such was the focus of a study by Neumark et al. [8], which considered, in particular, wage levels, working hours, employment rates. Neumark et al. find that countries where workers earn the minimum wage when hired are more profoundly affected by minimum wage increases, while higher wage earners in these states are less affected, thus reducing wage inequalities.

The other side of the relationship between economic growth and income inequality has also been the object of a sizeable literature. One may mention, in particular, a recent work by Molero-Simarro [14], which analyzes the relationship through the Bhaduri-Marglin Model, explaining growth in terms of the effect that factor shares have on aggregate demand.

The aim of the present paper is to draw on this literature on the relationship between income inequality, economic growth and wage distribution, using statistical dependencies in the case of Poland.

3 Data and Findings

As mentioned above, income inequality is on an upward trend globally—as can be seen by the average increase in the Gini index[1] across countries. This trend may seem paradoxical in countries with a positive growth in value-added, given the fact that the latter is usually associated with new job creations and lower unemployment, which in turn shall decrease unemployment and income inequality. Figure 1 shows the latest (as of 2017) value of the Gini index for OECD countries, along with the value of the index in the mid-1980s for the same countries.

As can be observed from Fig. 1, the Gini coefficient has not increased in all OECD countries; in some countries it is lower than in the mid-1980s, or than a more recent benchmark year (2007; see OECD [3]).

In fact, as shown in a recent country report on Poland for the OECD [6] economic growth does not affect all countries equally; and periods are not homogenous in the past three decades. Poland, in particular, experienced a sharp increase in its Gini index since the mid-1980s—mostly the impact of postsocialist transition [15]. However, since 2007 the Gini index has stagnated—registering a slight decrease over the past

[1]The Gini index, an indicator of household income inequality, may have values comprised between 0 and 1, where 0 is complete equality and 1 is complete inequality.

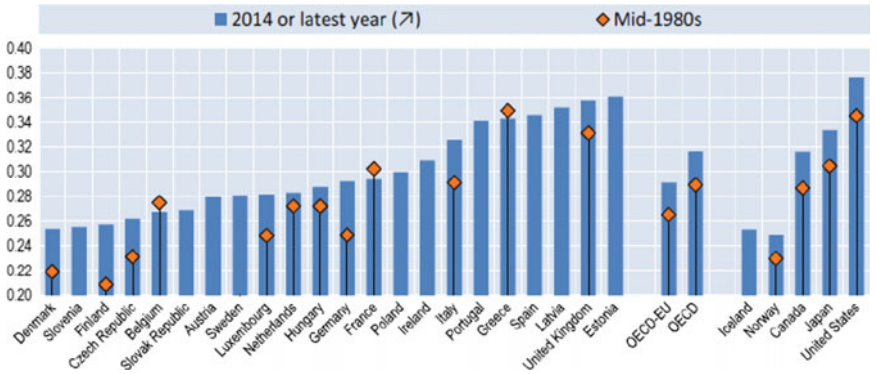


Fig. 1 Gini coefficient of disposable income inequalities. *Source* OECD [3]

decade (from 0.32 to 0.30). These two trends can be observed against the backdrop of constant, positive economic growth since 1990. Poland is the only EU country that did not experience global recession due to its preventive timely implemented fiscal central bank policies. In any case, according to several studies, the impact of the Great Recession on income inequalities was modest [16, 17]. In particular, according to Jenkins and colleagues, household incomes in many advanced economies were protected from the impact of the crisis by automatic stabilizers and welfare states [17].

As mentioned in the previous section, income inequality may be caused by a multiplicity of factors. Among such factors are the dynamics of wage growth and their differences across categories of wage earners.

Thus the remaining part of this paper analyzes the dynamics of gross wage growth in Poland over the period 2002–2016 with two different purposes: to calculate Gini coefficient and to identify wage inequalities in different time intervals. Gross wagees were chosen because of greater availability of statistical data, on the one hand, and the possibility of comparison with other countries, on the other hand. Polish statistics allow for 15 intervals [18–23]. The corresponding wage distribution is shown on Fig. 2.

Figure 2 illustrates the asymmetrical distribution of wages in Poland in 2012—the curve has a dome-shaped form; higher wages are associated with lower frequencies.

By using the appropriate accumulated values it is possible to determine, for Poland, the Lorenz curve and the Gini coefficient (see Fig. 3).

To simplify the determination of Gini coefficient, the Lorenz curve is approximated by the following polynomial function: $y = 0.008\ x^2 + 0.017\ x + 4.397$.

This enables to simplify the determination of the area under the Lorenz curve by integration $\int_0^{100} y(x)dx \int_0^{100} y(x)dx$, which equals 3191,367 for the chosen dimension (for Gini coefficient from 0 to 5000).

Then the Gini coefficient's values were calculated for other time intervals: $G_{2002} = 33.9\%$, $G_{2010} = 31.1\%$, $G9_{2012} = 36.2\%$, $G_{2014} = 36.8\%$, $G_{2016} = 34.6\%$.

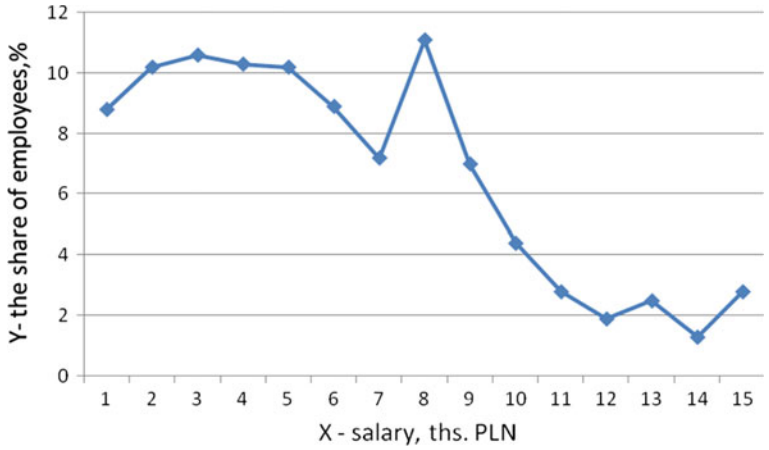


Fig. 2 Wage distribution in Poland, 2012. *Source* GUS [21]

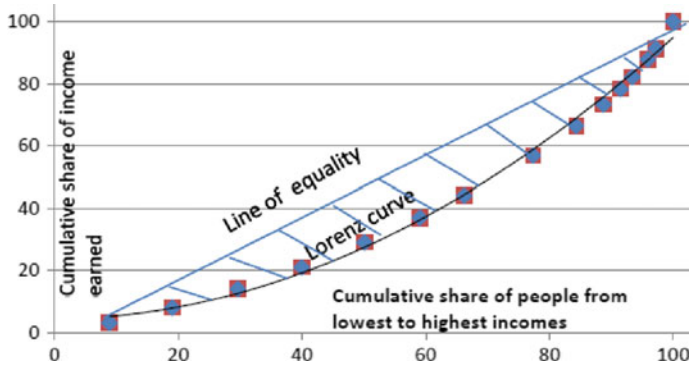


Fig. 3 Lorenz curve for the Polish data on gross wages GUS. *Source* Authors' own elaboration on GUS [21] data

Longitudinal fluctuations of Gini coefficient values (for wages) indicate a disproportionate increase in employees' wages. This finding led us to analyze wage dynamics across classes of wage-earners.[2]

With this aim, the primary array of data was divided in five groups according to relative wage size. The feature was a stable number of employees in time. In each of these groups, average wage levels were determined and thus, each year, five average salaries were obtained with the corresponding frequency distribution. Next, we wished to find the interdependence between the mean in adjacent periods of time. Such dependencies, based on the use of correlation methods in the form of linear dependencies, proved to be rather tight. Here are two examples of the following dependencies: $X(2014) = 1.039 \times 2012 + 0.174$; $R^2 = 0.999$; $X(2010) = 1.613 \times$

[2]Wages in this study mean gross wages (before tax).

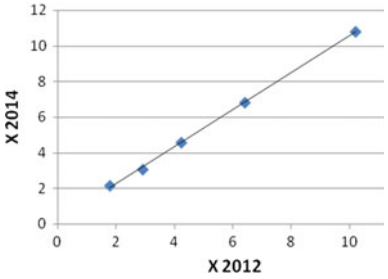


Fig. 4 Relationships of wages 2012 and 2014. *Source* Authors' own elaboration

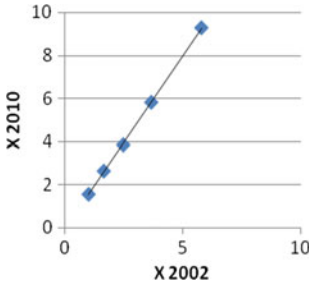


Fig. 5 Relationships of wages 2002 and 2010. *Source* Authors' own elaboration

2002 – 0.088; $R^2 = 0.999$. The high degree of correlation between wages in adjacent time periods suggests that wage growth is linearly determined by their basic value.

Below is a graphical representation of these dependencies (see Figs. 4 and 5).

Using the obtained equations, knowing the average wage in a category of employees in the base period, one can determine the average wage in the same category in the next period. We thus find that the dynamics of wage growth varies across categories of employees. For the five selected groups (representing different quintiles of wage earners, this dynamic is shown on Fig. 6.

According to our calculations based on GUS data, average wages of the lowest income earners (group 1) increased from 1.564 to 2.148 thousand PLN between 2010 and 2014—a 37% increase. During the same period, average wages in the higher earners' group (group 5) increased from 9.283 to 10.806 thousand PLN a 16% increase. Overall, average wages across the five groups (or categories) of employees increased from 3.373 thousand PLN in 2010 to 3.981 thousand PLN in 2014—a 18% increase. We observe, therefore, different wage growth dynamics across the different groups of income earners.

The largest difference in wage growth can be observed between the lowest and highest income earners.

The growth rates of wages can also be determined from the equations obtained as follows: $\frac{x_{2010}}{x_{2002}} = 1.613 - \frac{0.091}{x_{2002}}$; $\frac{x_{2014}}{x_{2012}} = 1.039 + \frac{0.174}{x_{2012}}$.

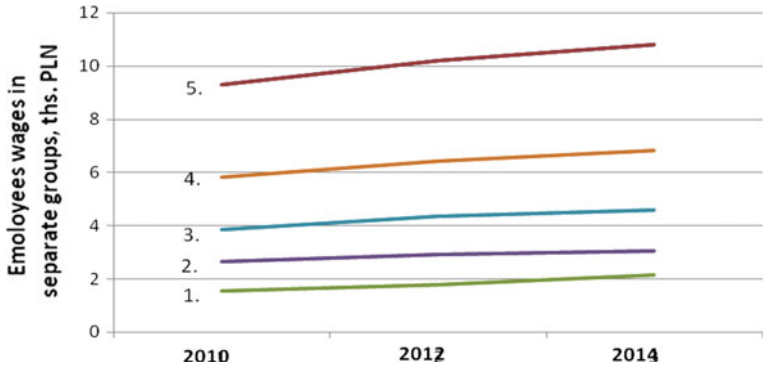


Fig. 6 The dynamics of wages in 5 major groups of employees. *Source* Authors' own elaboration on GUS (2004–2016) data

From these relations we can draw conclusions regarding the dynamic structure of wage growth in Poland. Where the constant coefficient of linear dependency is negative, lower wages are growing at a lower pace than higher ones. At the same time, interpretation with a positive constant coefficient depending on the trend will be reversed—lower wages will increase with greater intensity. Therefore, in a country such as Poland (with some degree of fluctuations in economic growth) there may be different (opposite) trends in the structural dynamics of wages. Obviously, there may be times when these dynamics are similar across wage levels.

Let us now consider the analytical aspects of wage distribution concerning the linear growth of wages for two adjacent periods of time, t and $t + 1$: $x_{t+1} = a + bx_t$, where a and b are constant coefficients. The property of such a relationship is that under condition that $\alpha \neq 0$ the growth rates of different wages are different. For $a > 0$, $b > 0$ lower wages grow more intensively than higher wages. For $a < 0$ and $b > 0$ there will be a reverse trend—lower wages increase at a lower pace. In the variation where $a = 0$, the growth of all wages is proportional to the same degree. This implies that in order to reach wage equality, or to reduce wage inequalities, it is necessary, ceteris paribus, that lower wages grow at a more rapid pace than higher wages. We have already showed the linear dependencies determined by statistical correlation above. But let us note that other approaches and analytical relations can be used to find stable coefficients of linear dependencies, especially for variants of close correlation between variables. For a linear dependency, the coefficient b is defined as follows: $b = \frac{\sigma x_{t+1}^2}{\sigma x_t^2}$, where standard deviations for x_t and x_{t+1} are used. Using also average values, the coefficient a is defined as follows: $a = \frac{1}{x_{t+1}^2} \frac{Vx_{t+1}^2}{(1-Vx_t^2)}$, where Vx_{t+1} is the coefficient of variation for $x_{t+1,\text{ and}}$ Vx_t is the coefficient of variation for x_t.

The expression for a implies that the sign of this coefficient depends on the ratio of the variation coefficients. In particular, this sign is positive when the coefficient of variation for x_{t+1} is smaller than the coefficient of variation for x_t. This is a condition for wages dynamics consistent with higher equity levels.

Table 1 Wage growth dynamics in Poland, 2002–2016

t	$\bar{X}$	σ_X	V_x	a	b
2002	2.156	1.2696	0.5888	–	–
2010	3.373	2.0147	0.5973	−0.0485	1.5869
2012	3.734	2.2639	0.6063	−0.0562	1.1237
2014	3.981	2.3583	0.5924	0.0913	1.0417
2016	4.197	2.4429	0.5820	0.0732	1.0359

Source Authors' own elaboration on GUS (2002–2017) data

By contrast, when Vxt + 1 > Vxt, wage growth is not consistent with achieving equity in the distribution of wages. In other words, the growth of lower wages is slower than the growth of higher wages. The proposed formulas, elaborated on the Polish data, result in the values shown in Table 1.[3]

The analysis of the Polish data shows that in the period from 2002 to 2012, the coefficient of variation achieves it maximum and then begins to decrease. At the same time, the constant a changes sign, from negative to positive. Thus, in recent years, wage growth in Poland has shifted in a way that is now consistent to higher equality in wage distribution. What may explain such a shift? Further analysis is required to unveil the relevant causal factors here. One decisive factor, we assume, consists in Poland's accession to the EU, with the subsequent, gradual process of economic integration it entailed, accompanied by sustained economic growth and, consequently, greater social protection of workers.

Accordingly, we perform the relevant calculations for 2012 and 2014, using aggregated data. The conditional total effect is calculated with the formula: $E = a \sum_{i=1}^{5} p_i \ln x_i$, where x_i is the wage of one employee in i-group and P_i—is the share of employees in i-group. It is assumed that the effect is proportional to some constant value of a. In addition to the actual data on wage growth in 2014, we also calculated the theoretical version of their proportional growth with the same pace. The latter is defined as the ratio of the total salary in 2014 to its value in 2012. The results of such calculation are shown in Table 2.

Comparison of 2012 data with 2014 data shows that the average salary increased in 2014 by 6.6%. Thus Table 3 incorporates an increase of average wages, for each group, of 6.6%.

The empirical analysis shows that the total effect is greater in the actual growth of wages than in the theoretical proportional version (127,1203 > 125,3472). This is due to the difference in the distribution structure. The actual distribution is more consistent with wage distribution equity. In the group of lower wage earners (group 1), average wages grew in 2014 to a greater extent than in the proportional increase. At the same time, in the group of higher wage earners, the trend is reversed. This result is consistent with theoretical considerations, as well as with the concepts of wage growth in line with linear laws. Thus, the deviation from the proportional

[3]Note that the arrays of primary data were used directly without their aggregation into groups.

Table 2 The calculation of the effect of the actual wage distribution on wage earners' utility function

Wage amount, ths. PLN, X_i	Wage share (%) P_i	P_iX_i	$\ln X_i$	$P_i \ln X_i$
2.1482	29.6	63.58672	0.76463	22.63306
3.0567	29.4	89.86698	1.117336	32.84968
4.5727	25.3	115.6893	1.520104	38.45863
6.8195	9.1	62.05745	1.919786	17.47005
10.8063	6.6	71.32158	2.380129	15.70885
–	Σ: 100	Σ: 402.522	–	Σ: 127.1203

Source: Author's elaboration on grouped actual data for Poland (2014), GUS

Table 3 The effect calculation under condition of proportional wages growth

Wage amount, ths. PLN, X_i	Wage share, (%) P_i	P_iX_i	$\ln X_i$	$P_i \ln X_i$
1.930908	29.6	57.21976	0.65799	19.4765
3.156025	29.4	92.78712	1.149313	33.7898
4.590196	25.3	116.132	1.523923	38.55525
6.957568	9.1	63.31387	1.93983	17.65245
11.07883	6.6	73.12027	2.405036	15.87324
–	100	402.5730	–	125.3472

Source: Author's elaboration on grouped actual data for Poland (2014), GUS

increase in salaries is an appropriate way towards a more equal distribution of wages at particular stages of economic growth after the transition period.

4 Conclusions

The study of dynamics of gross wages growth in Poland gives us important insights in the understanding of wage inequalities in dynamic terms. In particular, our analysis has showed that, in addition to changes in total earnings over time, different dynamics characterize the various groups of wage earners. Specifically, these changes have two variants. The first of these consists in the fact that lower wages grow at a lower pace than higher ones.

The second variant is characterized by a reverse trend—higher wages grow more intensively than lower wages. At certain periods of time there may also be a proportional increase in all wages (an indexation option). This is where the method chosen here, that of linear dependency, may be used proficuously with regard to the relationship between wage levels and growth rates in successive periods of time. In

accordance with the format of GUS Poland statistical data, this period was generally taken to be equal to two years.

The use of correlation methods revealed a close linear relationship of wages in successive time periods. At the same time, it is essential to describe this relationship by means of linear dependency with two stable coefficients, typical versions of the growth dynamics of the continuous set of wages. The variants are identified with a free constant coefficient sign in linear dependency. The presence of a positive sign means a more intense growth of lower earnings (wages). And the negative sign shows a more intense growth of higher wages. Under the condition of close linear relationships to assess the options for increasing wages, the variation rates for wages in the adjoining periods of time can be used. The advantage of this approach is that the primary array of statistics is used directly without aggregating and comparing the group average over time.

The distribution of wages can be considered from the standpoint of equality. The development of the G index for gross wages showed the existence of small fluctuations during the investigated time period. It should be noted that the principle of proportional payment for labour may prevail in the production sphere, which to some extent does not comply with the criterion of social justice. Also the political variables were not taken into account. Because of this, the Gini index can only be reduced to a certain limit. In general, we considered only particular earnings set transformation in the relation to the basic time moment. In the aspect of optimization, there is a problem regarding the distribution of the wage's fund growth in a narrow time interval. In turn, the incremental efficiency is associated with derivatives that are used to linearly approximate the efficiency function. The development of variants of possible active influence on the structure of wages and the dynamics of their growth requires further analysis.

References

1. Atkinson, A.B., Piketty, T., Saez, E.: Top incomes in the long run of history. J. Econom. Literat. **49**, 3–71 (2011)
2. Piketty, T.: Capital in the Twenty-First Century. Harvard University Press, Cambridge and London (2014)
3. OECD: Understanding the Socio-economic Divide in Europe. OECD, Paris (2017)
4. Glyn, A.: Functional distribution and inequality. In: Salvedra, W., Nolan, B., Smeeding, T.M. (eds.) The Oxford Handbook of Economic Inequality, pp. 101–126. Oxford University Press, Oxford (2009)
5. Tinbergen, J.: Income distribution. Amsterdam, North-Holland (1975)
6. Brzeziński, M., Jancewicz, B., Letki, N.: Gini Growing Inequalities' impacts. Warsaw Country Report, Poland (2014)
7. Inklaar, R., Prasada Rao, D.S.: Cross-country income levels over time: did the developing world suddenly become much richer? Am. Econ. J. Macroecon. **9**(1), 265–290 (2017)
8. Neumark, D., Schweitzer, M., Wascher, W.: The Effects of Minimum Wages Throughout the Wage Distribution. NBER Working Paper No. 7519 (2000)
9. Domar, E.: Expansion and Employment. Am. Econ. Rev. **37**, 34–55 (1947)
10. Kuznets, S.: Economic growth and income distribution. Am. Econ. Rev. **45**(1), 3–28 (1955)

11. Champernowne, C.G.: A model of income distribution. Econ. J. **63**(250) (1953)
12. Sattinger, M.: The distribution of wage rates. In: Unemployment, Choice and Inequality. Berlin, Springer
13. Voitchovsky, Sarah: Does the profile of income inequality matter for economic growth? J. Econ. Growth **10**(3), 273–296 (2005)
14. Molero-Simarro, Ricardo: Growth and inequality revisited: the role of primary distribution of income. A new approach for understanding today's economic and social crises. Camb. J. Econ. **41**, 367–390 (2017). https://doi.org/10.1093/cje/bew017
15. Bandelj, N., Mahutga, M.C.: How socio-economic change shapes income inequality in post-socialist Europe. Soc. Forces **88**(5), 2133–2161 (2010)
16. ILO: Global Wage Report 2010–2011. ILO, Geneva (2010)
17. Jenkins, S.P., Brandolini, A., Micklewright, J., Nolan, B. (eds.): The Great Recession and the Distribution of Household Income. Oxford University Press, Oxford (2012)
18. GUS, Rocznik statystyczny Rzeczypospolitej Polski/ Statistic yearbook of Republic of Poland (2004)
19. GUS, Rocznik statystyczny Rzeczypospolitej Polski/Statistic yearbook of Republic of Poland (2010)
20. GUS, Rocznik statystyczny Rzeczypospolitej Polski/ Statistic yearbook of Republic of Poland (2012)
21. GUS, Rocznik statystyczny Rzeczypospolitej Polski/ Statistic yearbook of Republic of Poland (2014)
22. GUS, Rocznik statystyczny Rzeczypospolitej Polski/ Statistic yearbook of Republic of Poland (2016)
23. GUS, Rocznik statystyczny Rzeczypospolitej Polski/ Statistic yearbook of Republic of Poland (2017)

Mathematical Statistics for Data Science

Unions of Orthogonal Arrays and Their Aberrations via Hilbert Bases

Roberto Fontana and Fabio Rapallo

Abstract We generate all the Orthogonal Arrays (OAs) of a given size n and strength t as the union of a collection of OAs which belong to an inclusion-minimal set of OAs. We derive a formula for computing the (Generalized) Word Length Pattern of a union of OAs that makes use of their polynomial counting functions. The best OAs according to the Generalized Minimum Aberration criterion can thereby be found simply by exploring a relatively small set of counting functions. The classes of OAs with 5 binary factors, strength 2, and sizes 16 and 20 are fully described.

Keywords Algebraic statistics · Counting function · Fractional factorial designs · Generalized word length pattern

1 Introduction

The design of factorial experiments plays a central role in several fields of Applied Statistics, from Biology to Engineering, from Computer Science to Economics. A comprehensive introduction to factorial experiments can be found in, e.g., [14]. In its simplest form, a designed experiment based on a factorial design consists in the measurement of a response variable at different levels of several explanatory variables (or factors), in order to decide what factors and interactions are actually significant and to estimate the coefficients of the resulting linear model. If all possible treatments (i.e., combinations of the factor levels) are considered in the design, it can be said that a full factorial design is used. But even in the simplest case of two-level factors, the full factorial design rapidly becomes very large when the number of factors included

R. Fontana
Department of Mathematical Sciences, Politecnico di Torino,
corso Duca degli Abruzzi 24, 10124 Torino, Italy
e-mail: roberto.fontana@polito.it

F. Rapallo (✉)
Department of Sciences and Technological Innovation, Università del Piemonte Orientale,
viale Teresa Michel 11, 15121 Alessandria, Italy
e-mail: fabio.rapallo@uniupo.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_31

in the experiment increases. Thus, we need to choose a fraction (i.e., a subset of the full factorial design). The need for efficient experimental designs has led to the definition of several criteria for the choice of the design points to be included in the fraction. All such criteria aim at producing the best estimates of the relevant parameters for a given sample size. When the effects to be estimated are known, the usual approach is a model-based one, leading to the definition of various optimality criteria (D-optimality, A-optimality, etc.). For screening designs the standard choice is to follow a model-free approach, which aims to minimize the aliasing between the effects. In this context, regular fractions defined by appropriate contrasts are popular but they have limitations on the choice of the sample size. To generalize the analysis to multilevel designs and to allow a wide range of possible sample sizes, non-regular designs and Orthogonal Arrays (OAs) have been introduced together with several criteria to define the best design in a given setting. Here we limit our attention to fractional factorial designs together with the Generalized Minimum Aberration (GMA) criterion.

Generalized Word-Length Pattern (GWLP) is an important tool for comparing fractional factorial designs in the framework of factorial experiments. Its origin goes back to Fries and Hunter [10] who proposed the concept of design aberration as a natural extension of the concept of design resolution. They considered two-level regular designs and defined the minimum aberration design as the design of maximum resolution which minimizes the number of words of minimum length in the defining relation. Non-regular multilevel designs were considered by Xu and Wu [16] which defined the concept of GWLP. Since the GWLP does not depend on the coding of the factor levels, Pistone and Rogantin [15] used the complex coding of the factor levels to express the basis of the polynomial complex functions over a design, and in particular of the counting function. When this coding is used, the coefficients of the counting function are closely related to aberrations and GWLP. Moreover, the coefficients of the counting function can be expressed in terms of the counts of the levels appearing in each simple or interaction term. General references for GWLP and its properties include [4, 12, 14].

In practice, GWLP is used to discriminate among different designs through the GMA criterion: given two designs $\mathscr{F}_1$ and $\mathscr{F}_2$ with m factors, the corresponding GWLPs are two vectors

$$A_{\mathscr{F}_i} = (A_0(\mathscr{F}_i) = 1, A_1(\mathscr{F}_i), \ldots, A_m(\mathscr{F}_i)) \qquad i = 1, 2\,.$$

The GMA criterion consists in the sequential minimization of the GWLPs: $\mathscr{F}_1$ is better than $\mathscr{F}_2$ if there exists j such that $A_0(\mathscr{F}_1) = A_0(\mathscr{F}_2), \ldots, A_j(\mathscr{F}_1) = A_j(\mathscr{F}_2)$ and $A_{j+1}(\mathscr{F}_1) < A_{j+1}(\mathscr{F}_2)$. The GMA criterion is usually applied to OAs, see [12].

In this work we use results from Combinatorics and Algebraic Geometry to ease the computation of GWLP. The connection between GWLP and the geometric structure of the design points is studied in [9], but here we adopt a different point of view. In particular, we show that the set of all OAs with given strength are the points with integer entries of a cone defined through linear constraints. This allows us to express each OA as the union of elements of the Hilbert basis of the cone. Moreover,

we show that the GWLP of the union of two or more fractions can be computed from their counting functions. The computation of the Hilbert basis is done through combinatorial algorithms and its complexity increases quickly with the number of factors and the number of factor levels. Thus, we illustrate explicit computations for relatively small designs. Nevertheless, the theory presented here may have also a theoretical interest and may be the basis of further developments.

2 Fractions, Counting Functions and Aberration

In this section, for ease in reference, we present some relevant results of the algebraic theory of Orthogonal Fractional Factorial Designs and we express the aberration of fractional designs using the coefficients of the polynomial counting function. This presentation is based on [7]. The interested reader can find further information, including the proofs of the propositions, in [8, 15].

2.1 Fractions of a Full Factorial Design

Let us consider an experiment which includes m factors $\mathscr{D}_j$, $j = 1, \ldots, m$. Let us code the s_j levels of the factor $\mathscr{D}_j$ by the s_jth roots of the unity

$$\mathscr{D}_j = \{\omega_0^{(s_j)}, \ldots, \omega_{s_j-1}^{(s_j)}\},$$

where $\omega_k^{(s_j)} = \exp\left(\sqrt{-1}\,\frac{2\pi}{s_j}\,k\right)$, $k = 0, \ldots, s_j - 1$, $j = 1, \ldots, m$.

The *full factorial design* with complex coding is $\mathscr{D} = \mathscr{D}_1 \times \cdots \mathscr{D}_j \cdots \times \mathscr{D}_m$. We denote its cardinality by $\#\mathscr{D}$, $\#\mathscr{D} = \prod_{j=1}^{m} s_j$.

Definition 1 A *fraction* $\mathscr{F}$ is a multiset $(\mathscr{F}_*, f_*)$ whose underlying set of elements $\mathscr{F}_*$ is contained in $\mathscr{D}$ and f_* is the multiplicity function $f_* : \mathscr{F}_* \to \mathbb{N}$ that for each element in $\mathscr{F}_*$ gives the number of times it belongs to the multiset $\mathscr{F}$.

The underlying set of elements $\mathscr{F}_*$ is the subset of $\mathscr{D}$ that contains all the elements of $\mathscr{D}$ that appear in $\mathscr{F}$ at least once. We denote the number of elements of a fraction $\mathscr{F}$ by $\#\mathscr{F}$, with $\#\mathscr{F} = \sum_{\zeta \in \mathscr{F}_*} f_*(\zeta)$.

In order to use polynomials to represent all the functions defined over $\mathscr{D}$, including multiplicity functions, we define

- X_j, the jth component function, which maps a point $\zeta = (\zeta_1, \ldots, \zeta_m)$ of $\mathscr{D}$ to its jth component,
$$X_j : \mathscr{D} \ni (\zeta_1, \ldots, \zeta_m) \longmapsto \zeta_j \in \mathscr{D}_j .$$

The function X_j is a *simple term* or, by abuse of terminology, a *factor*.

- $X^\alpha = X_1^{\alpha_1} \cdot \ldots \cdot X_m^{\alpha_m}$, $\alpha \in L = \mathbb{Z}_{s_1} \times \cdots \times \mathbb{Z}_{s_m}$ i.e., the monomial function

$$X^\alpha : \mathscr{D} \ni (\zeta_1, \ldots, \zeta_m) \mapsto \zeta_1^{\alpha_1} \cdot \ldots \cdot \zeta_m^{\alpha_m} .$$

The function X^α is an *interaction term*.

We observe that $\{X^\alpha : \alpha \in L = \mathbb{Z}_{s_1} \times \cdots \times \mathbb{Z}_{s_m}\}$ is a basis of all the complex functions defined over $\mathscr{D}$. We use this basis to represent the counting function of a fraction according to the following definition.

Definition 2 The *counting function* R of a fraction $\mathscr{F}$ is a complex polynomial defined over $\mathscr{D}$ so that for each $\zeta \in \mathscr{D}$, $R(\zeta)$ equals the number of appearances of ζ in the fraction. A $0-1$ valued counting function is called an *indicator function* of a single-replicate fraction $\mathscr{F}$. We denote by c_α the coefficients of the representation of R on $\mathscr{D}$ using the monomial basis $\{X^\alpha,\ \alpha \in L\}$:

$$R(\zeta) = \sum_{\alpha \in L} c_\alpha X^\alpha(\zeta),\ \zeta \in \mathscr{D},\ c_\alpha \in \mathbb{C} .$$

With Proposition 1 taken from [15], we link the orthogonality of two interaction terms with the coefficients of the polynomial representation of the counting function. We denote by $\overline{z}$ the complex conjugate of the complex number z.

Proposition 1 *If $\mathscr{F}$ is a fraction of a full factorial design $\mathscr{D}$, $R = \sum_{\alpha \in L} c_\alpha X^\alpha$ is its counting function and $[\alpha - \beta]$ is the m-tuple made by the componentwise difference in the rings $\mathbb{Z}_{s_j}$, $[\alpha - \beta] = \left([\alpha_1 - \beta_1]_{s_1}, \ldots, [\alpha_m - \beta_m]_{s_m}\right)$, then*

1. *the coefficients c_α are given by $c_\alpha = \frac{1}{\#\mathscr{D}} \sum_{\zeta \in \mathscr{F}} \overline{X^\alpha(\zeta)}$;*
2. *the term X^α is centered on $\mathscr{F}$ i.e., $\frac{1}{\#\mathscr{F}} \sum_{\zeta \in \mathscr{F}} X^\alpha(\zeta) = 0$ if, and only if, $c_\alpha = c_{[-\alpha]} = 0$;*
3. *the terms X^α and X^β are orthogonal on $\mathscr{F}$ if and only if $c_{[\alpha-\beta]} = 0$.*

We now define projectivity and, in particular, its relationship with OAs. Given $I = \{i_1, \ldots, i_k\} \subset \{1, \ldots, m\}$, $i_1 < \ldots < i_k$ and $\zeta = (\zeta_1, \ldots, \zeta_m) \in \mathscr{D}$ we define the projection $\pi_I(\zeta)$ as

$$\pi_I(\zeta) = \zeta_I = (\zeta_{i_1}, \ldots, \zeta_{i_k}) \in \mathscr{D}_{i_1} \times \ldots \times \mathscr{D}_{i_k} .$$

Definition 3 A fraction $\mathscr{F}$ *factorially projects* onto the I-factors, $I = \{i_1, \ldots, i_k\} \subset \{1, \ldots, m\}$, $i_1 < \ldots < i_k$, if the projection $\pi_I(\mathscr{F})$ is a multiple of a full factorial design, i.e., the multiset $(\mathscr{D}_{i_1} \times \ldots \times \mathscr{D}_{i_k}, f_*)$ where the multiplicity function f_* is constant over $\mathscr{D}_{i_1} \times \ldots \times \mathscr{D}_{i_k}$.

Definition 4 A fraction $\mathscr{F}$ is a *(mixed) Orthogonal Array (OA)* of strength t if it factorially projects onto any I-factors with $\#I = t$.

Proposition 2 *A fraction factorially projects onto the I-factors, $I = \{i_1, \ldots, i_k\} \subset \{1, \ldots, m\}$, $i_1 < \ldots < i_k$, if and only if all the coefficients of the counting function involving the I-factors only are* 0.

Proposition 2 can be immediately stated for mixed orthogonal arrays.

Proposition 3 *A fraction is an OA of strength t if and only if all the coefficients c_α, $\alpha \neq 0 \equiv (0, \ldots, 0)$ of the counting function up to the order t are* 0.

2.2 GWLP and Aberrations

Using the polynomial counting function, [3] provides the following definition of the GWLP $A_{\mathscr{F}} = (A_0(\mathscr{F}), \ldots, A_m(\mathscr{F}))$ of a fraction $\mathscr{F}$ of the full factorial design $\mathscr{D}$.

Definition 5 The Generalized Word-Length Pattern (GWLP) of a fraction $\mathscr{F}$ of the full factorial design $\mathscr{D}$ is the vector $A_{\mathscr{F}} = (A_0(\mathscr{F}), A_1(\mathscr{F}), \ldots, A_m(\mathscr{F}))$, where

$$A_j(\mathscr{F}) = \sum_{|\alpha|_0 = j} a_\alpha \quad j = 0, \ldots, m\,,$$

$$a_\alpha = \left(\frac{\|c_\alpha\|_2}{c_0}\right)^2, \tag{1}$$

$|\alpha|_0$ is the number of non-null elements of α, $\|z\|_2$ is the norm of the complex number z, and $c_0 = c_{(0,\ldots,0)} = \#\mathscr{F}/\#\mathscr{D}$.

We refer to a_α as the *aberration* of the interaction X^α. In Proposition 4 we provide a formula to compute a_α, and consequently $A_j(\mathscr{F})$, $j = 1, \ldots, m$, given a fraction $\mathscr{F}$ of $\mathscr{D}$. Notice that $A_0(\mathscr{F}) = 1$ for all $\mathscr{F}$. Moreover, in the case of binary designs, the coefficients of the counting function are real numbers and therefore the aberrations in Eq. (1) are simply

$$a_\alpha = \left(\frac{c_\alpha}{c_0}\right)^2.$$

Given a fraction $\mathscr{F}$ of the full factorial design $\mathscr{D}$, let us consider its counting function $R = \sum_{\alpha \in L} c_\alpha X^\alpha$. From item 1 of Proposition 1 the coefficients c_α are given by

$$c_\alpha = \frac{1}{\#\mathscr{D}} \sum_{\zeta \in \mathscr{F}} \overline{X^\alpha(\zeta)}$$

or equivalently

$$c_\alpha = \frac{1}{\#\mathscr{D}} \sum_{\zeta \in \mathscr{D}} R(\zeta)\overline{X^\alpha(\zeta)}\,.$$

To make the notation easier we use vectors and matrices and we make the non-restrictive hypothesis that both the runs ζ of the full factorial design $\mathscr{D}$ and the multi-indexes of $L = \mathbb{Z}_{s_1} \times \cdots \times \mathbb{Z}_{s_m}$ are considered in lexicographic order. We obtain

$$c_\alpha = \frac{1}{\#\mathscr{D}} \overline{X}_\alpha^T Y = \frac{1}{\#\mathscr{D}} Y^T \overline{X}_\alpha \,,$$

where X_α is the column vector $[\zeta^\alpha : \zeta \in \mathscr{D}]$, $\overline{X}_\alpha$ is the column vector $\left[\overline{\zeta}^\alpha : \zeta \in \mathscr{D}\right]$ Y is the column vector $[R(\zeta) : \zeta \in \mathscr{D}]$ and the exponent T denotes the transpose of a matrix. The square of the norm of a complex number z can be computed as $z\overline{z}$. It follows that

$$\|c_\alpha\|_2^2 = c_\alpha \overline{c_\alpha}$$

and therefore we get

$$(\#\mathscr{D})^2 \|c_\alpha\|_2^2 = (Y^T \overline{X}_\alpha)\overline{(\overline{X}_\alpha^T Y)} = Y^T \overline{X}_\alpha {X_\alpha}^T Y \,.$$

As in [5], we refer to Y as the *counting vector* of a fraction.

2.3 Counting Vector and Aberrations

Here we present some properties of the aberrations and some results about the relationships between the aberrations and the counting vector of a fraction. The results are adapted to the complex coding for multilevel factors.

Proposition 4 *Given a fraction $\mathscr{F}$ it holds:*

1. $a_\alpha = (Y^T \overline{X}_\alpha {X_\alpha}^T Y)/(\#\mathscr{F})^2$;
2. $\mathrm{Su}(\overline{X}_\alpha {X_\alpha}^T) = 0$, $\alpha \neq 0$ *where* $\mathrm{Su}(A)$ *is the sum of all the elements of the matrix* A;
3. $\sum_{j=0}^m A_j(\mathscr{F}) = \sum_{\alpha \in L} a_\alpha = (\#\mathscr{D} \sum_{\zeta \in \mathscr{D}} Y[\zeta]^2)/(\#\mathscr{F})^2$;
4. *if* $Y[\zeta] \in \{0, 1\}, \zeta \in \mathscr{D}$ *then* $\sum_{i=0}^m A_i(\mathscr{F}) = \#\mathscr{D}/\#\mathscr{F} = c_0^{-1}$.

Proof 1. From the definition of a_α we get

$$a_\alpha = \left(\frac{\|c_\alpha\|_2}{c_0}\right)^2 = \frac{(1/\#\mathscr{D})^2 Y^T \overline{X}_\alpha {X_\alpha}^T Y}{(\#\mathscr{F}/\#\mathscr{D})^2} = \frac{Y^T \overline{X}_\alpha {X_\alpha}^T Y}{(\#\mathscr{F})^2} \,.$$

2. Let us consider the full factorial design $\mathscr{D}$. Its counting vector is 1, i.e., the column vector with all the components equal to 1. The coefficients of its counting function are $c_0 = 1$ and $c_\alpha = 0$ for all $\alpha \neq 0$. We get $a_\alpha = 0$ for all $\alpha \neq 0$. It follows that the sum of all the elements of the matrix $\overline{X}_\alpha {X_\alpha}^T$ is

$$\mathrm{Su}(\overline{X}_\alpha {X_\alpha}^T) = 1^T \overline{X}_\alpha {X_\alpha}^T 1 = (\#\mathscr{D})^2 a_\alpha = 0, \ \alpha \neq 0 \,.$$

3. The sum of all the terms of the GWLP is

$$\sum_{j=0}^{m} A_j(\mathscr{F}) = \sum_{\alpha \in L} a_\alpha = \sum_{\alpha \in L} \frac{Y^T \overline{X}_\alpha {X_\alpha}^T Y}{(\#\mathscr{F})^2} =$$
$$= \frac{Y^T \sum_{\alpha \in L} (\overline{X}_\alpha {X_\alpha}^T) Y}{(\#\mathscr{F})^2} = \frac{Y^T \overline{X} X^T Y}{(\#\mathscr{F})^2} =$$
$$= \frac{\#\mathscr{D}\, Y^T Y}{(\#\mathscr{F})^2} = \frac{\#\mathscr{D} \sum_{\zeta \in \mathscr{D}} Y(\zeta)^2}{(\sum_{\zeta \in \mathscr{D}} Y(\zeta))^2},$$

where X is the orthogonal matrix whose columns are X_α, $\alpha \in L$.

4. It follows from item 3. by observing that $Y[\zeta] \in \{0, 1\}$, $\zeta \in \mathscr{D} \Rightarrow Y[\zeta]^2 = Y[\zeta]$ and then $\sum_{\zeta \in \mathscr{D}} Y[\zeta]^2 = \#\mathscr{F}$.

□

From items 3. and 4. of Proposition 4 we obtain that, for a given size n, the total aberration of a single-replicate fraction $\mathscr{F}_1$ (with counting vector Y_1) will be less than the total aberration of a fraction $\mathscr{F}_2$ (with counting vector Y_2) that admits replications. In fact, we get

$$\sum_{j=0}^{m} A_j(\mathscr{F}_1) = \frac{\#\mathscr{D}}{n}, \qquad \sum_{j=0}^{m} A_j(\mathscr{F}_2) = \frac{\#\mathscr{D} \sum_{\zeta \in \mathscr{D}} Y_2[\zeta]^2}{n^2}$$

and

$$\frac{\#\mathscr{D}}{n} \leq \frac{\#\mathscr{D}}{n} \frac{\sum_{\zeta \in \mathscr{D}} Y_2[\zeta]^2}{n}$$

because, given n, $\sum_{\zeta \in \mathscr{D}} Y_2[\zeta]^2 \geq n$.

Now, as in [11], let us consider the special case of OAs of size n and strength t (or equivalently with resolution $t + 1$), with $m = t + 1$ factors. Using the standard notation, we denote this class of OAs by $OA(n, s_1 \ldots s_m, m - 1)$. We can state the following proposition.

Proposition 5 *Let $\mathscr{F} \in OA(n, s_1 \ldots s_m, m - 1)$. Then*

$$A_m(\mathscr{F}) = \frac{\#\mathscr{D} \sum_{\zeta \in \mathscr{D}} Y[\zeta]^2 - n^2}{n^2}.$$

If $\mathscr{F}$ is a single-replicate OA (i.e. $Y[\zeta] \in \{0, 1\}$, $\zeta \in \mathscr{D}$) then

$$A_m(\mathscr{F}) = \frac{\#\mathscr{D} - n}{n}.$$

Proof Let us consider $\mathscr{F} \in OA(n, s_1 \ldots s_m, m - 1)$. Then

$$A_0(\mathscr{F}) = 1, A_1(\mathscr{F}) = \cdots = A_{m-1}(\mathscr{F}) = 0\,.$$

From item 3. of Proposition 4 we get

$$A_m(\mathscr{F}) = \sum_{j=0}^{m} A_j(\mathscr{F}) - \sum_{j=0}^{m-1} A_j(\mathscr{F}) = \\ = \frac{\#\mathscr{D}\sum_{\zeta\in\mathscr{D}} Y[\zeta]^2}{(\#\mathscr{F})^2} - 1 = \frac{\#\mathscr{D}\sum_{\zeta\in\mathscr{D}} Y[\zeta]^2 - (\#\mathscr{F})^2}{(\#\mathscr{F})^2}\,.$$

In the special case $Y[\zeta] \in \{0, 1\}$, $\zeta \in \mathscr{D}$ we get

$$A_m(\mathscr{F}) = \frac{\#\mathscr{D} - \#\mathscr{F}}{\#\mathscr{F}}\,.$$

□

We obtain a lower bound for $A_m(\mathscr{F})$ as in Theorem 5 of [11].

Proposition 6 *Let $\mathscr{F} \in OA(n, s_1 \ldots s_m, m-1)$. Then*

$$A_m(\mathscr{F}) \geq \frac{r(\#\mathscr{D} - r)}{n^2}\,,$$

where q and r are the quotient and the remainder when n is divided by $\#\mathscr{D}$, $n = q\#\mathscr{D} + r$ (and $q = 0$ when $n < \#\mathscr{D}$).

Proof From Proposition 5 we know that

$$A_m(\mathscr{F}) = \frac{\#\mathscr{D}\sum_{\zeta\in\mathscr{D}} Y[\zeta]^2 - n^2}{n^2}\,.$$

If we divide n by $\#\mathscr{D}$ we can write $n = q\#\mathscr{D} + r$. The counting vector $\tilde{Y}$ that minimizes $\sum_{\zeta\in\mathscr{D}} Y[\zeta]^2$ must be defined as

$$\tilde{Y}[\zeta] = \begin{cases} q+1 & \text{if } \zeta \in B_r \\ q & \text{if } \zeta \in \mathscr{D} - B_r \end{cases}$$

where B_r is any subset of $\mathscr{D}$ with r points. We obtain

$$\sum_{\zeta\in\mathscr{D}} \tilde{Y}[\zeta]^2 = \#\mathscr{D}q^2 + 2rq + r\,.$$

It follows that

$$A_m(\mathscr{F}) \geq \frac{\#\mathscr{D}(\#\mathscr{D}q^2 + 2rq + r) - (q\#\mathscr{D} + r)^2}{(q\#\mathscr{D} + r)^2}\,.$$

By simple algebra we obtain

$$A_m(\mathscr{F}) \geq \frac{r(\#\mathscr{D} - r)}{(\#\mathscr{F})^2}\,.$$

□

When we consider $m > t+1$ factors a lower bound for $A_{t+1}(\mathscr{F})$ can be obtained by summing up all the lower bounds that are obtained using Proposition 6 for all the $\binom{m}{t+1}$ subsets of $t+1$ factors of $\mathscr{D}_1, \ldots, \mathscr{D}_m$.

3 The GWLP of the Union of Fractions

In this section we analyze the behavior of the aberrations (and thus of the GWLP) of a fraction obtained by merging two or more fractions. In particular we focus on OAs which can be expressed as the union of other OAs.

First, it is worth noting that given a fraction $\mathscr{F}$ with counting function $R(\zeta)$, we can consider a fraction $\nu\mathscr{F}$ obtained by replicating ν times each design point of $\mathscr{F}$. In such a case, it is immediate to check that the counting function of $\nu\mathscr{F}$ is simply $\nu R(\zeta)$, and therefore all aberrations remain unchanged:

$$a_\alpha^{(\nu R)} = a_\alpha^{(R)}, \quad \text{for all } \alpha \in L\,.$$

In the following proposition we consider the union of k fractions, $k \geq 2$.

Proposition 7 *Let us consider fractions $\mathscr{F}_1, \ldots, \mathscr{F}_k$ with $n_1, \ldots, n_k$ design points, respectively. Let us denote by $R_i = \sum_{\alpha\in L} c_\alpha^{(i)} X^\alpha$ the counting function of $\mathscr{F}_i$, $i = 1, \ldots, k$, by $\mathscr{F}$ the union $\mathscr{F} = \mathscr{F}_1 \cup \cdots \cup \mathscr{F}_k$, by $R = \sum_{i=1}^k R_i = \sum_{\alpha\in L} c_\alpha^{(R)} X^\alpha$ the counting function of $\mathscr{F}$ and by n the size of $\mathscr{F}$, $n = n_1 + \ldots + n_k$.*

The jth element of the GWLP of $\mathscr{F}$ is

$$A_j(\mathscr{F}) = \sum_{i=1}^{k} \frac{n_i^2}{n^2} A_j(\mathscr{F}_i) + 2\frac{(\#\mathscr{D})^2}{n^2} \sum_{i_1<i_2} \sum_{|\alpha|_0=j} \mathrm{Re}\left(c_\alpha^{(i_1)} \overline{c}_\alpha^{(i_2)}\right), \quad j = 0, \ldots, m\,. \tag{2}$$

Proof Let us consider $k = 2$, i.e. $\mathscr{F} = \mathscr{F}_1 \cup \mathscr{F}_2$. The aberration $a_\alpha^{(R)}$ is

$$a_\alpha^{(R)} = \frac{\left(\|c_\alpha^{(1)} + c_\alpha^{(2)}\|_2\right)^2}{\left(c_0^{(1)} + c_0^{(2)}\right)^2}\,.$$

We obtain

$$\begin{aligned}\left(\left\|c_\alpha^{(1)}+c_\alpha^{(2)}\right\|_2\right)^2 &= \left(\left\|c_\alpha^{(1)}\right\|_2\right)^2+\left(\left\|c_\alpha^{(2)}\right\|_2\right)^2+2\,\mathrm{Re}\left(c_\alpha^{(1)}\overline{c}_\alpha^{(2)}\right)= \\ &= \left(\frac{n_1}{\#\mathscr{D}}\right)^2 a_\alpha^{(1)}+\left(\frac{n_2}{\#\mathscr{D}}\right)^2 a_\alpha^{(2)}+2\,\mathrm{Re}\left(c_\alpha^{(1)}\overline{c}_\alpha^{(2)}\right)= \\ &= \frac{1}{(\#\mathscr{D})^2}\left(n_1^2 a_\alpha^{(1)}+n_2^2 a_\alpha^{(2)}+2(\#\mathscr{D})^2\,\mathrm{Re}\left(c_\alpha^{(1)}\overline{c}_\alpha^{(2)}\right)\right)\end{aligned}$$

where $a_\alpha^{(i)}$ refers to $\mathscr{F}_i$, i = 1,2. We also obtain

$$\left(c_0^{(1)}+c_0^{(2)}\right)^2=\left(\frac{n_1}{\#\mathscr{D}}+\frac{n_2}{\#\mathscr{D}}\right)^2=\frac{n^2}{(\#\mathscr{D})^2}\,.$$

It follows

$$a_\alpha^{(R)}=\frac{1}{n^2}\left(n_1^2 a_\alpha^{(1)}+n_2^2 a_\alpha^{(2)}+2(\#\mathscr{D})^2\,\mathrm{Re}\left(c_\alpha^{(1)}\overline{c}_\alpha^{(2)}\right)\right)$$

and

$$\begin{aligned}A_j(\mathscr{F}) &= \sum_{|\alpha|_0=j} a_\alpha^{(R)}= \\ &= \left(\frac{n_1}{n}\right)^2 A_j(\mathscr{F}_1)+\left(\frac{n_2}{n}\right)^2 A_j(\mathscr{F}_2)+2\left(\frac{\#\mathscr{D}}{n}\right)^2\sum_{|\alpha|_0=j}\mathrm{Re}\left(c_\alpha^{(1)}\overline{c}_\alpha^{(2)}\right)\end{aligned}$$

for $j=0,1,\ldots,m$.

The generalization of this formula to the case $k>2$ is straightforward. □

In case of two-level designs, $c_\alpha\in\mathbb{R}$ and thus Eq. (2) becomes

$$A_j(\mathscr{F})=\sum_{i=1}^{k}\frac{n_i^2}{n^2}A_j(\mathscr{F}_i)+2\frac{(\#\mathscr{D})^2}{n^2}\sum_{i_1<i_2}\sum_{|\alpha|_0=j}c_\alpha^{(i_1)}c_\alpha^{(i_2)},\quad j=0,\ldots,m\,.$$

The term $\sum_{|\alpha|_0=j}c_\alpha^{(i_1)}c_\alpha^{(i_2)}$ can be viewed as a kind of covariance between the coefficients of order j of the two counting functions R_{i_1} and R_{i_2}.

To illustrate the use of Proposition 7 on a very small example, let us consider the two regular fractions of the 2^3 design, whose union is the full-factorial:

$$\mathscr{F}_1=X_1X_2X_3=-1\qquad R_1=\frac{1}{2}(1-X_1X_2X_3)\,;$$

$$\mathscr{F}_2=X_1X_2X_3=+1\qquad R_2=\frac{1}{2}(1+X_1X_2X_3)\,.$$

In this case we have

$$A_0(\mathscr{F}_1) = 1, A_1(\mathscr{F}_1) = A_2(\mathscr{F}_1) = 0, A_3(\mathscr{F}_1) = 1\,;$$
$$A_0(\mathscr{F}_2) = 1, A_1(\mathscr{F}_2) = A_2(\mathscr{F}_2) = 0, A_3(\mathscr{F}_2) = 1\,.$$

As expected we obtain $A_0(\mathscr{F}) = 1$, $A_1(\mathscr{F}) = A_2(\mathscr{F}) = 0$ and

$$A_3(\mathscr{F}) = \left(\frac{4}{8}\right)^2 A_3(\mathscr{F}_1) + \left(\frac{4}{8}\right)^2 A_3(\mathscr{F}_2) + 2\left(\frac{8}{4}\right)^2 c_{111}^{(1)} c_{111}^{(2)} = 0$$

because $c_{111}^{(1)} = -1/2$ and $c_{111}^{(2)} = 1/2$.

4 The Hilbert Basis for Orthogonal Arrays

In this section we define the set $OA(\bullet, \mathscr{D}, t)$ of all the OAs with strength t of the full design $\mathscr{D}$ and we study its combinatorial and geometric properties. With respect to the standard notation, we allow the cardinality to vary, because our study will concern the union of two or more OAs, and thus we use the symbol $\bullet$ in place of the cardinality of the fraction. In the case of binary designs, this set has already been considered in [2], where the reader can find also a simple and comprehensive summary of the basic definitions from Combinatorics used here. The generalization to mixed-level designs can be found in [6].

As a preliminary remark, notice that to the set $OA(\bullet, \mathscr{D}, t)$ can be associated in a natural way the set of the corresponding counting functions. With as slight abuse of notation, we use the same notation for both these sets.

Lemma 1 *The set $OA(\bullet, \mathscr{D}, t)$ can be written in the form*

$$OA(\bullet, \mathscr{D}, t) = C \cap \mathbb{N}^{\#\mathscr{D}} \tag{3}$$

where C is a polyhedral cone in $\mathbb{R}^{\#\mathscr{D}}$.

Proof Recall that a subset of $\mathbb{R}^k$ is a cone if for all $x, y \in C$ and for all $\lambda, \mu \in \mathbb{R}$ we have $\lambda x + \mu y \in C$, and it is a polyhedral cone if in addition it can be written in the form

$$C = \left\{x \in \mathbb{R}^k \;:\; Ax \geq 0\right\}\,. \tag{4}$$

In this setting it is enough to define the matrix A in such a way all the t-marginals of x are constant (i.e., the difference of any two elements in a t-marginal is equal to 0). □

In Combinatorics, objects like $OA(\bullet, \mathscr{D}, t)$ expressed as the lattice points of a cone as in Eq. (3) are widely studied. See, e.g., Chap. 6 in [13] for a general introduction

to semigroups, lattice ideals, and Hilbert bases. In this paper, we focus on the notion of Hilbert basis of a lattice, and we specialize its definition.

Definition 6 A Hilbert basis of $OA(\bullet, \mathscr{D}, t)$ is an inclusion-minimal finite set of OAs $\mathscr{B}_1, \ldots, \mathscr{B}_r$ such that each OA $\mathscr{F} \in OA(\bullet, \mathscr{D}, t)$ is

$$\mathscr{F} = c_1\mathscr{B}_1 + \cdots + c_r\mathscr{B}_r$$

with coefficients $c_1, \ldots, c_r \in \mathbb{N}$.

Under mild conditions, which are satisfied by $OA(\bullet, \mathscr{D}, t)$, the Hilbert basis exists and is unique.

The Hilbert basis of $OA(\bullet, \mathscr{D}, t)$ depends on the matrix A in Eq. (4), which in turn depends on the t-marginals of the OAs. Thus, we have a different Hilbert basis for different $\mathscr{D}$ and t. From the computational point of view, there are specific algorithms to efficiently compute Hilbert bases. Such algorithms are available by means of specialized software. Currently, two choices are available: `4ti2`, see [17], and the more recent package `normaliz`, see [1]. For our purpose, the use of one or the other software is equivalent. In our examples, we have used `4ti2`, but the use of both these software is very easy. It is enough to input the matrix A defining the polyhedral cone and the software returns the corresponding Hilbert basis.

Using the elements of the Hilbert basis, we can build all OAs of any given sample size. As noticed in the Introduction, the limitation of our approach is due to the fact the computation of Hilbert bases is very intensive and the computational cost grows very fast when the full design becomes large. Therefore, the computations are limited to relatively small cases, which are to be considered as illustrative examples.

5 Computations

We consider OAs of strength 2 for 5 factors, each with 2 levels, $OA(\bullet, 2^5, t)$. The Hilbert Basis for this problem contains 26, 142 different elements which can be classified according to their size as reported in Table 1.

First, we focus on the OAs of size 16. There are 162 OAs of size 16 in the Hilbert Basis. The remaining 16-run OAs can be generated considering all possible unions of two OAs of size 8. We denote these OAs as $(8 + 8)$-run OAs. There are 60 8-run OAs and therefore $60 + \binom{60}{2} = 1{,}830$ possible different $(8 + 8)$-run OAs. We find 1,770 different $(8 + 8)$-run OAs. The classification of the $162 + 1{,}770 = 1{,}932$ OAs of size 16 according to the values of $A_3(\mathscr{F})$ is reported in Table 2.

From Table 2 we immediately see that there are 12 designs with $A_3(\mathscr{F}) = 0$. We can choose the best design(s) among these 12 fractions. We find two OAs of the 16-run type for which $A_1(\mathscr{F}) = A_2(\mathscr{F}) = A_3(\mathscr{F}) = A_4(\mathscr{F}) = 0$ and $A_5(\mathscr{F}) = 1$.

As a second example, we consider OAs with 20 runs. There are 960 OAs of size 20 in the Hilbert Basis. The remaining 20-run OAs can be generated by considering

Table 1 The elements of the Hilbert basis for $OA(\bullet, 2^5, 2)$ classified with respect to their sample size

Size	N
8	60
12	224
16	162
20	960
24	7680
28	8384
32	5760
36	2912

Table 2 Distribution of $OA(16, 2^5, 2)$ with respect to $A_3(\mathscr{F})$

	$A_3(\mathscr{F})$							
Type	0	0.25	0.5	0.75	1	1.5	2	Total
16-run	2	80	0	80	0	0	0	162
(8 + 8)-run	10	0	240	0	1,220	240	60	1,770

Table 3 Distribution of $OA(20, 2^5, 2)$ with respect to $A_3(\mathscr{F})$

	$A_3(\mathscr{F})$			
Type	0.4	0.72	1.04	Total
20-run	480	0	480	960
(8 + 12)-run	1,632	4,800	3,360	9,792

all possible unions of two OAs, one of size 8 and one of size 12. We denote these OAs as (8 + 12)-run OAs. There are 60 8-run OAs and 224 12-run OAs and therefore $60 \cdot 224 = 13{,}440$ possibly different (8 + 12)-run OAs. We find 9,792 different (8 + 12)-run OAs. The classification of the $960 + 9{,}792 = 10{,}752$ OAs of size 20 according to the values of $A_3(\mathscr{F})$ is reported in Table 3.

If we proceed as we did for OAs of size 16, focusing on the 2,112 OAs with $A_3 = 0.4$, we find 192 GMA-optimal OAs. These are of the (8 + 12)-run type and their Word Length Pattern is $A_1(\mathscr{F}) = A_2(\mathscr{F}) = 0$, $A_3(\mathscr{F}) = 0.4$, $A_4(\mathscr{F}) = 0.2$ and $A_5(\mathscr{F}) = 0$.

Acknowledgements Both authors are partially supported by a INdAM GNAMPA 2017 project. RF acknowledges that the present research has been partially supported by MIUR grant Dipartimenti di Eccellenza 2018-2022 (E11G18000350001)

References

1. Bruns, W., Ichim, B., Römer, T., Sieg, R., Söger, C.: Normaliz algorithms for rational cones and affine monoids. https://www.normaliz.uni-osnabrueck.de
2. Carlini, E., Pistone, G.: Hilbert bases for orthogonal arrays. J. Stat. Theory Pract. **1**(3–4), 299–309 (2007)
3. Cheng, S.W., Ye, K.Q.: Geometric isomorphism and minimum aberration for factorial designs with quantitative factors. Ann. Stat. **32**(5), 2168–2185 (2004)
4. Dey, A., Mukerjee, R.: Fractional Factorial Plans. Wiley, New York (2009)
5. Fontana, R.: Counting vectors for orthogonal fractional factorial design generation. AIP Conf. Proc. **1368**(1), 327–330 (2011)
6. Fontana, R.: Algebraic generation of minimum size orthogonal fractional factorial designs: an approach based on integer linear programming. Comput. Stat. **28**(1), 241–253 (2013)
7. Fontana, R.: Generalized minimum aberration mixed-level orthogonal arrays: A general approach based on sequential integer quadratically constrained quadratic programming. Commun. Stat. Theory Methods **46**(9), 4275–4284 (2017)
8. Fontana, R., Pistone, G., Rogantin, M.P.: Classification of two-level factorial fractions. J. Stat. Plann. Inference **87**(1), 149–172 (2000)
9. Fontana, R., Rapallo, F., Rogantin, M.P.: Aberration in qualitative multilevel designs. J. Stat. Plann. Inference **174**, 1–10 (2016)
10. Fries, A., Hunter, W.G.: Minimum aberration 2^{k-p} designs. Technometrics **22**(4), 601–608 (1980)
11. Grömping, U., Xu, H.: Generalized resolution for orthogonal arrays. Ann. Stat. **42**(3), 918–939 (2014)
12. Hedayat, A.S., Sloane, N.J.A., Stufken, J.: Orthogonal Arrays: Theory and Applications. Springer, New York (2012)
13. Miller, E., Sturmfels, B.: Combinatorial Commutative Algebra. Springer, New York (2006)
14. Mukerjee, R., Wu, C.F.J.: A Modern Theory of Factorial Design. Springer, New York (2007)
15. Pistone, G., Rogantin, M.P.: Indicator function and complex coding for mixed fractional factorial designs. J. Stat. Plann. Inference **138**(3), 787–802 (2008)
16. Xu, H., Wu, C.F.J.: Generalized minimum aberration for asymmetrical fractional factorial designs. Ann. Stat. **29**(4), 1066–1077 (2001)
17. 4ti2 team: 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. www.4ti2.de

A Copula-Based Hidden Markov Model for Toroidal Time Series

Francesco Lagona

Abstract Toroidal time series are temporal sequences of bivariate angular observations that often arise in environmental and ecological studies. A hidden Markov model is proposed for segmenting these data according to a finite number of latent classes, associated with copula-based toroidal densities. The model conveniently integrates circular correlation, multimodality and temporal auto-correlation. A computationally efficient EM algorithm is proposed for parameter estimation. The proposal is illustrated on a time series of wind and sea wave directions.

Keywords Copula · Hidden Markov model · Segmentation · Toroidal data

1 Introduction

Bivariate sequences of angles are often referred to as toroidal time series, because the pair of two angles can be represented as a point on a torus. These data often arise in environmental and ecological studies. Examples include time series of wind and wave directions [9], time series of wind mean directions and directions of the maximum gust observed each day [2] and time series of turning angles in studies of animal movement [12].

The analysis of toroidal time series is complicated by the difficulties in modeling the dependence between angular measurements over time [8]. An additional complication is given by the multimodality of the marginal distribution of the data, because environmental toroidal data are observed under time-varying heterogeneous conditions.

This paper introduces a hidden Markov model (HMM) that simultaneously accounts for dependence across circular measurements, temporal auto-correlation, multimodality and latent time-varying heterogeneity. Under this model, the distribution of toroidal data is approximated by a mixture of copula-based toroidal den-

F. Lagona (✉)
University of Roma Tre, via G. Chiabrera 199, Rome, Italy
e-mail: francesco.lagona@uniroma3.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_32

sities, whose parameters depend on the evolution of a latent Markov chain. While the copula-based toroidal density accommodates dependence between two circular variables, a mixture of copula-based densities allows for multimodality and, finally, a latent Markov chain accounts for temporal correlation and, simultaneously, for time-varying heterogeneity.

Following an approach that has been recently suggested to model time series of cylindrical data [7], this model extends previous proposals that are either based on mixtures of conditionally independent circular densities [11] or based on mixtures of bivariate von Mises densities [1, 10] It provides an intuitively appealing framework where the data are modeled by integrating conventional tools of statistical analysis: a copula, a mixture and a Markov chain. It is furthermore numerically tractable, by exploiting a suitable Expectation Maximization (EM) algorithm for parameter estimation.

The rest of the paper is organized as follows. Section 2 introduces the proposed model. Section 3 is devoted to maximum likelihood parameter estimation and Sect. 4 illustrates the proposal on a case study of wave and wind directions. Relevant points of discussion are finally summarized in Sect. 5.

2 A Copula-Based Toroidal Hidden Markov Model

Let $\mathbf{z} = (x, y)$ be a pair of angles, $x, y \in [0, 2\pi)$. Moreover, let $f(x; \alpha)$ and $f(y; \beta)$ be two circular densities, respectively known up to the parameters α and β. Further, let $F(x; \alpha)$ and $F(y; \beta)$ be the two cumulative distribution functions of x and y, defined with respect to a fixed, although arbitrary, origin. Finally, let $g(u; \gamma)$, $u \in [0, 2\pi)$ be a parametric circular density, known up to a parameter γ. Then,

$$f_q(\mathbf{z}; \theta) = 2\pi g\left(2\pi \left(F(x; \alpha) - qF(y; \beta)\right) f(x; \alpha)\right) f(y; \beta)) \quad q = \pm 1 \tag{1}$$

is a parametric toroidal density with support $[0, 2\pi)^2$, known up to the parameter vector $\theta = (\alpha, \beta, \gamma)$, having the marginal densities $f(x; \alpha)$ and $f(y; \beta)$ [3]. Equation (1) is a typical example of a copula-based construction of a bivariate density, obtained by decoupling the margins from the joint distribution. When the binding density g is the uniform circular distribution, say $g(x) = (2\pi)^{-1}$, then Eq. (1) reduces to the product of the marginal densities. Otherwise, the dependence between x and y is captured by the concentration of g: when g is highly concentrated, the dependence is high; when g is more diffuse, dependence is low. Finally, the constant $q = \pm 1$ determines whether the dependence between x and y is positive ($q = 1$) or negative ($q = -1$). Additional details on copula-based methods that use a circular binding density to specify bivariate and multivariate densities can be found in [4].

The proposed hidden Markov model can be described as a dynamic mixture of copula-based toroidal densities. To illustrate, let $\mathbf{z} = (\mathbf{z}_t, t = 1, \ldots, T)$, $\mathbf{z}_t = (x_t, y_t)$, $x_t, y_t \in [0, 2\pi)$, be a toroidal time series. We assume that the distribution of the data is driven by the evolution of an unobserved Markov chain with K states,

which represents (time-varying) latent classes and can be specified as a sequence $\mathbf{u} = (\mathbf{u}_t, t = 1, \ldots, T)$ of multinomial variables $\mathbf{u}_t = (u_{t1} \ldots u_{tK})$ with one trial and K classes, whose binary components represent class membership at time t. The joint distribution $p(\mathbf{u}; \pi)$ of the chain is fully known up to a parameter π that includes K initial probabilities $\pi_k = P(u_{1k} = 1), k = 1, \ldots, K, \sum_k \pi_k = 1$, and K^2 transition probabilities $\pi_{hk} = P(u_{tk} = 1 | u_{t-1,h} = 1), h, k = 1, \ldots, K, \sum_k \pi_{hk} = 1$. Formally, we assume that

$$p(\mathbf{u}; \pi) = \prod_{k=1}^{K} \pi_k^{u_{1k}} \prod_{t=2}^{T} \prod_{h=1}^{K} \prod_{k=1}^{K} \pi_{hk}^{u_{t-1,h} u_{tk}}. \tag{2}$$

The specification of the HMM is completed by assuming that the observations are conditionally independent, given a realization of the Markov chain. As a result, the conditional distribution of the observed process, given the latent process, takes the form of a product density, say

$$f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K) = \prod_{t=1}^{T} \prod_{k=1}^{K} f(\mathbf{z}_t; \theta_k)^{u_{tk}}, \tag{3}$$

where $f(\mathbf{z}; \theta_k)$, $k = 1, \ldots, K$ are the K cylindrical densities defined by (1) and known up to a vector of parameters θ_k.

The likelihood function of the model is therefore obtained by integrating the joint density of the observed data and the unobserved class memberships with respect to the segmentation $\mathbf{u}$, namely

$$L(\pi, \theta; \mathbf{z}) = \sum_{\mathbf{u}} p(\mathbf{u}; \pi) f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K). \tag{4}$$

By computing the maximum likelihood estimate $\hat{\theta}$, the cylindrical time series can be then segmented according to the posterior probabilities of class membership

$$\hat{\pi}_{tk} = P(u_{tk} = 1 \mid \mathbf{z}; \hat{\theta}), \tag{5}$$

based on $\hat{\theta}$. More precisely, the observation at time t can be allocated to class $k^\star$ if $\hat{\pi}_{tk^\star} \geq \hat{\pi}_{th}$, for each $h = 1 \ldots K$ (maximum a posterior, MAP, allocation).

When the transition probability matrix has equal rows, the model reduces to a mixture model where observations are clustered by ignoring the information redundancy that is due to temporal correlation. In general, the proposed HMM segments the series by accounting not only for similarities in the variable space but also in a temporal neighborhood.

3 Parameter Estimation

An EM algorithm can be exploited to maximize the likelihood function (4). It is based on the following complete-data log-likelihood function

$$\log L_{\text{comp}}(\theta, \mathbf{u}, \mathbf{z}) = \sum_{k=1}^{K} u_{1k} \log \pi_k + \sum_{t=2}^{T} \sum_{h=1}^{K} \sum_{k=1}^{K} u_{t-1,h} u_{t,k} \log \pi_{hk} + \sum_{t=1}^{T} \sum_{k=1}^{K} u_{tk} \log f(\mathbf{z}_t; \theta_k). \tag{6}$$

The algorithm is iterated by alternating an expectation (E) and a maximization (M) step. Given the estimates $\hat{\pi}_s$ and $\hat{\theta}_s$, obtained at the end of the sth iteration, the $(s+1)$th iteration is initialized by the E-step, which evaluates the expected value of the complete data log-likelihood (6) with respect to the conditional distribution of the missing values u_{tk} given the observed data.

The E step reduces to the computation of the univariate posterior probabilities of each latent state at time t, $\hat{\pi}_{tk} = P(u_{tk} = 1 \mid \mathbf{z}, \hat{\pi}_s, \hat{\theta}_s) \quad k = 1 \ldots K, t = 1 \ldots T$, and the computation of the bivariate posterior probabilities of each pair of states in two adjacent times, say $\hat{\pi}_{t-1,t,hk} = P(u_{t-1,h} = 1, u_{tk} = 1 \mid \mathbf{z}, \hat{\pi}_s, \hat{\theta}_s) \quad h, k = 1 \ldots K, t = 2 \ldots T$. The task of computing these posterior probabilities from an estimate $(\hat{\pi}_s, \hat{\theta}_s)$ is generally referred to as the HMM-smoothing numerical issue and it is typically solved by specifying the posterior probabilities in terms of suitably normalized functions, which can be computed recursively, avoiding unpractical summations over the state space of latent Markov chain and numerical under- and over-flows. In this paper, we exploited the HMM-smoothing algorithm that is described by [1].

The M-step of the algorithm updates the estimate $(\hat{\pi}_s, \hat{\theta}_s)$ with a new estimate $(\hat{\pi}_{s+1}, \hat{\theta}_{s+1})$, by maximizing the expected value of the complete data log-likelihood (6), obtained from the previous E step. This expected value is the sum of functions that depend on independent sets of parameters and can therefore be maximized separately. Maximization with respect to the transition probabilities π_{hk}, under the constraints $\sum_{k=1}^{K} \pi_{hk} = 1, h = 1 \ldots K$, provides the closed-form updating formula

$$\hat{\pi}_{hk(s+1)} = \frac{\sum_{t=1}^{T} \hat{\pi}_{t-1,t,hk}(\hat{\pi}_s, \hat{\theta}_s)}{\sum_{t=1}^{T} \hat{\pi}_{t-1,h}(\hat{\pi}_s, \hat{\theta}_s)}, \qquad h, k = 1, \ldots, K.$$

Maximization with respect to the parameters θ_k of the kth copula-based cylindrical components reduces to maximize

$$\sum_{t=1}^{T} \hat{\pi}_{tk} f(\mathbf{z}_t; \theta_k). \tag{7}$$

We can maximize (7) with respect to all the parameters or, more efficiently, we can take a IFM (inference function for margins [6]) approach. Precisely, (7) can be written as the sum of three components, namely

$$\sum_{t=1}^{T} \hat{\pi}_{tk} f(\mathbf{z}_t; \theta_k) = \sum_{t=1}^{T} \hat{\pi}_{tk} \log g(2\pi(F(x_t; \alpha) - q F(y_t; \beta)); \gamma) \tag{8}$$

$$+ \sum_{t=1}^{T} \hat{\pi}_{tk} f(x_t; \alpha_k) \tag{9}$$

$$+ \sum_{t=1}^{T} \hat{\pi}_{tk} f(y_t; \beta_k) \tag{10}$$

Accordingly, IFM proceeds by finding the parameter values $\hat{\alpha}$ and $\hat{\beta}$ that respectively maximize (9) and (10) and then maximizing function (8), evaluated at $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$, to obtain an estimate of γ.

The procedure outlined above does not produce confidence intervals of the estimates, which however can be computed by taking a parametric bootstrap approach, by re-fitting the model from a number R of bootstrap samples, simulated from the estimated model parameters, and computing, for example, the 2.5% and the 97.5% quantiles of the empirical distribution of each bootstrap estimate.

Simulation of the model is straightforward. First, a sequence of states is simulated from a Markov chain with the desired transition probabilities, by repeatedly drawing samples from a multinomial distribution with K states. Given a sequence of states, a toroidal observation at time t is obtained by exploiting one of the algorithm suggested by [4].

4 Application

The proposed methods have been implemented to segment a time series of $T = 1326$ semi-hourly wind and wave directions, taken in wintertime by the buoy of Ancona, which is located in the Adriatic Sea at about 30 km from the coast. Figure 1 displays the scatter plot of the data. Point coordinates indicate the direction (in radians) from which winds blow and waves travel. For simplicity, these bivariate observations are plotted on the plane, although data points are actually on a torus. The interpretation of these data is not easy. While in the ocean wind and wave directions are strongly correlated, this is not necessarily true in the Adriatic Sea, due to the complex orography of the basin. The Adriatic Sea is a semienclosed, long narrow basin, extending for about 800 km along the major axis from SE to NW, with a width of about 200 km. In winter, relevant wind events in the Adriatic Sea are typically generated by the Bora wind, which in the Ancona area blows north northwesterly along the major axis of the basin, and by the Sirocco wind, which blows

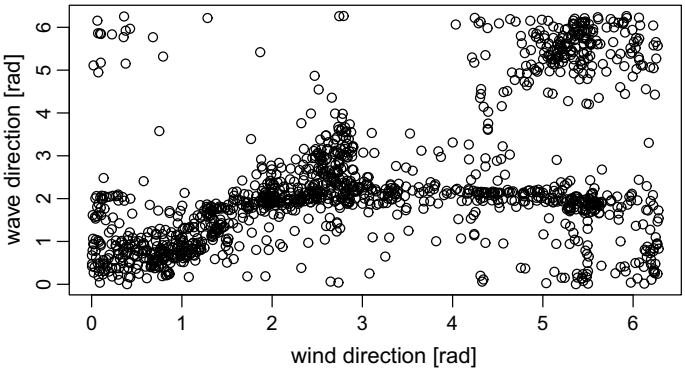


Fig. 1 Wave directions and heights, as observed by the buoy of Ancona in wintertime (0, $\pi/2$, π and $3\pi/4$ respectively indicate North, East, South and West). For simplicity, the data are plotted on the plane, although they are points on the torus $[0, 2\pi)^2$

southeasterly. Waves generated by these winds travel in the same direction of the winds or slightly rotate along the major axis of the basin. In addition, there are winds that blow northwesterly, westerly and southwesterly from the Italian coast, along the minor axis of the basin. Coastal winds generate synchronized waves only when the waves travel unobstructed, that is, either northwesterly or southeasterly, along the major axis of the basin. In the case of western winds, waves travel southwesterly. When, however, coastal winds rotate clockwise, waves tend to travel from north. This explains the clusters shown in Fig. 1 and suggests the occurrence of a number of latent wind wave regimes. Estimation of an HMM from these data can be helpful in clustering the data into a number of toroidal clusters, each associated with a specific wind-wave distribution.

The proposed HMM requires a parametric specification of the toroidal density (1), which reduces to the choice of the binding density g and the choice of the marginal densities $f(x;\alpha)$ and $f(y;\beta)$ that respectively model the marginal distribution of the wind and wave direction.

However, depending on the choice of the binding density, the density (1) can be multimodal [4]. Using multimodal densities in segmentation and classification problems, such as the one motivating this paper, may unnecessarily complicate the interpretation of the results. Unimodal densities can however be obtained by using the wrapped Cauchy as a binding density g [4].

Accordingly, for this study, the binding density has been specified as a centered wrapped Cauchy

$$g(u;\gamma) = \frac{1}{2\pi}\frac{1-\gamma^2}{1+\gamma^2-2\gamma\cos(u)} \quad u \in [0, 2\pi).$$

This circular density depends on a single concentration parameter $\gamma \in [0, 1)$ and reduces to the uniform circular density when $\gamma = 0$.

Wrapped Cauchy densities that include additional location parameters α_1 and β_1 have been instead exploited to model the marginal distributions of wind and wave direction, say

$$f(x;\alpha) = \frac{1}{2\pi}\frac{1-\alpha_2^2}{1+\alpha_2^2-2\alpha_2\cos(y-\alpha_1)} \quad x \in [0, 2\pi) \tag{11}$$

$$f(y;\beta) = \frac{1}{2\pi}\frac{1-\beta_2^2}{1+\beta_2^2-2\beta_2\cos(y-\beta_1)} \quad y \in [0, 2\pi) \tag{12}$$

The proposed toroidal density is therefore obtained by taking a wrapped Cauchy density that binds wrapped Cauchy marginals, a model known as the bivariate wrapped Cauchy model [5].

A number of models have been estimated from these data, by varying the number K of components from 2 to 5, and associating each component with $q = \pm 1$. The BIC statistic suggested to segment the data according to 4 regimes that are respectively associated with $q = 1, 1, -1, 1$. Table 1 displays the estimates under these four latent states, along with bootstrap percentiles, computed by simulating 400 samples.

Table 1 displays the estimates of the parameters of the 4 toroidal densities and Fig. 2 shows the shapes of the related distributions and the segmented observations. We can observe that the estimated transition probability matrix (Table 1) is essentially diagonal, suggesting that the assumption of independent samples (i.e. a transition probability matrix with equal rows) is, in this example, unrealistic. The model clusters the data into well-separated groups, which can be interpreted as latent wind wave regimes. Components 1 and 4 are, respectively, associated with Bora and Sirocco events. In the Ancona area, Bora blows north northeasterly along the major axis of the basin, while Sirocco blows southeasterly. Waves generated by these winds travel in the same direction of the winds or slightly rotate along the major axis of the basin. Components 2 and 3 are instead associated with coastal winds, which generate waves that tend to travel along the major axis of the basin. As a result, waves travel in a direction that is weakly correlated with the wind direction. Overall, the model describes the plasticity of the wind wave interaction in the Adriatic Sea, indicating that the joint distribution of wind and wave data changes under different environmental regimes. Regime switching changes not only the modal directions and concentrations around these modes but also, and more interestingly, the correlation structure of the data. As a result, on the one side, the (marginal) weak correlation between wind and wave directions is explained by the presence of coastal winds (component 1). On the other side, the model indicates that the wind direction is an accurate predictor of the wave direction during Bora and Sirocco episodes, but that the level of accuracy decreases in the presence of coastal winds. In summary, wind directions should not be used to predict wave directions, without accounting for the latent, environmental heterogeneity of the data under study.

The rows at the bottom of Table 1 include the estimated transition probabilities of the latent Markov chain. The transition probability matrix is essentially diagonal, reflecting the temporal persistence of the classes. Such a persistence is shown by Fig. 3, which displays the posterior probabilities $\hat{\pi}_{tk}$ that have been obtained at the

Table 1 Parameter estimates and bootstrap quantiles of a 4-state toroidal hidden Markov model

State 1	Parameter	Estimate	2.5% Quantile	97.5% Quantile
Wind	location	0.796	0.654	0.863
	concentration	0.693	0.682	0.718
Wave	location	0.821	0.654	0.946
	concentration	0.780	0.682	0.788
Copula	dependence	0.283	0.115	0.354
	q	1		
State 2	Parameter	Estimate	2.5% Quantile	97.5% Quantile
Wind	location	5.367	4.981	6.091
	concentration	0.764	0.755	0.899
Wave	location	5.681	4.654	6.263
	concentration	0.653	0.622	0.718
Copula	dependence	0.203	0.090	0.210
	q	1		
State 3	Parameter	Estimate	2.5% Quantile	97.5% Quantile
Wind	location	5.255	3.112	6.006
	concentration	0.520	0.210	0.656
Wave	location	1.975	1.254	2.063
	concentration	0.863	0.782	0.901
Copula	dependence	0.368	0.217	0.375
	q	–1		
State 4	Parameter	Estimate	2.5% Quantile	97.5% Quantile
Wind	location	2.546	2.112	2.916
	concentration	0.732	0.110	0.956
Wave	location	2.129	1.954	2.763
	concentration	0.763	0.682	0.798
Copula	dependence	0.079	0.017	0.125
	q	1		
Destination				
Origin	State 1	State 2	State 3	State 4
State 1	0.972	0.007	0.000	0.021
State 2	0.015	0.962	0.012	0.010
State 3	0.015	0.029	0.943	0.013
State 4	0.000	0.000	0.031	0.969

convergence of the EM algorithm. The adopted segmentation model hence confirms that the sea surface in the study area tends to alternate relevant marine events with periods of good sea conditions.

Figure 2 displays the contour plots of the 4 toroidal densities (right side) and the scatterplot of the points (left side), colored with grey levels that are proportional

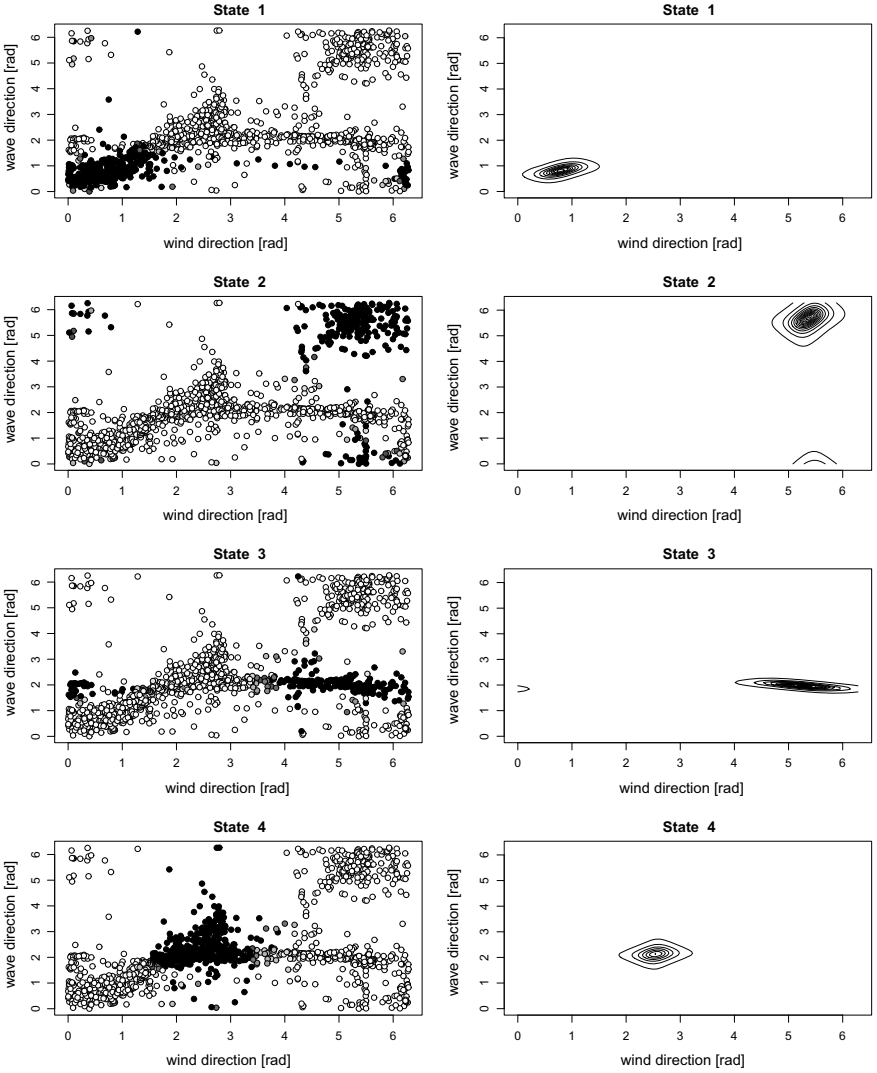


Fig. 2 Segmentation of a time series of wind and wave directions. Left: observations colored with grey levels according to the estimated membership probabilities of each class (black indicates a probability equal to 1). Right: contour plot of state-specific copula-based densities

to the estimated posterior probabilities (5) of class membership (black indicates 1). Remarkably, most points are black, indicating that the model segments the data according to well separated latent classes.

By computing the proportion p_k of the data that have been MAP-allocated to each class k, we obtain the estimated marginal distributions $\sum_k p_k f(y, \hat{\beta}_k)$ and $\sum_k p_k f(y, \hat{\beta}_k)$ of the data, overlapped on the observed histograms in Fig. 4. These

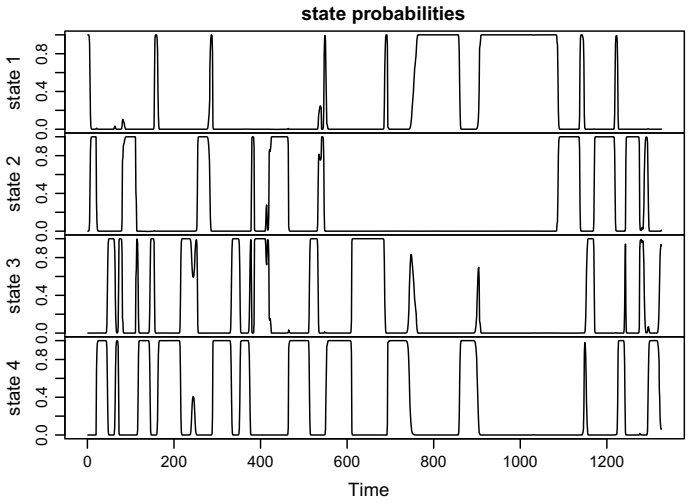


Fig. 3 The estimated posterior probabilities of the four latent states for each half hour in the study period

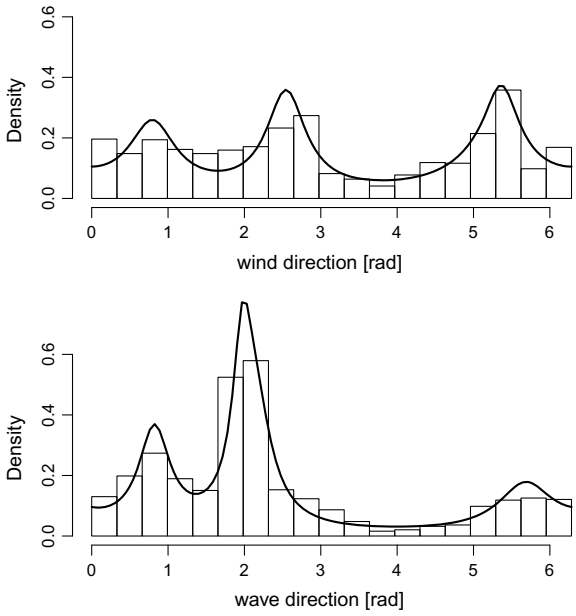


Fig. 4 Marginal distributions of the data: observed (histogram) and estimated by the model (continuous bold line) for wind (top) and wave (bottom) directions

pictures indicate a reasonable goodness of fit of the estimated marginal distributions, which can be improved by choosing a larger number K of components, if desired. Figure 5 indicates the estimated marginal densities under each state.

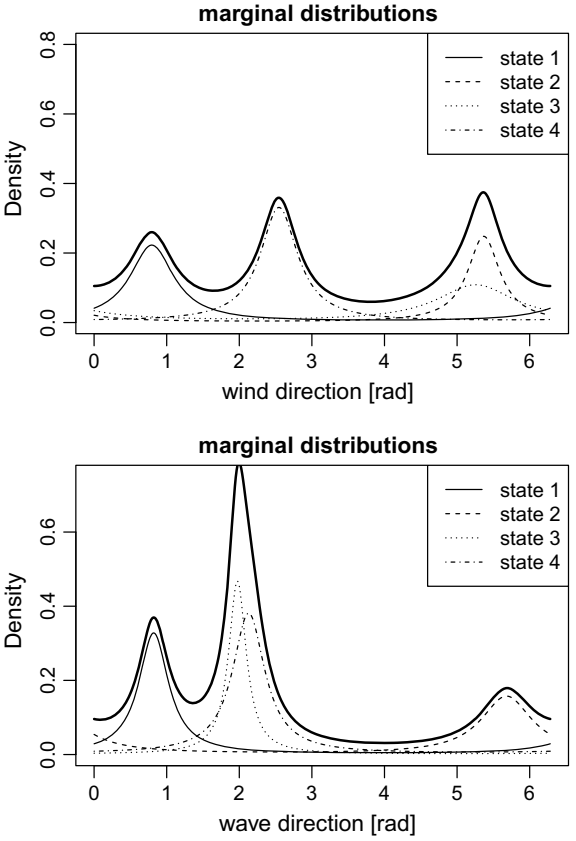


Fig. 5 The estimated marginal density (continuous bold line) and the estimated marginal densities under each state for wind (top) and wave (bottom) directions

5 Discussion

A novel HMM is introduced for segmenting toroidal times series according to a finite number of latent classes, associated with toroidal densities that describe the distribution of the data under each class. The model parsimoniously accommodates temporal auto-correlation, multimodality and circular correlation. It flexibly allows for any marginal distributions that is required by a specific case study. Parametric inference is relatively inexpensive from a computational viewpoint. In a case study of wind-wave data, the model segmented a time series of wave and wind directions according to intuitively appealing latent classes, providing a parsimonious description of wave dynamics in terms of interpretable environmental regimes.

Acknowledgements Francesco Lagona was supported by the 2015 PRIN supported project 'Environmental processes and human activities: capturing their interactions via statistical methods', funded by the Italian Ministry of Education, University and Scientific Research.

References

1. Bulla, J., Lagona, F., Maruotti, A., Picone, M.: A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. J. Agric., Biol. Environ. Stat. **17**, 544–567 (2012)
2. Coles, S.: Inference for circular distributions and processes. Stat. Comput. **8**, 105–113 (1998)
3. Johnson, R.A., Wehrly, T.E.: Some angular-linear distributions and related regression models. J. Am. Stat. Assoc. **73**, 602–606 (1978)
4. Jones, M.C., Pewsey, A., Kato, S.: On a class of circulas: copulas for circular distributions. Ann. Inst. Stat. Math. **67**, 843–862 (2015)
5. Kato, S., Pewsey, A.: A Möbius transformation-induced distribution on the torus. Biometrika **102**, 359–370 (2015)
6. Kim, G., Silvapulle, M., Silvapulle, P.: Comparison of semiparametric and parametric methods for estimating copulas. Comput. Stat. Data Anal. **51**, 2836–2850 (2007)
7. Lagona, F.: Copula-based segmentation of cylindrical time series. Stat. Probab. Lett. **144**, 16–22 (2019)
8. Lagona, F.: Correlated cylindrical data. In: Ley, C., Verdebout, T. (eds.) Applied Directional Statistics: Modern Methods and Case Studies, Chapman and Hall/CRC, New York, pp. 45–59 (2018)
9. Lagona, F., Picone, M., Maruotti, A., Cosoli, S.: A hidden Markov approach to the analysis of space-time environmental data with linear and circular components. Stoch. Environ. Res. Risk Assess. **29**, 397–409 (2014)
10. Lagona, F., Picone, M.: Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data. J. Stat. Comput. Simul. **83**, 1223–1237 (2013)
11. Lagona, F., Picone, M.: A gaussian-von mises hidden markov model for clustering multivariate linear-circular data. In: Giudici, P., Ingrassia, S., Vichi, M. (eds.) Statistical Models for Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Heidelberg, pp. 171–179 (2013a)
12. Mastrantonio, G.: The joint projected normal and skew-normal: a distribution for poly-cylindrical data. J. Multivar. Anal. **165**, 14–26 (2018)

A Biased Kaczmarz Algorithm for Clustered Equations

Alessandro Lanteri, Mauro Maggioni and Stefano Vigogna

Abstract The Kaczmarz method is an iterative algorithm for solving overdetermined linear systems by consecutive projections onto the hyperplanes defined by the system equations. The method has a wide range of applications in signal processing, notably for biomedical imaging in X-ray tomography. It has been shown that selecting the hyperplane randomly at each iteration guarantees exponential convergence to the solution. We propose here a new implementation of the Kaczmarz method for clustered equations. When the hyperplanes are grouped into directional clusters, we draw the projection promoting sparse high-variance clusters. This leads to an improvement in performance, as we show in several numerical experiments. Some applications to image reconstruction are presented.

Keywords Image reconstruction · Kaczmarz method · Randomized algorithm · Overdetermined linear systems

1 Introduction

In many applications, one aims to recover a signal $x \in \mathbb{C}^n$ from m linear measurements

$$b_r = a_r^* x \qquad a_r \in \mathbb{C}^n \qquad r = 1, \dots, m. \tag{1}$$

A. Lanteri (✉)
University of Torino, Turin, Italy
e-mail: alessandro.lanteri@unito.it

M. Maggioni
Johns Hopkins University, Baltimore, USA
e-mail: mauro.maggioni@jhu.edu

S. Vigogna
University of Genova, Genoa, Italy
e-mail: vigogna@dibris.unige.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_33

For example, b_r may be cross-sectional scans of some object x. When the number of observations is large comparing to the dimension, i.e. $m \gg n$, the implementation of standard linear solvers such as Gaussian elimination and singular-value decomposition may be prohibitive, requiring $O(mn^2)$ operations.

The Kaczmarz method, introduced in [3] and rediscovered later in [2], is an iterative method which approximates the solution of a linear system without loading the whole matrix and within a number of iterations possibly independent of the number of rows. This feature makes it very appealing when dealing with large-scale overdetermined systems

$$Ax = b, \tag{2}$$

where $A = [a_1 \cdots a_m]^*$ is a full rank $m \times n$ matrix and $b = [b_1, \ldots, b_m]^T$. The basic idea is beautifully simple: since x is the intersection of the hyperplanes defined by the Eqs. (1), one will get close to x starting from an initial guess and projecting successively onto such hyperplanes in iterative fashion. In formulas, the $(i+1)$th iteration of the Kaczmarz method is

$$x^{i+1} := x^i + (b_{r_i} - a_{r_i}^* x^i) \frac{a_{r_i}}{\|a_{r_i}\|_2^2}. \tag{3}$$

Note that the solver takes one projection per iteration, picking one row a_{r_i} at a time.

In its classical version, the algorithm runs cyclically through the equations in the given order:

$$r_i := i \mod m + 1. \tag{4}$$

It turns out, however, that shuffling the sequence can have a dramatic impact on how fast (3) will approximate the solution (Fig. 1). For example, drawing the hyperplane randomly at each iteration (with or without replacement) may speed up the conver-

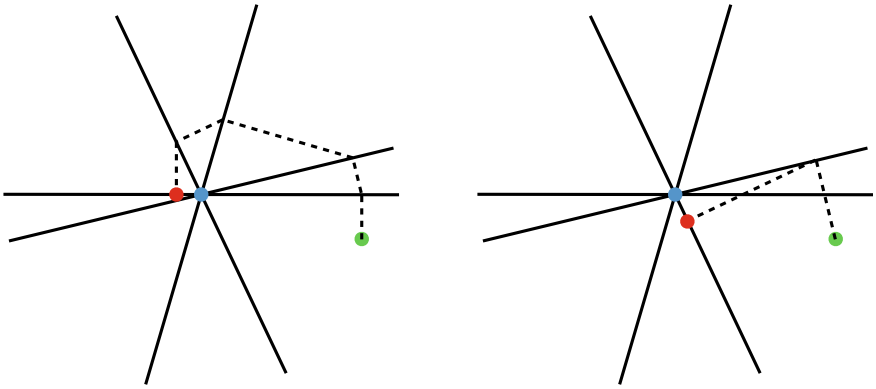


Fig. 1 Intuitive representation of the Kaczmarz method. The number of iterations needed to achieve a desired accuracy strongly depends on the order in which the hyperplanes are picked

gence. In this direction, Strohmer and Vershynin proposed a randomized Kaczmarz algorithm (RKA) [4], where at each iteration the hyperplane is selected with probability proportional to its directional energy, namely

$$\mathbb{P}\{r_i = r\} := \|a_r\|_2^2 / \|A\|_F^2. \tag{5}$$

By means of such randomization, they prove that the mean squared error (MSE) of the Nth iterand obeys

$$\mathbb{E}\|x^N - x\|_2^2 \leq \|x^0 - x\|_2^2 \, (1 - \kappa(A)^{-2})^N, \tag{6}$$

where $\kappa(A) := \|A\|_F \|A^{-1}\|_2 \geq \sqrt{n}$ is the scaled condition number of A. This estimate shows that the Kaczmarz approximation converges exponentially to the true solution. Viewed differently, this estimates says that the expected number of iterations N_ε needed to achieve a desired accuracy ε is

$$N_\varepsilon \leq O\left(\kappa(A)^2 \log \frac{1}{\varepsilon}\right).$$

Thus, if the matrix A is well conditioned, say $\kappa(A)^2 = O(n)$, RKA will converge in $O(n)$ iterations, independently of the number of equations m. Since each iteration (3) takes $O(n)$ time and the computation of the density (5) costs $O(mn)$, RKA will compute a solution in $O(mn + n^2)$ operations, as opposed to the usual $O(mn^2)$ of Gaussian elimination. Furthermore, RKA works on one n-dimensional array at a time, drawn from an m-dimensional array of probabilities, reducing the auxiliary space complexity from $O(mn)$ to $O(m + n)$.

While RKA as proposed by [4] picks the rows according to (5), other densities are of course possible, and may be better depending on specific properties of the matrix A. As it has been observed in [1], the solution to (2) is invariant to independent scalings of the rows of A, hence (6) can be generalized for arbitrary row-selection laws. Indeed, for a generic distribution

$$p_r := \mathbb{P}\{r_i = r\} \qquad r = 1, \ldots, m\,, \tag{7}$$

one has

$$\mathbb{E}\|x^N - x\|_2^2 \leq \|x^0 - x\|_2^2 \, (1 - \kappa(A, p)^{-2})^N \tag{8}$$

with $\kappa(A, p) := \| (\mathrm{diag}(\|a_1\|_2, \ldots, \|a_m\|_2)A)^{-1} \, \mathrm{diag}(p_1, \ldots, p_m)^{-1/2} \|_2$. This begs the question of finding a density (7) maximizing the convergence of (8) for a given matrix A. In fact, the MSE can be computed exactly, and the research of the best probabilities can be formulated as a convex optimization problem (see [1]). However, solving such problem might be impractical, and a principled, more direct strategy to choose a good performing distribution may be preferable in dealing with a class of

coefficient matrices, rather than computing the optimal solution for each matrix in the class.

We propose a choice of row-selection law under the assumption that the matrix A features a particular clustered structure. We think of A as a measurement setting. Whether such setting has been given or designed, some of its measurements might be more alike, and some groups of similar measurements might have larger or smaller cardinality than others. For instance, the rows of A may represent measurements taken from several groups of sensors, each of which spanning more or less directions. Or again, different subsets of equations may correspond to different frequency bands sampled at various rates. In such configurations, the solver could favor blocks of equations with higher linear independence, and counterbalance the effect of oversampled redundant information.

To this end, we need to choose a notion of distance defining the clusters, and work out quantifiable properties of sparsity and linear independence. Consisting of consecutive projections, the Kaczmarz procedure is directional by nature, hence it is reasonable to group together hyperplanes having similar orientation. The rows of A define such orientations up to an arbitrary scaling factor, therefore we seek clusters $C_1, \dots, C_K$ of normalized rows as clusters of points on the $(n-1)$-dimensional unit sphere. The statistical variance σ_k^2 of a cluster is a measure of the linear independence of the corresponding block of rows, while the cardinality m_k is indicative of its redundancy. In light of these considerations, we will define the extraction probability of a row to be directly proportional to the standard deviation and inversely proportional to the cardinality of the cluster containing that row:

$$\mathbb{P}\{r_i = r\} := \frac{\sigma_{k(r)}/m_{k(r)}}{\sum_{r=1}^{m} \sigma_{k(r)}/m_{k(r)}}, \tag{9}$$

where $k(r)$ is the index of the cluster of the r-th row, i.e. $a_r \in C_{k(r)}$, and $\sigma^2_{k(r)}$, $m_{k(r)}$ denote the variance and the cardinality, respectively, of $C_{k(r)}$. The distribution (9) will encourage directional variability and compensate for unbalanced sampling. We call our method a Biased Kaczmarz Algorithm (BKA). Note that, unlike RKA, the implementation of BKA is completely invariant under arbitrary scalings.

The paper is organized as follows. In Sect. 2 we illustrate the algorithm and discuss its time and space complexity. In Sect. 3 we show the empirical convergence of our algorithm and compare its performance with the Random Kaczmarz Algorithm. We use synthetic data in Sect. 3.1, and apply the method to Lenna's picture and a snapshot of the moon in Sect. 3.2. We finally collect some considerations and draw our conclusions in Sect. 4.

2 Biased Kaczmarz Algorithm

We describe here our Biased Kaczmarz Algorithm (BKA, Algorithm 1). BKA requires as input the linear system to be solved, that is, a coefficient matrix $A \in \mathbb{C}^{m\times n}$ and constant terms $b \in \mathbb{C}^m$, and the number of iterations N to be performed, the default value being $10n$. If available, it can be provided with the clustering structure of A, encoded in the clusters labels $k \in \{1, \ldots, K\}^m$. Given these inputs, the first step of the algorithm consists in normalizing each equation $a_r^* x = b_r$ by $\|a_r\|_2$. If the cluster labels k are not assigned, BKA will perform a K-means step (cross-validating if the number of clusters K is also unknown). Once k is determined, the algorithm will compute the total variance $\widehat{\sigma}_k^2$ and the cardinality $\widehat{m}_k$ of each cluster. The extraction probabilities $\widehat{p}_r$ will then be computed according to Eq. (9). After a random initialization, BKA will iteratively perform Eq. (3), drawing the rows according to the density computed in the previous step. The output of the algorithm is the approximate solution $\widehat{x} \in \mathbb{C}^n$ of the system $Ax = b$ obtained at the Nth iteration.

Algorithm 1: Biased Kaczmarz Algorithm

Input : $A \in \mathbb{C}^{m\times n}$: coefficient matrix
$b \in \mathbb{C}^m$: constant terms
$N \geq 1$: number of iterations (default value is $10n$)
$K \geq 1$: number of clusters (optional)
$k \in \{1, \ldots, K\}^m$: cluster labels (optional)
Output: $\widehat{x} \in \mathbb{C}^n$: approximate solution of $Ax = b$

normalize each row: $[a_r^*, b_r] \leftarrow [a_r^*, b_r]/\|a_r\|_2, r = 1, \ldots, m$;
if *K is provided and k is not provided* **then**
  compute k using the K-means algorithm;
end
if *K and k are not provided* **then**
  compute K and k using a cross-validated K-means algorithm;
end
compute each cluster total variance $\widehat{\sigma}_k^2$ and cardinality $\widehat{m}_k, k = 1, \ldots, K$;
compute density weights: $\widehat{p}_r \leftarrow \widehat{\sigma}_{k(r)}/\widehat{m}_{k(r)}, r = 1, \ldots, m$;
initialize x^0 arbitrarily;
for $i = 1$ *to* N **do**
  $r_0 \leftarrow r$ with probability $\widehat{p}_r / \sum_{r=1}^m \widehat{p}_r$;
  $x^{i+1} \leftarrow x^i + (b_{r_0} - a_{r_0}^* x^i) a_{r_0}$;
end
return $\widehat{x} \leftarrow x^N$.

Computational Complexity

The time complexity of BKA for a well conditioned matrix is $O(mn + n^2)$, of which $O(mn)$ to compute row norms, clusters and variances, and $O(n^2)$ to iterate (3). When the clustering structure is known a priori, the auxiliary space complexity is $O(m + n)$ as for RKA. If necessary, the K-means step increases the space complexity

from $O(m+n)$ to $O(mn+m+n)$. Nevertheless, the clusters need to be learned only once for a fixed setting A, and can then be used to resolve several signals x. Therefore, after a one-off step demanding $O(mn)$ memory, each resolution will only need to store $O(m+n)$ numbers.

3 Numerical Results

In this section we present some experiments to show the performance of our algorithm in comparison with the Randomized Kaczmarz Algorithm [4], both on synthetically generated data (Sect. 3.1) and on real images (Sect. 3.2).

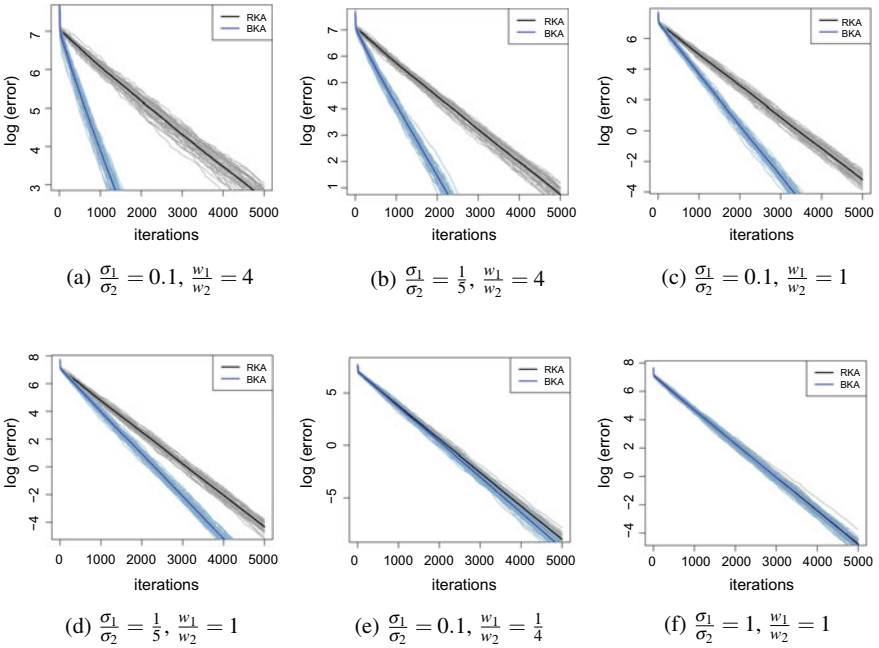
3.1 Empirical Rates

In our numerical simulations we generate A with $n = 100$, $m = 5{,}000$ and $K = 2$. Each row a_r^* is generated from a distribution

$$\frac{w_1\mathcal{N}(\mu_1, I_n\sigma_1^2) + w_2\mathcal{N}(\mu_2, I_n\sigma_2^2)}{w_1 + w_2},$$

and then normalized by its ℓ^2-norm. We draw x from the standard n-dimensional normal distribution and take b as $b = Ax$. We apply K-means to the rows of A in order to obtain two clusters, and compute their respective empirical standard errors $\widehat{\sigma}_1, \widehat{\sigma}_2$ and cardinalities $\widehat{m}_1, \widehat{m}_2$. We apply both the Randomized and the Biased Kaczmarz Algorithm. In RKA, the rows are selected at random with uniform probability, while BKA picks a row with probability proportional to $\widehat{\sigma}_1/\widehat{m}_1$ if it belongs to the first cluster, or $\widehat{\sigma}_2/\widehat{m}_2$ if it belongs to the second.

In Fig. 2 we display the results of our numerical simulations for different choices of ratios σ_1/σ_2 and w_1/w_2. We fixed μ_1 and μ_2 as the versors of the first two coordinate axes. For each setting, the experiment is repeated 50 times. Each plot in Fig. 2 shows, for both RKA and BKA, all 50 "error trajectories", the "average error trajectory" and the standard deviation bands. One "error trajectory" is the natural logarithm of the error $\|x - x^i\|_2$ against the iteration i. As expected, when one cluster has higher variability and a lower number of points than the other cluster (Fig. 2a, b), BKA significantly outperforms RKA. When $w_1 = w_2$, but there is a consistent difference between the variability of the two clusters (Fig. 2c, d), BKA still significantly outperforms RKA. In the case where the cluster with more variability also have more points (Fig. 2e) BKA performs slightly better than RKA. As expected, when $\sigma_1 = \sigma_2$ and $w_1 = w_2$ (Fig. 2f) the two algorithms perform in the same way, being the extraction density of BKA nearly uniform. Overall, the simulations shown in Fig. 2 confirm



(a) $\frac{\sigma_1}{\sigma_2} = 0.1$, $\frac{w_1}{w_2} = 4$ (b) $\frac{\sigma_1}{\sigma_2} = \frac{1}{5}$, $\frac{w_1}{w_2} = 4$ (c) $\frac{\sigma_1}{\sigma_2} = 0.1$, $\frac{w_1}{w_2} = 1$

(d) $\frac{\sigma_1}{\sigma_2} = \frac{1}{5}$, $\frac{w_1}{w_2} = 1$ (e) $\frac{\sigma_1}{\sigma_2} = 0.1$, $\frac{w_1}{w_2} = \frac{1}{4}$ (f) $\frac{\sigma_1}{\sigma_2} = 1$, $\frac{w_1}{w_2} = 1$

Fig. 2 Comparison of the rates of convergence of BKA and RKA applied to a matrix with $m = 5{,}000$ rows and $n = 100$ columns. These figures show, in different experiment settings, the "error trajectories" of 50 randomly-selected trials of BKA and RKA (light blue and light gray, respectively), the average trajectories (solid blue for BKA and solid black for RKA) and the standard deviation bands (dotted blue for BKA and dotted black for RKA)

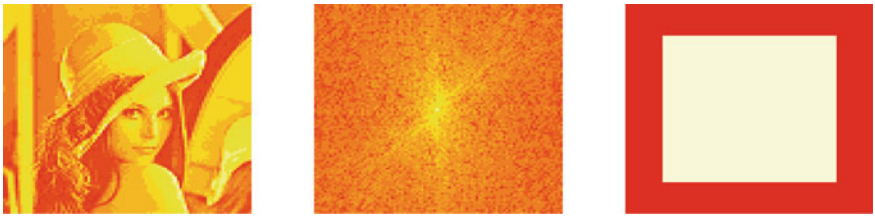
that a biased approach promoting the extraction of rows from clusters with higher variability can improve considerably the performance of the Kaczmarz method.

3.2 Applications to Image Reconstruction

We now apply our method for the reconstruction of images from irregular redundant equations clustering in separate regions of the frequency (Sect. 3.2.1) or space domain (Sect. 3.2.2).

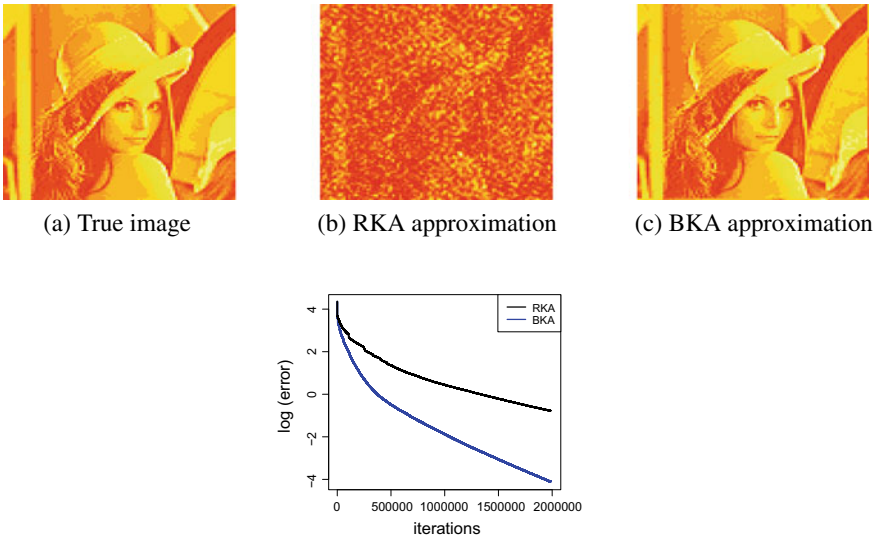
3.2.1 Clusters in Frequency

Our goal is to reconstruct a picture from a sequence of integrals of its Fourier transform. If such integrals are supported on different frequency bands, the reconstruction procedure should first seek to recover the lowest frequencies, where most of the infor-



(a) Image (b) Fourier transform (c) Box and annulus

Fig. 3 Lenna's picture (**a**), the logarithmic magnitude of its Fourier transform (**b**), and the decomposition of the frequency plane into a low frequency box and a high frequency annulus (**c**)



(a) True image (b) RKA approximation (c) BKA approximation

(d) Rates

Fig. 4 True image (**a**) at a resolution of 100 by 100 pixels, and approximations obtained after 2,000,000 iterations of RKA (**b**) and BKA (**c**). BKA reconstructs the image with very few imperfections, while the RKA approximation is still very poor. **d** shows the "error trajectories" of RKA and BKA (black and blue line, respectively)

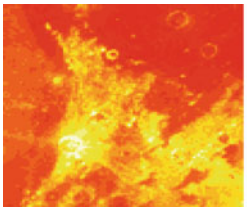
mation is stored. When using a randomized Kaczmarz method, an unbalanced set of measurements may adversely affect the selection of low frequency integrals, forcing the solver to persist in exploring uninformative directions. An excess of high frequency observations may in fact bias the extraction density, taking up the computing resources with the retrieval of small coefficients. In the following we show how our algorithm can prevent this issue, and compare the result with what is obtained by drawing uniformly.

For our experiment we picked x as the Fourier transform of a reduced grayscale version of Lenna's picture (Fig. 3a). The image is 100 by 100 pixels, hence $n =$

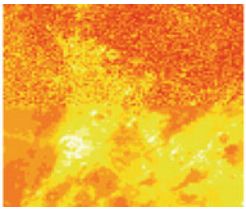
10,000. The magnitude of x, in logarithmic scale, is shown in Fig. 3b. We divided the frequency plane in two regions of approximately even area: a central low frequency box of side length 70, and the complimentary high frequency annulus (Fig. 3c). Each equation in the system is a weighted sum of pixels intensities supported in one of the two regions. We took $m = 5n$ equations, drawing the non-zero coefficients from a Gaussian distribution $\mathcal{N}(1, 0.5)$, and then normalizing. We generated 0.2 m equations on the low frequency box, and 0.8 m equations on the high frequency annulus. In Fig. 4 we compare the approximations of the image obtained using RKA (Fig. 4b) and BKA (Fig. 4c). After 2,000,000 iterations, BKA has recovered the image almost perfectly, while RKA is still far. Figure 4d displays the "error trajectories" of RKA and BKA.

3.2.2 Clusters in Space

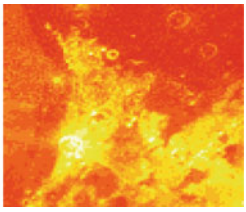
In this example we simulate a situation where a geospatial analysis has collected uneven information, capturing unequal amounts of data on equally important regions of space. Our method removes the statistical bias leading to a better uniform reconstruction.



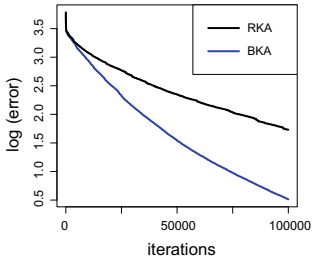
(a) True image



(b) RKA approximation



(c) BKA approximation



(d) Rates

Fig. 5 True image (**a**) at a resolution of 100 by 100 pixels, and approximations obtained after 100,000 iterations of RKA (**b**) and BKA (**c**). BKA reconstructs the image with very few imperfections, while the RKA approximation is poor in the upper region and less accurate in the lower one. **d** shows the "error trajectories" of RKA and BKA (black and blue line, respectively)

In this test x is as a grayscale 100 by 100 picture of lunar craters (Fig. 5a). Each of the $m = 5n$ ($n = 10{,}000$) equations is a weighted sum of pixels intensities supported either in the upper or lower half of the image. The non-zero coefficients are generated from a Gaussian distribution $\mathcal{N}(1, 5)$, and then normalized. The upper and lower regions have been sampled with 0.2m and 0.8 m equations, respectively. Figure 5 shows the results obtained from the application of RKA (Fig. 5b) and BKA (Fig. 5c). After 100,000 iterations, BKA recovers the picture pretty well. On the other hand, RKA does a poor job on the upper half and is still less accurate in the lower half. We plot the "error trajectories" in Fig. 5d.

4 Conclusions

We presented a new scale-invariant randomized implementation of the Kaczmarz method with the aim of improving its performance on sets of equations featuring a clustered directional structure. Our row-selection law is designed to favor blocks of equations with higher linear independence and level out the bias coming from redundant or poor local sampling. If the clustering structure of the system is known, our algorithm has the same computational complexity as the Randomized Kaczmarz Algorithm [4], otherwise it requires in addition a preliminary K-means step, to be performed once for a fixed set of coefficients and all desired unknowns. Our numerical experiments show that our algorithm achieves faster convergence rates than [4] in several configurations of the aforementioned setting.

Acknowledgements The authors are thankful for partial support under awards AFOSR FA9550-17-1-0280 and NSF ATD 1737984.

References

1. Agaskar, A., Wang, C., Lu, Y. M.: Randomized Kaczmarz algorithms: exact MSE analysis and optimal sampling probabilities. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP) Atlanta, GA (2014)
2. Gordon, R., Herman, G.T., Johnson, S.A.: Image reconstruction from projections. Sci. Am. **233**(4), 56–71 (1975)
3. Kaczmarz, S.: Angenäherte Auflösung von Systemen linearer Gleichungen. Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques **35**, 355–357 (1937)
4. Strohmer, T., Vershynin, R.: A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl. **15**, 262–278 (2009)

Nearly Unbiased Probability Plots for Extreme Value Distributions

Antonio Lepore

Abstract Probability plots allow for a straightforward analysis of the data and interpretation of results also by non-statisticians and still play a central role in today's software. In this chapter, probability plots for extreme value (EV) distributions are developed based on the generalized least-squares distribution fitting method and on convenient approximations of the first two moments of order statistics from the standard EV distributions. The proposed probability plots lead to graphical estimators of parameters that are shown to be nearly unbiased through the use of pivotal indices that avoid the massive numerical investigations usually presented for similar purposes in the recent literature. Although more efficient biased solutions can be theoretically found, the obtained parameter estimators achieve also adequate performances in terms of mean square deviation with respect to those derived through probability plots that have been presented separately in the literature as the most effective for EV distributions. Lastly, a real-case study is presented concerning wind speed data collected at a candidate wind farm site in Southern Italy. The results demonstrate how the proposed probability plot can effectively support EV analysis and assist practitioners in the selection of the turbine class to be installed.

Keywords Graphics and data visualization · Linear unbiased estimators · Location-scale distributions · Extreme value distribution · Gumbel distribution

1 Introduction

Practitioners still use software tools that adopt probability plots to check the fit provided by the selected model graphically and, in general, to have deeper insight and visual understanding of statistical information (e.g., outliers and leverage points). Classical probability plots are essentially obtained by reporting ordered observations

A. Lepore (✉)
Department of Industrial Engineering, University of Naples Federico II,
P.le V. Tecchio 80, 80125 Naples, Italy
e-mail: antonio.lepore@unina.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_34

of a random variable (i.e., empirical data) against the corresponding estimates $\hat{F}_i$ of the parent cumulative distribution function (cdf) (i.e., the plotting position) on axes that are properly scaled to achieve linearity. Then, as the parent distribution is required to belong to the location-scale family, the distribution parameters are estimated through the slope and the intercept of the line of best fit [12].

However, the choice of the distribution fitting method, of the response variable and the corresponding relative accuracy are not always clear [11]. This has given rise to recent controversial contributions on probability plot and plotting position definitions [2, 3, 5–8, 12, 17–21]. Relevant and comprehensive overview of probability plots and plotting positions can be found in [11, 15]. This work, instead, is focused on the most recent and relevant contributions in the special case of extreme value (EV) distributions.

Pirouzi Fard and Holmquist [21] define plotting positions based on simple approximations of variances and covariances for order statistics from the standard EV distribution for maxima. Pirouzi Fard [20] provides a comparison between the ordinary least-squares (OLS) and the generalized least-squares (GLS) distribution fitting methods for probability plots when the data set arises from the standard EV distribution for minima. Cook and Harris [3] find out in the case of the EV distribution for maxima that the classical Gringorten estimator [9] of the order statistic mean gives satisfactory results only asymptotically, even if it is commonly used for small sample sizes. Fuglem et al. [8] support previous work by Cunnane [4] and state that plotting position should be only defined according to the selected parent distribution. On the opposite side, Makkonen et al. [18, 19] support the classical distribution-free approach [10]. In this work, the rationale behind the graphical best linear unbiased estimators (BLUEs) [7], which have recently appeared in the literature, is exploited and elaborated for the particular case of EV distributions.

The remainder of the chapter is as follows. In Sect. 2 the problem is mathematically stated and the new probability plots are proposed based on convenient approximations of the first two moments of order statistics for the EV distributions. In Sect. 3 the estimators of the distribution parameters obtained by the proposed probability plot are shown to be nearly unbiased and are compared through proper pivotal (parameter-free) indices that avoid the massive numerical investigation usually presented for similar purposes even in the recent literature. In Sect. 4, a real-case study is presented concerning wind speed data collected at a candidate wind farm site in Southern Italy. The results show how the proposed probability plot can effectively support EV analysis and assist practitioners in the selection of the turbine class to be installed. Section 5 draws conclusions and practical directions.

2 Approximation of the BLUEs of Extreme Value Distribution Parameters via Probability Plots

The EV cdf for minima (referred to as extreme value distribution in [20, 21]) and for maxima (referred to Gumbel as in [3, 11, 12]) are, respectively, as follows

$$F_m(x; a, b) = 1 - e^{-e^{\frac{x-a}{b}}}, \qquad F_M(x; a, b) = e^{-e^{-\frac{x-a}{b}}}; \qquad b > 0. \tag{1}$$

As is known, EV standard cdf's can be obtained by setting $a = 0$ and $b = 1$ and have inverse functions that are infinitely differentiable. Performances of graphical approaches for EV distributions are influenced by the choice of the plotting position formula, the distribution fitting method, as well as the covariance matrix (or its approximation) especially if the sample size is small [11]. In general, the best results are achieved by using the ordered observations of a sample of size N, $x_{(1)}, \ldots, x_{(i)}, \ldots, x_{(N)}$, as response variable and the mean of the standard order statistics, $\mu_{(1)}, \ldots, \mu_{(i)}, \ldots, \mu_{(N)}$, as explanatory variables. This choice is mandatory for the GLS distribution fitting method, which explicitly requires the specification of the covariance $\sigma_{(i,j)}$ between the i-th and j-th standard order statistics ($1 \le i \le j \le N$) in order to obtain BLUEs of distribution parameters [7].

In this chapter the approximations suggested in [7] for $\mu_{(i)}$ and $\sigma_{(i,j)}$ truncated to the fourth order term and given by

$$\begin{aligned}\tilde{\mu}_{(i)} =& G^{-1}(p_i) + \frac{1}{2} G^{-1(2)}(p_i) \frac{p_i(1-p_i)}{(N+2)^2} + \frac{1}{3} G^{-1(3)}(p_i) \frac{p_i(1-p_i)(1-2p_i)}{(N+2)^2} \\ &+ \frac{1}{8} G^{-1(4)}(p_i) \frac{p_i^2(1-p_i)^2}{(N+2)^3}\end{aligned} \tag{2}$$

$$\begin{aligned}\tilde{\sigma}_{(i,j)} =& \frac{p_i(1-p_i)}{N+2} G^{-1}(p_i) + \frac{p_i(1-p_j)}{(N+2)^2} \Big[(1-2p_i) G^{-1(2)}(p_i) G^{-1}(p_j) \\ &+ (1-2p_j) G^{-1(2)}(p_j) G^{-1}(p_i) + \frac{1}{2} p_i (1-p_i) G^{-1(3)}(p_i) G^{-1}(p_j) \\ &+ \frac{1}{2} p_j (1-p_j) G^{-1(3)}(p_j) G^{-1}(p_i) + \frac{1}{2} p_i (1-p_j) G^{-1(2)}(p_i) G^{-1(2)}(p_j) \Big]\end{aligned} \tag{3}$$

are conveniently elaborated for the EV distribution for minima (resp. maxima), where $G(x) = F_m(x; 0, 1)$ (resp. $G(x) = F_M(x; 0, 1)$), $p_i = i/(N+1)$, and $G^{-1(k)}(x)$ is the k-th derivative of the inverse function $G^{-1}(x)$. In order to do that, it is worth noting that simple closed forms are available for $\mu_{(1)}$ and $\sigma_{(1,1)}$ in the case of the EV distribution for minima

$$\mu_{(1)} = -\gamma - \ln N, \qquad \sigma_{(1,1)} = \pi^2/6 \tag{4}$$

and for $\mu_{(N)}$ and $\sigma_{(N,N)}$, in the case of the EV distribution for maxima

$$\mu_{(N)} = \gamma + \ln N, \qquad \sigma_{(N,N)} = \pi^2/6 \tag{5}$$

where γ is the Euler's constant. Therefore, expressions (4) and (5) can be more opportunely used in place of (2) and (3). Note that the GLS regression of $\tilde{\mu}_{(i)}$ on the sample observations, through the covariance approximation $\tilde{\sigma}_{(i,j)}$, lead to graphical estimators for a and b that are not unbiased because of the approximations. Hence, based on the results drawn in [11], it can be of interest to compare the latter approach with the most effective ones among those mentioned in the introduction and summarized in the first two rows of Table 1, namely Pirouzi Fard (PF) [20] and Hong and Li (HL) [12]. In general, each approximation $\tilde{\mu}_{(i)}$ is associated with a plotting position $\hat{F}_i = G^{-1}\left(\tilde{\mu}_{(i)}\right)$ and vice versa.

The last two rows of Table 1 report the Cook and Harris (CH) [3] and the classical Gumbel (GU) [10] plotting position approach that rely instead on the use of the OLS method. Note that PF only applies to the EV distribution for minima, whereas HL and CH only apply to that for maxima.

3 Simulation Study and Results

A simulation study is carried out by drawing $M = 10^5$ pseudo-random samples from the EV distributions for minima and maxima at sample sizes $N = 5$ and $N = 30$ to compare

(i) the goodness of the approximations used for $\mu_{(i)}$
(ii) (when applicable) the goodness of the approximations used for $\sigma_{(i,j)}$
(iii) the bias and the efficiency of graphical estimators for a and b

of the proposed approach and its competitors listed in Table 1.

Table 1 Summary of the plotting positions $\hat{F}_i = G^{-1}\left(\tilde{\mu}_{(i)}\right)$ used by the competing probability plots ($1 \le i \le j \le N$). Covariance approximations $\hat{\sigma}_{(i,j)}$ indicate the use of GLS distribution fitting method instead of OLS. The correction factors γ_{Nk} ($k = 1, \ldots, 5$) are defined as in [12]

	$\hat{F}_i$	$\hat{\sigma}_{(i,j)}$
PF	$\begin{cases} 1-e^{\frac{-e^{-\gamma}}{N}} & i=1 \\ \frac{i-0.4866}{N+0.1840} & elsewhere \end{cases}$	$\begin{cases} \pi^2 6 & i=j=1 \\ \frac{(i-0.469)(N+0.831-i)^{-1}(N+0.073)^{-1}}{\ln\frac{N+0.779-j}{N+0.356}\ln\frac{N+0.8314-i}{N+0.356}} & elsewhere \end{cases}$
HL	$\begin{cases} e^{-\frac{e^{-\gamma}}{N}} & i=N \\ \frac{i-0.37+0.232/\sqrt{N}}{N+0.144+0.232/\sqrt{N}} & elsewhere \end{cases}$	$\begin{cases} \pi^2 6 & i=j=N \\ \frac{N+1-j-\gamma_{N1}}{(N+2-\gamma_{N2})(j-\gamma_{N3})\ln\frac{i-\gamma_{N5}}{N+1-\gamma_{N4}}\ln\frac{i-\gamma_{N3}}{N+1-\gamma_{N4}}} & elsewhere \end{cases}$
CH	$\frac{i-0.439+0.466/\ln(N)}{N+0.113+0.466/\ln(N)}$	–
GU	$\frac{i}{N+1}$	–

Table 2 $RMSE$ and MAD achieved at different sample sizes by the proposed approach and the four alternatives listed in Table 1 (bold text highlights the smallest value of each column)

	EV distribution for minima				EV distribution for maxima			
	RMSE		*MAD*		*RMSE*		*MAD*	
	$N = 5$	$N = 30$	$N = 5$	$N = 30$	$N = 5$	$N = 30$	$N = 5$	$N = 30$
Proposed	**0.00355**	**0.00057**	**0.01453**	**0.00193**	**0.00355**	**0.00057**	**0.01453**	**0.00193**
PF	–	–	–	–	0.01564	0.00288	0.02531	0.00324
HL	0.00756	0.00282	0.24872	0.32144	–	–	–	–
CH	0.04349	0.00743	–	–	–	–	–	–
GU	0.23592	0.12330	–	–	0.23592	0.12332	–	–

Slightly differently from [11, 12, 21], the following root mean square error ($RMSE$) index is used to compare (i)

$$RMSE = \sqrt{\sum_{i=1}^{N} (\mu_{(i)} - \tilde{\mu}_{(i)})^2 / N}, \tag{6}$$

whereas the maximum absolute deviation (MAD) is used to compare (ii)

$$MAD = \max_{1 \le i \le j \le N} |\sigma_{(i,j)} - \tilde{\sigma}_{(i,j)}|. \tag{7}$$

Note that RMSE defined by (6) can be determined for any $\mu_{(i)}$, differently to that used in [11]. Then, to perform comparison, the actual distribution parameters are chosen equal to standard values. This is necessary as the indices in (6) and (7) are still not pivotal (parameter-free), they may vary according to the actual distribution parameters. The exact evaluation of $\mu_{(i)}$ and $\sigma_{(i,j)}$ in (6) is obtained through numerical integration [16]. The lower the $RMSE$ and the MAD, the better the proposed approximation of $\mu_{(i)}$ and $\sigma_{(i)}$ are, respectively. The values of $RMSE$ and MAD achieved by $\tilde{\mu}_{(i)}$and $\tilde{\sigma}_{(i,j)}$ calculated through the proposed approach are compared in Table 2 with those calculated through the probability plots listed in Table 1.

The proposed approximations for the means and the covariances of the order statistics from the EV distributions achieve the best performances at each considered sample size ($N = 5, 30$) both in terms of $RMSE$ and MAD.

Finally, the deviations

$$e_{(i,N)} = \mu_{(i)} - \hat{\mu}_{(i)}, \qquad i = 1, \ldots, N, \tag{8}$$

which are used by Hong and Li [12] and contribute to Eq. (6), are further considered in order to compare (i) resulting from the proposed probability plots and its competitors reported in Table 1 for every rank i and at different sample sizes N. Usually (i) is referred to as the descriptive ability of the plotting position.

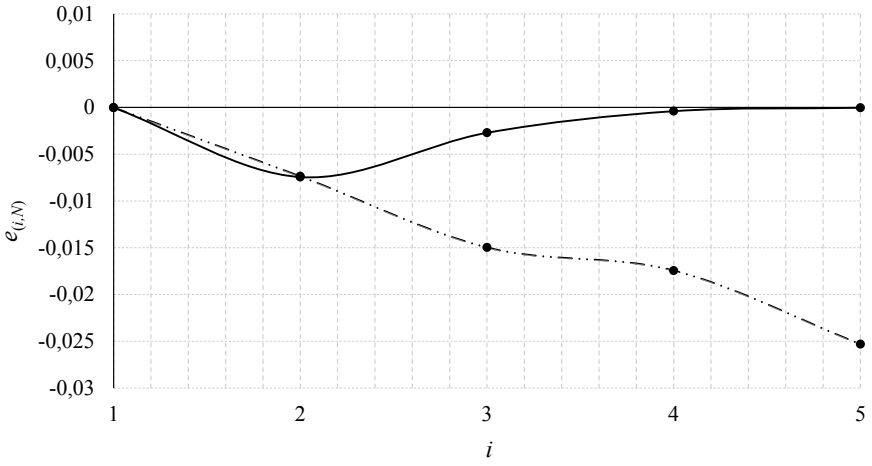


Fig. 1 Deviation $e_{(i,N)}$ achieved by PF (dot-dashed line) and the proposed (solid line) approximation of $\mu_{(i)}$ at sample size $N = 5$ for the EV distribution for maxima

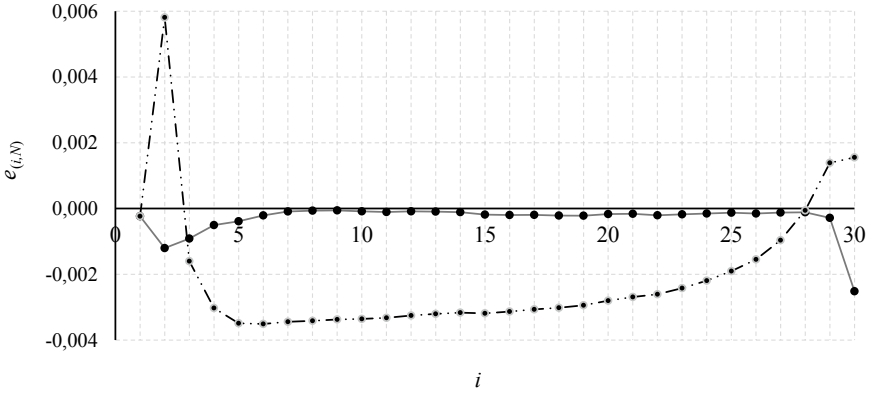


Fig. 2 Deviation $e_{(i,N)}$ achieved by PF (dot-dashed line) and the proposed (solid line) approximation of $\mu_{(i)}$ at sample size $N = 30$ for the EV distribution for maxima

In Figs. 1 and 2, the deviations $e_{(i,N)}$ defined in Eq. (8) achieved by the proposed approximation and PF are plotted versus i at sample sizes $N = 5$ and $N = 30$, respectively, in the case of EV distribution for maxima. From these figures, it is clear that the proposed approximations outperform the PF ones, which drastically overestimate $\mu_{(i)}$ as i increases, whereas the proposed one tends to zero. In particular, Fig. 2 shows that, when $N = 30$, the proposed approximation is very close to the exact value at each i (unless $i = 30$); whereas the one corresponding to the PF probability plot underestimates $\mu_{(i)}$ at $i = 2$ and $i = 30$ and considerably overestimates $\mu_{(i)}$ elsewhere. Trivially note that when $i = 1$ both the PF approximations and the proposed ones achieve the same and exact value (Eq. (5)).

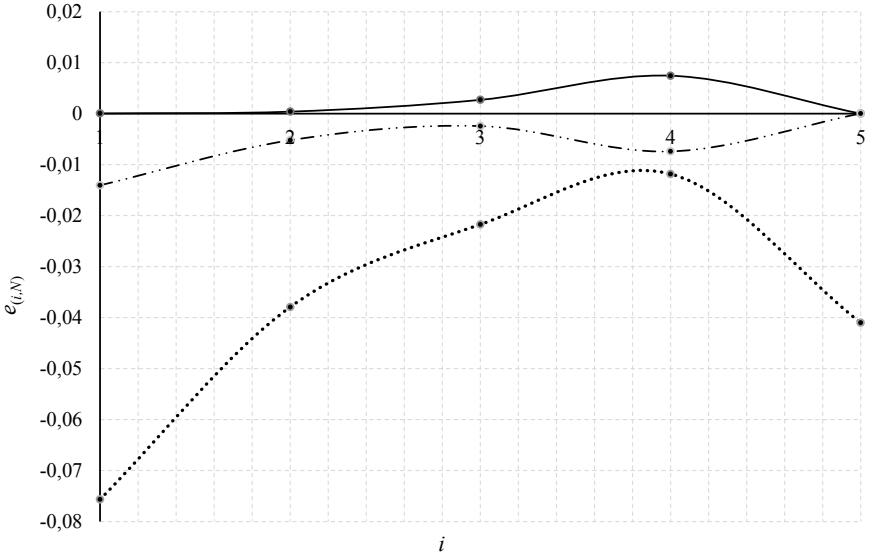


Fig. 3 Deviation $e_{(i,N)}$ achieved by HL (dashed line), CH (dotted line), and the proposed (solid line) approximation of $\mu_{(i)}$ at sample size $N = 5$ for the EV distribution for minima

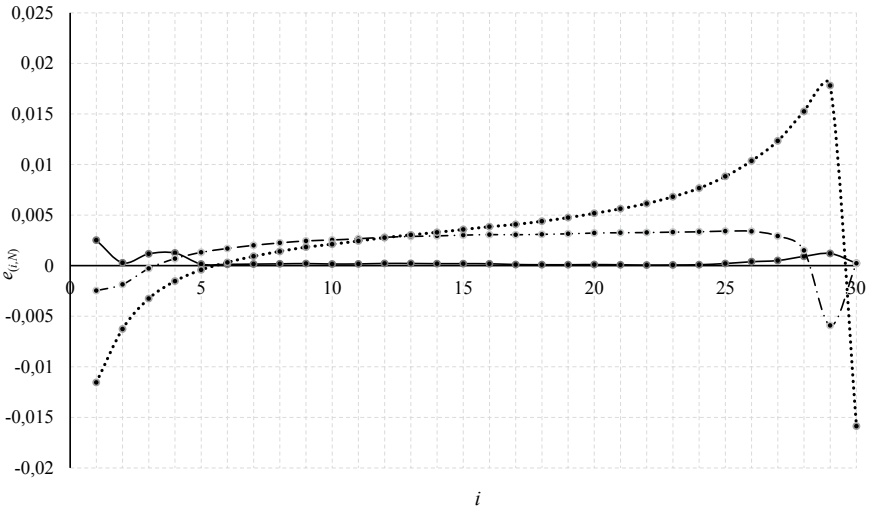


Fig. 4 Deviation $e_{(i,N)}$ achieved by HL (dashed line), CH (dotted line), and the proposed (solid line) approximation of $\mu_{(i)}$ at sample size $N = 30$ for the EV distribution for minima

Similarly, in the case of EV distribution for minima, Figs. 3 and 4 display the deviations $e_{(i,N)}$ achieved by the proposed approximation and the competitor HL and CH versus i at sample sizes $N = 5$ and $N = 30$, respectively. In particular, at sample size $N = 5$ (Fig. 3) the proposed plotting positions are clearly shown to provide the more accurate approximation of the first moment of the EV distribution for minima with respect to the competing ones. Whereas, the deviations achieved by the CH approximation are always the larger. At sample size $N = 30$ (Fig. 4), the better performance of the proposed approximation is generally confirmed unless $i = 1, 3, 4$ where the HL approximation performs slightly better.

Moreover, the following indices, namely the pivotal root deviation (PRD) and the pivotal absolute bias (PAB) of estimators $\hat{a}$ and $\hat{b}$, are introduced in order to compare (iii)

$$PRD\left(\hat{a}\right) = \sqrt{E\{\left(\hat{a} - a\right)^2\}/b^2}, \qquad PRD\left(\hat{b}\right) = \sqrt{E\{\left(\hat{b} - b\right)^2\}/b^2} \tag{9}$$

$$PAB\left(\hat{a}\right) = |E\{\hat{a}\} - a|\big/ b, \qquad PAB\left(\hat{b}\right) = |E\{\hat{b}\} - b|/b. \tag{10}$$

It is trivial to show that (9) and (10) are pivotals (see, e.g., [7, 14]) and therefore, the obtained results hold for any parameter. The indices PRD and PAB of the estimators $\hat{a}$ and $\hat{b}$ resulting from the proposed approach are reported in Table 3 with those resulting from the competitors listed in Table 1, as well as those of the classical maximum likelihood estimators (MLEs). As anticipated, note that CH and GU approaches do not involve the approximation of $\sigma_{(i,j)}$, thus do not apply for MAD (see Eq. (7)).

Table 3 PRD and PAB of $\hat{a}$ and $\hat{b}$ at different sample sizes by the proposed approach and the four alternatives listed in Table 1 (bold text highlights the smallest value of each column; MLEs are highlighted in italic text and are excluded from the comparison)

		EV distribution for minima				EV distribution for maxima			
		$PRD\left(\hat{a}\right)$	$PAB\left(\hat{a}\right)$	$PRD\left(\hat{b}\right)$	$PAB\left(\hat{b}\right)$	$PRD\left(\hat{a}\right)$	$PAB\left(\hat{a}\right)$	$PRD\left(\hat{b}\right)$	$PAB\left(\hat{b}\right)$
$N = 5$	Proposed	**0.480**	**0.002**	**0.408**	**0.001**	**0.480**	**0.002**	0.408	**0.001**
	PF	–	–	–	–	0.482	0.019	**0.404**	0.010
	HL	0.480	0.008	0.415	0.004	–	–	–	–
	CH	0.481	0.046	0.462	0.006	–	–	–	–
	GU	0.484	0.004	0.614	0.249	0.484	0.004	0.614	0.249
	MLE	*0.493*	*0.008*	*0.377*	*0.158*	*0.493*	*0.008*	*0.377*	*0.158*
$N = 30$	Proposed	**0.193**	**0.001**	**0.147**	**0.000**	**0.193**	**0.001**	**0.147**	**0.000**
	PF	–	–	–	–	0.193	0.002	0.147	0.002
	HL	0.194	0.002	0.149	0.002	–	–	–	–
	CH	0.197	0.005	0.185	0.001	–	–	–	–
	GU	0.195	0.006	0.216	0.090	0.194	0.006	0.216	0.090
	MLE	*0.194*	*0.013*	*0.145*	*0.025*	*0.194*	*0.013*	*0.145*	*0.025*

Table 3 confirms that the graphical estimators of distribution parameters resulting from the proposed approach achieve the smallest bias (PAB) and the highest efficiency (i.e., the smallest PRD) even when plugging in the proposed approximations for large sample sizes ($N = 30$). However, as expected, some rather biased estimators can be slightly more efficient at small sample sizes ($N = 5$), namely PF and HL. According to [11], note that graphical estimators of distribution parameters that rely on the OLS instead of the GLS distribution fitting method, namely CH and GU, are always the least efficient.

4 Real-Case Study: Wind Speed Data and Wind Turbine Classification

Many structural design criteria and engineering applications are based on the statistical analysis of EVs. In particular, the selection of the optimal class of turbines to be installed in a wind farm is based on the analysis of the wind speed maxima, and thus can be supported by the proposed probability plot. In particular, in this case study, wind speed data are collected from March 2013 to April 2017 at a Southern Italian site that is a candidate for the construction of a wind farm. The 4 years' worth of data consist of monthly maxima of 10 min average wind speeds.

The regulation IEC 61400-1 [13] of the International Electrotechnical Commission (IEC) specifies the design classes with respect to wind speed site-specific conditions. The practical goal is to identify the optimal wind turbine class that has both adequate robustness with respect to the site-specific wind loads and the higher energy production. In fact, the higher the class, the higher the energy production, but the lower the robustness of the wind turbine. Table 4 reports the wind turbine classification that appears in [13] based on the maximum acceptable *reference wind speed*, V_{ref}, which is defined as the wind speed percentile with a return period $RP = 50$ $years$. To carry out the analysis, the R software environment [22] is employed to extract the monthly maxima and obtain the probability plot in Fig. 5, as well as the 95% generalized prediction intervals constructed as in [7]. This figure shows the data are satisfactorily explained by the EV distribution for maxima. Accordingly, location and scale parameters $\hat{a}$ and $\hat{b}$ are calculated as well as the coefficient of determination R^2, in the special case of tied values and the GLS distribution fitting method [1]. The attained population line then allows practitioners to calculate, for the candidate site, the aforementioned reference wind speed $V_{ref} = 37.2$ m/s, which has return period $RP = 50$ $years$ (600 $months$) (i.e., the wind speed with percentile rank at $1 - 1/600 = 0.9983$). From Table 4, we see that the obtained reference wind

Table 4 V_{ref} parameter for wind turbine class determination [13]. The higher the class, the higher the energy production, the lower the robustness of the wind turbine

Wind turbine class	I	II	III	IV
V_{ref} (m/s)	50	42.5	37.5	30

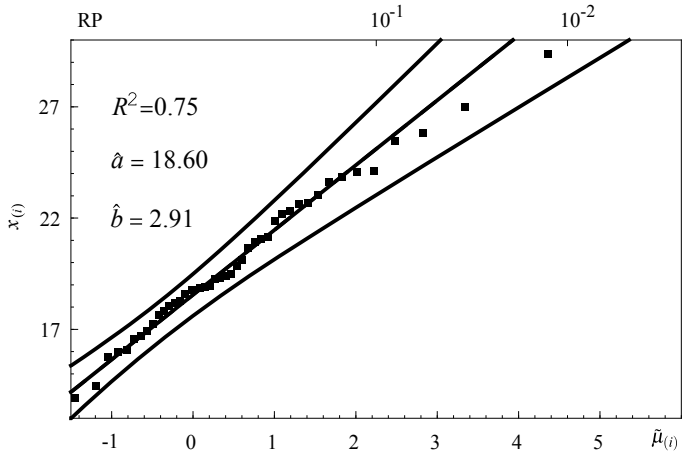


Fig. 5 Monthly maximum wind speeds from March 2013 to April 2017 reported on by Nearly unbiased EV probability plot with corresponding 95% confidence bands and generalized coefficient of determination R^2

speed value is located under the threshold 37.5 m/s and thus turbines of class III must be selected for the candidate site. This is indeed perfectly consistent with the experts' usual choice of turbines adopted in similar neighboring wind areas.

5 Conclusions

By conveniently approximating the first two moments of the standard extreme value distributions for minima and maxima, a new probability plot has been proposed. A simulation study has shown that the location and scale parameter estimators derived from this probability plot (i.e., graphical estimators) outperform the usual estimators obtained through the most popular competing probability plots appearing in the literature at all the considered sample sizes ($N = 5, 30$) and their efficiency is (comparatively) satisfactory. Moreover, the proposed probability plots are shown to have higher descriptive ability than competitors that have been presented separately in the literature as the most effective for extreme value distributions for minima and maxima. In other words, the resulting population line drawn by the proposed probability plot does not suffer from the typical bias related to classical probability plots, which is relevant especially for small sample sizes. In view of these results, the proposed probability plots can be regarded as straightforward tools for the analysis of the data and transfer of the results also to non-statisticians. In this way, practitioners do not necessarily need to abandon graphical methods, which are easy to compute and interpret, and opt for analytical solutions, irrespective of the purpose of the analysis. Nevertheless, in all the simulation scenarios, the worst results have been mostly

achieved by the classical Gumbel plotting positions. Therefore, this conclusion incidentally disproves the claim for the exclusive use of distribution-free approaches in the plotting position controversy raised in the last decade.

Acknowledgements The author is extremely grateful to the Editor and the anonymous reviewers for their valuable suggestions as well as to Professor Pasquale Erto for his continuous criticism and insight that significantly contributed to improve the chapter. The author also deeply thanks Ten Project S.r.l. (www.tenproject.it) for providing with the anemometric data and the engineer Massimo Lepore (Renewable Energy Source Systems) for his experienced discussion useful in defining the case study.

References

1. Buse, A.: Goodness-of-fit in the seemingly unrelated regressions model: a generalization. J. Econ. **10**(1), 109–113 (1979)
2. Cook, N.: Rebuttal of Problems in the extreme value analysis. Struct. Saf. (2012)
3. Cook, N.J., Harris, R.I.: The Gringorten estimator revisited. Wind Struct. An Int. J. **16**(4), 355–372 (2013). https://doi.org/10.12989/was.2013.16.4.355
4. Cunnane, C.: Unbiased plotting positions a review. J. Hydrol. **37**(3–4), 205–222 (1978). https://doi.org/10.1016/0022-1694(78)90017-3
5. Erto, P., Lepore, A.: A note on the plotting position controversy and a new distribution-free formula. In: Proceeding of the 45th Scientific Meeting of the Italian Statistical Society, pp. 16–18 (2010)
6. Erto, P., Lepore, A.: New distribution-free plotting position through an approximation to the beta median. Adv. Theor. Appl. Stat., 23–27 (2013). https://doi.org/10.1007/978-3-642-35588-2_3
7. Erto, P., Lepore, A.: Best unbiased graphical estimators of location-scale distribution parameters: application to the Pozzuoli's bradyseism earthquake data. Environ. Ecol. Stat. **23**(4), 605–621 (2016). https://doi.org/10.1007/s10651-016-0356-9
8. Fuglem, M., Parr, G., Jordaan, I.: Plotting positions for fitting distributions and extreme value analysis. Can. J. Civ. Eng. **40**(2), 130–139 (2013). https://doi.org/10.1139/cjce-2012-0427
9. Gringorten, I.I.: A plotting rule for extreme probability paper. J. Geophys. Res. **68**(3), 813–814 (1963). https://doi.org/10.1029/JZ068i003p00813
10. Gumbel, E.: Statistics of Extremes, vol. 247. Columbia University Press, New York (1958)
11. Hong, H.P.: Selection of regressand for fitting the extreme value distributions using the ordinary, weighted and generalized least-squares methods. Reliab. Eng. Syst. Saf. **118**, 71–80 (2013). https://doi.org/10.1016/j.ress.2013.04.003
12. Hong, H.P., Li, S.H.: Plotting positions and approximating first two moments of order statistics for Gumbel distribution: estimating quantiles of wind speed. Wind Struct. An Int. J. **19**(4), 371–387 (2014). https://doi.org/10.12989/was.2014.19.4.37
13. International Electrotechnical Commission (IEC): Wind Turbines – Part 1: Design Requirements. IEC 61400-1, 3rd edn (2005)
14. Lawless, J.: Confidence interval estimation for the Weibull and extreme value distributions. Technometrics **20**(4), 355–364 (1978)
15. Leon Harter, H.: Another look at plotting positions. Commun. Stat.-Theory Methods **13**(13), 1613–1633 (1984). https://doi.org/10.1080/03610928408828781
16. Lieblein, J.: Efficient methods of extreme-value methodology. Technical Report, Institute for Applied Technology, National Bureau of Standards, Washington, D.C. (1976)
17. Makkonen, L.: Bringing closure to the plotting position controversy. Commun. Stat.-Theory Methods **37**(3), 460–467 (2008). https://doi.org/10.1080/03610920701653094

18. Makkonen, L., Pajari, M., Tikanmäki, M.: Closure to Problems in the extreme value analysis (Struct. Safety : 30: 405–419). Saf. Struct., 2013 (2008)
19. Makkonen, L., Pajari, M., Tikanmäki, M.: Discussion on plotting positions for fitting distributions and extreme value analysis. Can. J. Civil. Eng. **40**(9), 927–929 (2013)
20. Pirouzi Fard, M.N.: Probability plots and order statistics of the standard extreme value distribution. Comput. Stat. **25**(2), 257–267 (2010). https://doi.org/10.1007/s00180-009-0174-8
21. Pirouzi Fard, M.N., Holmquist, B.: Approximations of variances and covariances for order statistics from the standard extreme value distribution. Commun. Stat. - Simul. Comput. **37**(8), 1500–1506 (2008). https://doi.org/10.1080/03610910802244059
22. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2016)

Estimating High-Dimensional Regression Models with Bootstrap Group Penalties

Valentina Mameli, Debora Slanzi and Irene Poli

Abstract Currently many research problems are addressed by analysing datasets characterized by a huge number of variables, with a relatively limited number of observations, especially when data are generated by experimentation. Most of the classical statistical procedures for regression analysis are often inadequate to deal with such datasets as they have been developed assuming that the number of observations is larger than the number of the variables. In this work, we propose a new penalization procedure for variable selection in regression models based on Bootstrap group Penalties (BgP). This new family of penalization methods extends the bootstrap version of the LASSO approach by taking into account the grouping structure that may be present or introduced in the model. We develop a simulation study to compare the performance of this new approach with respect several existing group penalization methods in terms of both prediction accuracy and variable selection quality. The results achieved in this study show that the new procedure outperforms the other penalties procedures considered.

Keywords Bi-level selection · Bootstrap · High-dimensionality · Regression models · Variable selections

V. Mameli (✉)
Department of Environmental Sciences, Informatics and Statistics,
Ca' Foscari University of Venice, via Torino 155, Mestre (IT), Italy
e-mail: mameli.valentina@virgilio.it

V. Mameli · D. Slanzi · I. Poli
European Centre for Living Technology, S. Marco 2940, Venice (IT), Italy

I. Poli
e-mail: irenpoli@unive.it

D. Slanzi
Department of Management, Ca' Foscari University of Venice, Cannareggio 873,
30121 Venice (IT), Italy
e-mail: debora.slanzi@unive.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_35

1 Introduction

New powerful technologies can produce datasets characterized by a huge number of variables, for example in fields such as genomics and micro-array experimentation. Such datasets motivate the recent development of efficient new statistical tools for modelling and inference. Recent research focuses on variable selection procedures based on different families of penalizations for regression models, and these procedures seem to provide estimated models with good predictive performances (see [8] and [14] for reviews of this research). We can identify three main classes of methods in this research. The first class is related to individual variable selection; among the procedures of this class, the Least Absolute Shrinkage Selection Operator (LASSO) proposed in [20] is surely the most used and well-known. In LASSO the number of selected variables is limited by the sample size and it presents a penalty that tends to select only one or a few from a set of highly correlated relevant variables. The second class is related to group variable selection and the third class to bi-level selection procedures. When a grouping structure is introduced into a model, interest may rely entirely on selecting relevant groups and not individual variables, but when both individual variables and groups are relevant, bi-level selection procedures can be adopted to select both the relevant groups and variables within these groups. Examples of procedures in these two classes include the group LASSO method [22], the Smoothly Clipped Absolute Deviation penalty [7], the Minimax Concave Penalty method [23], the composite MCP [4], the group Bridge penalty [12] and the group exponential LASSO [3]. These selection procedures have been introduced with the aim of overcoming some limitations of the original LASSO approach and present a number of appealing properties in terms of both estimation accuracy and variable selection properties.

Addressing the problem of estimating regression models with high dimensionality and a small number of observations, as in problems where data are generated by laboratory experimentation, it can be useful to adopt bootstrap re-sampling techniques [6, 9]. These techniques are in fact able to change the initial dataset and gain information from the multiple pseudo-datasets resulting from the bootstrap procedure. This approach was suggested in a LASSO framework by [2]. In this paper we make a further development to this approach by introducing a new family of penalization procedures obtained by combining the properties of penalized group procedures with bootstrap re-sampling methods. We call this approach Bootstrap group Penalties (BgP). BgP is based on the idea that group sparsity can be a very useful informative element in inferring statistical model with high dimensionality. In fact, this approach combines and extends the concepts of the individual and group variables penalties with re-sampling techniques. We evaluated the performance of BgP conducting some simulation studies and we noticed that this new approach is able to capture the benefits of sparsity in high-dimensional settings both at individual and group level.

The paper is organized as follows. In Sect. 2 we present the model and shortly review the most relevant penalized regression procedures. In Sect. 3 we introduce the novel BgP family of penalized procedures and in Sect. 4 we evaluate the performance of the approach in simulation studies. Section 5 presents some concluding remarks.

2 The Regression Model and Variable Selection Penalties

We consider a multiple linear regression model

$$y = X\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $y \in \mathbb{R}^n$ is the response vector, X is the $n \times p$ design matrix, n denotes the sample size and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ represents the vector of regression coefficients where the number of covariates p is large and exceeds the number of observations n ($p > n$). Moreover, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ is the error vector, and we assume ε_i, for $i = 1, \ldots, n$ have 0 mean, constant variance σ^2 and they are independent with normal distribution, $\varepsilon \sim N(0, \sigma^2 I)$. The model is generally referred as the high-dimensional regression model.

To address the estimation problem of this model, the assumption that is commonly adopted is that the parameter $\boldsymbol{\beta}$ is sparse in the sense that many of its elements are zero, i.e. most of the covariates have small or no effect on the response variable.

Several penalized regression procedures, also known as regularized regression methods, have been proposed in the statistical literature to address the inferential problem of regression models with the number of covariates much larger than the number of observations. In these procedures, the vector of regression coefficients $\boldsymbol{\beta}$ is estimated by minimizing the penalized least squares criterion $Q(\cdot)$ composed of two elements: the least square loss function, $\frac{1}{2n}(y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta})$, and a penalty function $P(\cdot)$:

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}(y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta}) + P(\boldsymbol{\beta}|\lambda). \tag{2}$$

The penalty function $P(\cdot)$ controls the complexity of the model. There are several possible choices for the penalty function tailored to the scientific problem under consideration. The parameter λ is a tuning parameter which can be assessed by using cross validation technique or information criteria such as the Akaike or the Bayesian information criteria [1, 21]. Depending on the type of variable selection, penalized regression approaches can be classified into three wide classes: individual, group and bi-level variable selection procedures.

In the following we shortly describe the main variable selection procedures based on penalizations.

2.1 Individual Variable Selection

Among the most prominent penalized procedures for individual variable selection, we mention the Least Absolute Shrinkage Selection Operator (LASSO) proposed by [20] which is based on the L_1 penalty, i.e. $P(\boldsymbol{\beta}|\lambda) = \lambda \sum_{j=1}^{p} |\beta_j|$. One characteristic of the LASSO penalty is the ability to allow both continuous shrinkage and automatic variable selection; it is able to exclude irrelevant variables and produce sparse estimators. Despite its good properties, the LASSO procedure has some drawbacks as described by [7]. LASSO tends in fact to select as informative variables also those variables not actual relevant for the model. Moreover, it is well known that LASSO does not achieve selection consistency properties when the parameter λ is chosen by minimizing the prediction error [16]. It also presents some difficulties in case of correlated covariates.

A different penalty for individual variable selection is the Smoothly Clipped Absolute Deviation (SCAD) penalty function [7], defined as:

$$P(\boldsymbol{\beta}|\lambda,\gamma) = \lambda \sum_{j=1}^{p} \int_0^{|\beta_j|} \mathbb{I}_{\{t\leq\lambda\}} + ((\gamma\lambda - t)_+/(\gamma-1)\lambda)\mathbb{I}_{\{t>\lambda\}} dt, \text{ with } \lambda \geq 0 \text{ and } \gamma > 2, \tag{3}$$

where $\mathbb{I}_A$ denotes the indicator function of a set A and a_+ represents the non-negative part of a, and λ e γ are two tuning parameters.

Finally, we mention the Minimax Concave Penalty (MCP) described by [23]

$$P(\boldsymbol{\beta}|\lambda,\gamma) = \lambda \sum_{j=1}^{p} \int_0^{|\beta_j|} (1 - (t/\gamma\lambda))_+ dt, \text{ with } \lambda \geq 0 \text{ and } \gamma > 1. \tag{4}$$

The parameter γ controls the concavity in both SCAD and MCP penalties: small values of γ indicate that the penalty tends to be concave. It is interesting to note also that when $\gamma \to \infty$ both SCAD and MCP reduce to the LASSO penalty.

2.2 Group Variable Selection

In high dimensional regression settings, approaches based on the identification of groups of covariates has been proposed in the literature to reduce the dimensionality of the model [4, 17]. The information contained in the identified grouping structures can be exploited in the regression model in order to enhance its prediction capacities. Considering p covariates grouped into K non-overlapping clusters, the multivariate linear regression model is described in the following form:

$$y = \sum_{k=1}^{K} \tilde{X}_k \tilde{\beta}_k + \varepsilon, \tag{5}$$

where $\tilde{X}_k$ is the $n \times d_k$ design matrix formed by the d_k covariates belonging to the k-th cluster, $\tilde{\beta}_k = (\beta_{k1}, \ldots, \beta_{kd_k}) \in \mathbb{R}^{d_k}$ is the vector of regression coefficients of the k-th cluster and ε is the error vector.

One of the earliest group penalizations was proposed by [22] as an extension of the LASSO. This procedure, called group LASSO (`gLASSO`), penalizes the L_1 norms of the groups variables coefficients as follows

$$P(\boldsymbol{\beta}|\lambda) = \lambda \sum_{k=1}^{K} c_k ||\tilde{\beta}_k||_{R_k}. \tag{6}$$

Here, the coefficients c_k are introduced to adjust the procedure for the group size. In addition, R_k are $d_k \times d_k$ positive definite matrices ([22]), which satisfies $||\tilde{\beta}_k||_{R_k} = \tilde{\beta}_k^T R_k \tilde{\beta}_k$. A common choice for R_k is the Gram matrix based on $\tilde{X}_k$, i.e. $R_k = \tilde{X}_k^T \tilde{X}_k / n$. The group LASSO shows excellent properties in terms of both prediction and estimation errors, and its selection consistency relies on the assumption that the design matrix satisfies a particular condition (the *irrepresentable condition* as defined in [24]) which becomes infeasible in the high-dimensional setting. Moreover, the group LASSO shows superior performance with respect to the standard LASSO when the *strong group sparsity* condition is fulfilled; see [13].

Following the group LASSO, other group penalizations have been introduced in the literature. We mention the group Bridge penalty (`gBridge`) proposed by [12] and defined as follows:

$$P(\boldsymbol{\beta}|\lambda, \gamma) = \lambda \sum_{k=1}^{K} \left(|\beta_{k1}| + \cdots + |\beta_{kd_k}| \right)^{\gamma}, \tag{7}$$

where $0 < \gamma < 1$ is a tuning parameter.

A class of penalties with group variable selection properties, which encompasses the group LASSO, could be obtained by considering the family of the following penalties proposed by [14] and defined as follows:

$$P(\boldsymbol{\beta}|\lambda, \gamma) = \sum_{k=1}^{K} \rho_{\tilde{\lambda},\gamma}(||\tilde{\beta}_k||_{R_k}), \tag{8}$$

where $\rho_{\tilde{\lambda},\gamma}(\cdot)$ is a concave function with $\tilde{\lambda} = c_k \lambda$. Some possible choices for ρ include the MCP and the SCAD penalties, which applied to (8) lead to the group MCP (`gMCP`) and group SCAD, respectively, as derived in [14].

2.3 Bi-level Variable Selection

Bi-level variable selection penalties can be obtained by combining individual and group variable penalties [4], and they are defined as

$$P(\boldsymbol{\beta}|\lambda, \gamma_O, \gamma_I) = \sum_{k=1}^{K} \rho_{\lambda,\gamma_O} \left(\sum_{j=1}^{d_k} \rho_{\lambda,\gamma_I} \left(|\beta_{kj}| \right) \right) \tag{9}$$

where the penalty $\rho_{\lambda,\gamma_O}(\cdot)$, called the outer penalty, incorporates the information present in the group structure, while the penalty $\rho_{\lambda,\gamma_I}(\cdot)$, called inner penalty, incorporates information on the individual covariates. The two penalties are able to identify relevant variables by exploiting the information contained in the cluster structure both at individual and at group levels. The parameters γ_O and γ_I are tuning parameters with $\gamma_O = d_k \gamma_I \lambda/2$. Simulation studies on the tuning parameters γ_O and γ_I could be found in [4] and references therein. Possible choices for the outer and inner penalties include the MCP and the SCAD penalties. In particular, an interesting special case of this class of penalties is the composite Minimax Concave Penalty (`cMCP`), defined by [4], which can be obtained from equation (9) by using as inner and outer penalties the MCP penalty given in equation (4). It should be also noted that the group Bridge and the group LASSO can be embedded into the framework of penalties as in equation (9). In fact, the group Bridge can be represented in this framework by assuming as outer penalty the Bridge penalty and as inner penalty the LASSO penalty, the group LASSO can be constructed by assuming the Bridge penalty and the ridge penalty as ρ_O and ρ_I, respectively. An alternative approach for constructing bi-level selection penalties has been developed in [18] by considering convex combination of individual- and group-level variable selection methods. Another penalty belonging to this third class of penalties is the group exponential LASSO (`gel`) proposed by [3]; this penalty belongs to the class of concave 1-norm group penalties [14], and is defined as

$$P(\boldsymbol{\beta}|\lambda, \tau) = \sum_{k=1}^{K} \rho_{\lambda,\tau} \left(||\tilde{\beta}_k||_1 \right), \tag{10}$$

where ρ is the exponential penalty, i.e. $\rho_{\lambda,\tau}(\theta) = \frac{\lambda^2}{\tau} \left\{ 1 - \exp\left(-\frac{\tau\theta}{\lambda} \right) \right\}$, λ and τ are two tuning parameters; τ represents the rate of exponential decay. If $\tau < 1$ the objective function in equation (2) is strictly convex with a unique global minimum; see [3].

These selection procedures have been introduced in literature with the main objective to overcome some limitations of the LASSO estimator. Aim of this contribution is in fact to improve the performances of these procedures in a high-dimensional regression setting, when the number of observations is small.

3 The Family of Bootstrap Group Penalties (BgP)

In order to address the estimation problem of the multiple linear regression model characterized by a large number of covariates and a small number of observations, we introduced a novel family of regression penalties. This family is based on bootstrap re-sampling technique in combination with group and bi-level variable selection penalties. We call this new family Bootstrap group Penalties (BgP).

Under the general structure of the multiple regression model defined in equation (1), we consider n observations $(X_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, and B bootstrap pseudo-replications of the n pairs (X_i, y_i) and we build the novel procedure according to the following steps:

1. We derive B bootstrap replications of the n pairs (X_i, y_i), that is for $b = 1, \ldots, B$, the subset $(X_{bi}, y_{bi}) \in \mathbb{R}^p \times \mathbb{R}$ is sampled at random with replacement from the original training set (X_i, y_i), $i = 1, \ldots, n$. Then at each bootstrap iteration, we estimate the regression parameters β_j, for $j = 1, \ldots, p$, by using a penalized group (or bi-level) selection procedure.
2. We identify the set J_b of the indices of the covariates selected by the penalized group (or bi-level) selection procedure at each bootstrap iteration b, namely the sets
$$J_b = \{j | \hat{\beta}_j^b, j = 1, \ldots, p\}, \quad b = 1, \ldots, B,$$
where $\hat{\beta}_j^b$ for $j = 1, \ldots, p$ are the non-zero coefficients of the covariates selected at each bootstrap iteration b.
Among all the B sets J_b, only the covariates with a frequency higher than a defined threshold π in the B bootstrap replications were identified as the covariates to include into the model.
3. We estimate the regression model through a penalized group (or bi-level) selection procedure using only the previously selected covariates.

4 Simulation Studies

We evaluate the performances of the Bootstrap group Penalties family approach by conducting simulation studies. In particular, we develop two comparative studies to evaluate the group penalization procedures and the novel corresponding bootstrap procedures. Among the penalties described in Sect. 2, we consider: the group Bridge (`gBridge`), the composite MCP (`cMCP`), the group Maximum Concave Penalty (`gMCP`), the group exponential LASSO (`gel`) and the group LASSO (`gLASSO`). The novel penalties belonging to the BgP family are identified as follows: the Bootstrap group Bridge (`BgBridge`), the Bootstrap composite cMCP(`BcMCP`), the Bootstrap group MCP (`BgMCP`), the Bootstrap group exponential lasso (`Bgel`) and the

Bootstrap group LASSO (BgLASSO). We select $B = 500$ bootstrap samples, and to evaluate the robustness of the approach we run 1000 replications for the first simulation and 500 replications for the second simulation. In the second simulation study the number of replications is fixed to 500 due to higher computational burden. The performance of the approaches is then evaluated with measures of prediction accuracy and variable selection efficiency: the Predictive Mean Square Error (PMSE), the Sensitivity measure (the ratio between the number of selected relevant variables and the number of defined relevant variables), and the Specificity measure (the ratio between the number of removed non relevant variables and the number of non relevant defined variables) as defined in [11].

4.1 First Simulation Study: Uncorrelated Covariates

In this simulation we assume a multivariate linear regression model as described in equation (1) where $\varepsilon_i \sim N(0, \sigma^2)$ and σ takes the value 3. We also assume that covariates were generated from the normal distribution as in the study proposed by [3]. For the grouping structure, we consider the following setup: 10 groups with 20 variables in each group ($p = 200$), $n = 100$, the number of non zero coefficients is 4 and all the non zero coefficients belong to the same group. To evaluate the prediction accuracy we split the data into training and testing datasets. The results of this comparison are presented in Table 1.

Comparing the different model penalties we can notice that Bootstrap group Penalties family (BgBridge, BcMCP, BgMCP, Bgel, BgLasso) is able to achieve much better performances in almost all the comparisons proposed. We can also highlight the very good results in prediction of Bgel with respect to all the other approaches.

Table 1 Comparison of the performance of penalties procedures based on Predictive Mean Square Error (PMSE), Sensitivity and Specificity (1000 replications). In bold we present the best performance of the models based on the selected penalties

Model penalties	PMSE	Sensitivity	Specificity
gBridge	0.885 (0.176)	0.934 (0.122)	0.843 (0.025)
BgBridge	**0.835** (0.293)	**1.000** (0.000)	**0.907** (0.031)
cMCP	0.936 (0.207)	0.797 (0.190)	0.849 (0.014)
BcMCP	**0.839** (0.150)	**0.910** (0.136)	**0.913** (0.014)
gMCP	1.281 (0.257)	1.000 (0.000)	0.489 (0.077)
BgMCP	**1.038** (0.911)	1.000 (0.000)	**0.899** (0.040)
gel	0.949 (0.185)	1.000 (0.000)	0.462 (0.083)
Bgel	**0.755** (0.608)	0.9998(0.008)	**0.937**(0.044)
gLASSO	**0.924** (0.221)	0.554 (0.497)	**0.916** (0.158)
BgLASSO	1.293 (0.544)	**1.000** (0.000)	0.652 (0.104)

This penalization is able to select the actually relevant variables of the model as suggested by its sensitivity and specificity.

4.2 Second Simulation Study: Correlated Covariates

This simulation, based on the same structure of the multiple linear regression model as described in the previous simulation study, has been conducted to evaluate the performance of the approach when a correlation structure between covariates is present. This is motivated by the fact that there are many scientific domains where structures of grouped predictors arise. Examples include genetic association studies, where genetic markers from the same gene can be considered a group as in fMRI data analysis; see [19]. It is obvious that variables belonging to the same group have similar characteristics, therefore, an higher within-group correlation among the members of a group is expected. In this simulation study we consider the following setup: 10 groups with 50 variables in each group ($p = 500$), $n = 350$. We assume covariates are generated from a multivariate normal distribution with zero mean and covariance matrix $\Sigma_{p\times p} = \Sigma_{base} + blockdiagonal(\Sigma_1, \ldots, \Sigma_{10})$. According to [15], we also assume a correlation structure between covariates: the Σ_{base} is a $p \times p$ symmetric matrix with correlation among covariates $\rho = 0.1$, Σ_k is a 50×50 symmetric covariance matrix with within-group correlation $\rho = 0.6$, for $k = 1, \ldots, 10$. The vector of regression coefficients is defined as follows $(\beta_1, \ldots, \beta_{60}) = (a, -a, a, -a, a, -a)$ and $(\beta_{61}, \ldots, \beta_{500}) = (0, \ldots, 0)$ with $a = (0, 0, 0.2, 0.25, 0.5, 0, 0, 0.2, 0.25, 0.5)$, and with a total number of non zero coefficients equals to 36. We pointed out that in this simulation the relevant covariates belongs to two different groups. The number of predictors and the number of non-zero coefficients is the same as proposed in the paper [15]. To evaluate the prediction accuracy we split the data into training and testing datasets. The results of this simulation are presented in Table 2.

Table 2 Comparison of the performance of penalties procedures based on Predictive Mean Square Error (PMSE), Sensitivity and Specificity (500 replications). In bold we present the best performance of the models based on the selected penalties

Model penalties	PMSE	Sensitivity	Specificity
`gBridge`	1.39 (0.17)	**0.99** (0.01)	0.98 (0.01)
`BgBridge`	**1.31** (0.15)	0.96 (0.03)	**1.00** (0.00)
`cMCP`	1.56 (0.23)	0.94 (0.04)	0.98 (0.01)
`BcMCP`	**1.43** (0.18)	**0.95** (0.03)	**0.99** (0.00)
`gMCP`	3.52 (5.05)	1.00 (0.00)	0.82 (0.10)
`BgMCP`	**1.79** (0.23)	1.00 (0.00)	**0.86** (0.01)
`gel`	3.43 (4.08)	**1.00** (0.01)	0.77 (0.38)
`Bgel`	**1.52** (0.26)	0.93 (0.02)	**0.96** (0.00)
`gLASSO`	**2.45** (0.63)	**1.00** (0.01)	0.86 (0.02)
`BgLASSO`	3.41 (1.55)	0.91 (0.08)	**0.91** (0.05)

From this simulation we can notice that the Bootstrap group Penalties family is able to improve the PMSE and the Specificity measures in almost all the compared model penalties as highlighted in bold in Table 2. The exception is just related to `gLASSO` as in the previous simulation study. Moreover, the complex structure setup assumed for this simulation shows the difficulties of `gMCP` and `gel` in prediction and the advantages of the corresponding bootstrap procedures. At last, we would like to highlight the very good performance of the proposed approach for the `BgBridge` in particular achieving the best values in prediction and Specificity.

5 Conclusion

In several fields of research where group structure can be identified in the collected dataset and the number of observations is too small with respect to the number of covariates, the combination of different methodologies can help in deriving effective regression models. The classical statistical procedures for regression can be enhanced for the analysis and modelling of such dataset. In this work we propose to combine the penalized group and bi-level variable selection approaches with bootstrap methods to handle high-dimensional datasets and small number of observations. The results of the simulation studies that we conducted suggest that this approach is promising in estimating reliable and efficient regression models when the number of covariates exceeds the number of the observations and group structures are detected in the data. This method could be easily adapted to handle other penalization procedures and other re-sampling techniques allowing the construction of regression models also in difficult contexts.

Moreover, the approach can be extended to overlapping groups as they are frequently observed in several research fields.

Appendix

Computational tools: the statistical analyses were performed using the `R`-project free software environment for statistical computing. In particular, we use the `R`-package `grpreg` to fit the group-penalized regression models, such as the `gel`, `cMCP`, `gMCP`, `gBridge`, `gLASSO` ([3–5]). `LASSO` was fitted by using functions of the `R`-package `glmnet` ([10]).

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (eds.) 2nd International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)
2. Bach, F.R.: Bolasso: model consistent lasso estimation through the bootstrap. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland (2008)
3. Breheny, P.: The group exponential lasso for bi-level variable selection. Biometrics **71**, 731–740 (2015)
4. Breheny, P., Huang, J.: Penalized methods for bi-level variable selection. Stat. Its Interface **2**(3), 369–380 (2009)
5. Breheny, P., Huang, J.: Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat. Comput. **25**, 173–187 (2015)
6. Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. **7**(1), 1–26 (1979)
7. Fan, J., Li, R.: Variable selection via non-concave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)
8. Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. Stat. Sin. **20**, 101–148 (2010)
9. Fang, K., Ma, S.: Analyzing large datasets with bootstrap penalization. Biom. J. **59**(2), 358–376 (2017)
10. Friedman, J., Hastie, T., Robert Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)
11. Geng, Z., Wang, S., Yu, M., Monahan, P.O., Champion, V., Wahba, G.: Group variable selection via convex log-exp-sum penalty with application to a breast cancer survivor study. Biometrics **71**(1), 53–62 (2015)
12. Huang, J., Ma, S., Xie, H., Zhang, C.: A group bridge approach for variable selection. Biometrika **9**, 339–355 (2009)
13. Huang, J., Zhang, T.: The benefit of group sparsity. Ann. Stat. **38**, 1978–2004 (2010)
14. Huang, J., Breheny, P., Ma, S.: A selective review of group selection in high-dimensional models. Stat. Sci. **27**(4), 481–499 (2012)
15. Jiang, D., Huang, J.: Concave 1-norm group selection. Biostatistics **16**(2), 252–267 (2015)
16. Leng, C., Lin, Y., Wahba, G.: A note on the lasso and the related procedures in model selection. Stat. Sin. **16**, 1273–1284 (2006)
17. Mameli, V., Lunardon, N., Khoroshiltseva, M., Slanzi, D., Poli, I.: Reducing dimensionality in molecular systems: a bayesian non-parametric approach. In: Rossi F., Piotto S., Concilio S. (eds.) Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry, WIVACE 2016. Communications in Computer and Information Science, vol. 708, pp. 114–125. Springer, Cham (2017)
18. Noah, S., Friedman, J., Hastie, T., Tibshirani, R.: A sparse-group lasso. J. Comput. Graph. Stat. **22**(2), 231–245 (2013)
19. Rish, I., Grabarnik, G.: Sparse Modeling: Theory, Algorithms, and Applications. CRC Press Inc, Boca Raton (2014)
20. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B **58**(1), 267–288 (1996)
21. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
22. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc., Ser. B **68**, 49–67 (2006)
23. Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)
24. Zhao, P., Yu, B.: On model selection consistency of Lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)